Evaluating Gender Bias in AI Applications Using Household Survey Data

Executive Summary

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

Executive Summary

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.

Introduction

Over the past year, AidData, in partnership with CDD-Ghana, has worked to evaluate the role of gender and 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, and household surveys, along with CDD-Ghana's knowledge of the local context 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, utilizing 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 AI 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.

The project's efforts to address these challenges and maintain the standards of a truly equitable AI approach ultimately produced valuable insight into the influence of gender on AI-based wealth estimates and how gender bias may be evaluated and factored into future applications. 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.

Core lessons learned over the course of this project can provide others working towards equitable applications of AI with critical knowledge they can apply to their own work. Lessons learned included: 1) Equitable AI is not always easy - figuring out how to evaluate gender or other potential bias can be a complex exercise. 2) First steps matter - even if an initial approach to implement a more equitable AI solution is not perfect, it serves as a

necessary stepping stone and opens a conversation around the importance of equity and implications of less equitable applications. 3) Research requires context - involving local partners like CDD-Ghana that understand and belong to the population which data and AI models are based on and may impact is critical to building both accurate and equitable AI solutions. Each of these these lessons are explored in greater depth later in this document

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 estimate use into their work. This executive summary and all other project outputs have been made available via aiddata.org/projects/equitable-ai. In the remainder of the executive summary, we will provide a brief overview of the activities implemented, what we learned, and implications for future work.

Technical Approach

The DHS Wealth Index (WI), an asset-based metric of household wealth, is one of the most widely used sources of training data for AI models which estimate wealth and is available with all contemporary DHS household surveys. Household and subnational level wealth estimates are a critical resource leveraged by governments and development organizations to target and evaluate efforts to improve wellbeing in developing nations. AI-based wealth estimation models trained on satellite derived geospatial data and existing DHS WI data help fill gaps where DHS survey data used for traditional wealth estimates does not exist. AI wealth estimation models have been shown to perform well in general across numerous studies, yet no work has explored their effectiveness at accurately capturing conditions for subpopulations, such as women, or their relative accuracy across subpopulations (e.g., women vs men). To explore variation in model performance or gender bias, we classify households surveyed in the 2014 Ghana DHS survey by gender and train separate AI models in order to compare them.

Since DHS assets are only recorded for the entire household, capturing gender-specific conditions can be difficult. The approaches used for classifying the gender of a household are informed by expert knowledge of existing methods and local conditions provided by CDD-Ghana. The baseline classification approach is based on the self-reported gender of the head of household from the DHS survey, while alternatives are derived leveraging country-specific context and trends related to gendered asset ownership and control identified by CDD-Ghana in a detailed report[1]. Gender-specific Random Forests (a type of AI model) are then trained using the DHS WI as the dependent variable, and a range of geospatial data (e.g., nighttime lights, population, land cover) as the independent variables (see Figure 1).

Alongside exploring varying household gender classification approaches, we develop and test models using a range of conditions or parameters. These include, among others:

[1] Highly gendered assets (based on control/decision making) identified in the report were combined with gendered asset ownership trends in the DHS data to select assets used to classify households by gender under one of the household classification approaches.

testing models training using different hyperparameters - settings which influence how models are built and trained; adjusting the data being used to train the model - such as modifying the ratio of households associated with each gender to explore the impact of gender-specific sample size and address sampling bias; and evaluating the use of different sets of geospatial features available to the model to learn from.

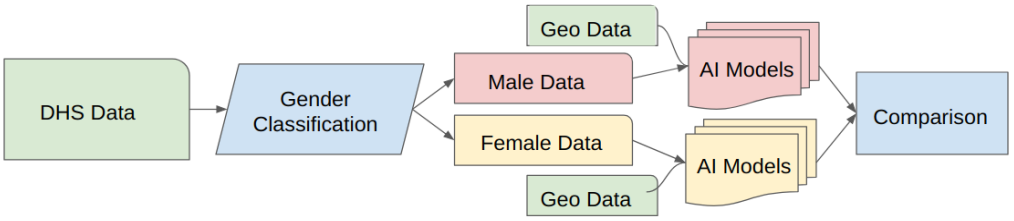

Figure 1. Workflow illustrating how data and models are prepared to evaluate the role of gender in wealth estimates.

Finally, to explore the influence of gender on the creation of the DHS WI itself, we also evaluate two alternative approaches to developing a wealth index based on asset ownership from the DHS data. We test 1) a gender-specific implementation of the DHS WI and 2) the International Wealth Index (IWI). The gender-specific DHS WI is useful because it uses the same wealth index construction as the standard DHS WI (which is influenced by the data for which the wealth index is being built on) but based only on each gendered subpopulation. The IWI is useful because it is based on a standardized construction that does not vary given the subpopulation.

Findings

We test the performance of random forest models trained across a range of parameters and inputs that impact model behavior, using data for each gender. Across gender agnostic and gender-specific tests, models perform well during cross validation and indicate good generalizability. However, models trained on male household data consistently outperform models trained on female household data (see Figure 2). Model performance is measured using a model's R^2 value - or the ability of a model to account for variation in the WI across clusters using the geospatial variables provided.

A critical consideration is the difference between the number of male and female households used in models. As there are typically more male households, we run robustness tests to address sampling bias in which we limit the number of households to be equal across genders. We find that when using equal household counts, male models still outperform female models, suggesting gender is still meaningfully influencing model performance.

A valuable feature of random forest models is the ability to measure the importance of specific features (geospatial variables) used to estimate wealth in the models. Male and female models had similar feature importance overall, and were most heavily influenced by a similar subset of features, including nighttime lights, urban area coverage, and population. The specific order of importance did vary slightly between genders. The ability

to use a smaller set of features for training is significant as it can facilitate practical applications by reducing data requirements.

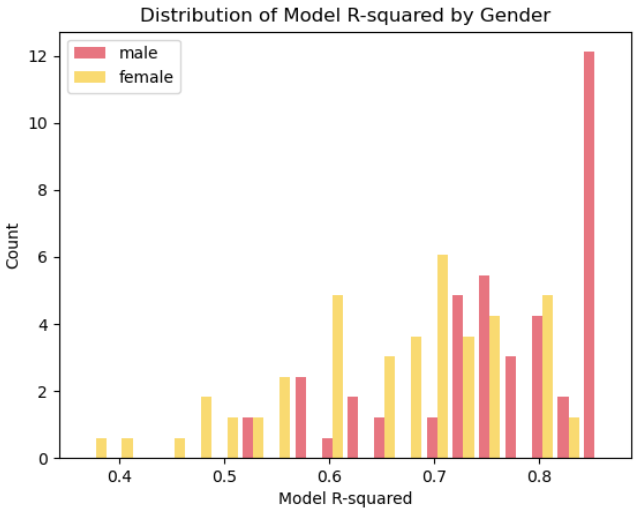

Figure 2. Comparison of the distribution of model R^2 by gender.

Additional efforts to understand the influence of gender on the DHS Wealth Index itself involved rebuilding the index using gender-specific data. After recreating the index using the gender-specific approach, we found that many female-led households in the lower wealth quintiles were classified as even less wealthy than in the original DHS WI. Similar results were found when comparing wealth index quintiles generated based on the IWI. While the differences between the original DHS WI and gender-specific WI or IWI are notable, there is insufficient information available without a "ground truth" to say which is more accurate.

Key Lessons

Equitable AI is still a young and evolving focal area in which there is still much to learn, particularly when it comes to specific use cases and applications. 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 of the data 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 more equitable use of AI-based estimates of wealth.

Lesson 1. Equitable AI is not always easy.
Not all machine learning training data is conducive to easy assessment of gender bias. Traditional examples of dealing with gender bias in AI models and training data typically have individual observations that are discrete and easily gendered, which makes identifying basic sampling gender bias - resulting in imbalance or skewed model behavior - more

straightforward. The data we use on wealth is more complicated because 1) it is collected at the household level and is not clearly tied to one gender and 2) household locations are aggregated to the cluster (i.e., roughly village level) to protect respondents, which limits the accuracy of household geospatial variables (e.g., nighttime lights). In this application we are also concerned with the influence of gender bias at a societal level - influencing the data itself and possibly what AI models learn - rather than just at a gender-of-household sampling level. Sampling bias (e.g., using more male households in a model) can be directly evaluated by balancing sampling across genders, after which remains the influence of the broader societal bias which we aim to evaluate. In addition to genuine gender discrepancies as they relate to individual wealth, there is the potential that the current DHS survey data collection introduces other types of gender bias such as how the head of household is determined. It can be difficult to truly isolate the influences of just one of these types of bias, but considering the multiple ways in which gender bias can be introduced to AI models is an important, and often application specific, aspect of Equitable AI.

Lesson 2. First steps matter.
Even when working towards an equitable version of an AI application is difficult, there is value in still making an effort to evaluate potential bias - even if the approach is not perfect. This is most critical in areas where gender / equitability have not been explored at all, as it provides a starting point for additional work, and informs end users that a lack of analysis does not imply there is no bias. Early efforts also create an opportunity to discuss Equitable AI in a specific sector / space, and open up important conversations around what are the tradeoffs involved in using AI for certain applications, the cost of allowing bias to remain in the real world, and whether AI should be used at all if it cannot control for equity.

Lesson 3. Research requires context.
The first piece of this lesson, as is becoming increasingly established in Equitable AI practices, is that involving the populations that data is based on and it will impact - and doing so early and throughout the lifecycle of an AI product - is essential. For wealth estimation and gender bias in particular, knowledge of the local context around gender conditions is critical to producing accurate models and bias assessment and requires involvement of informed local partners. For example, the practical meaning of being a head of household, and how it relates to survey responses, alone could significantly redefine how our approach was implemented, and may vary from one country to another. CDD-Ghana's role in this project was critical to the development of the household gender classification approaches used to evaluate the influence of gender in the AI wealth estimates.

While our approach was in many ways intended to be replicable and extensible to other applications, there is a very real reliance on local knowledge to inform critical aspects such as how households are classified by gender. But even with local involvement, translating findings from research and theory into practice is not always straightforward. Disparate influences of 1) external efforts to use AI to inform decision-making (e.g., by researchers or aid organizations in the global north), 2) local societal norms around gender and equitability, and 3) practical focus of AI use by in-country actors can make advancing Equitable AI challenging.

Outreach & Dissemination

Knowledge sharing and accessibility are critical to the growth of Equitable AI. To foster deeper engagement with our work across a range of relevant audiences, we leveraged multiple avenues of outreach and dissemination. Our efforts focused on 1) local engagement with stakeholders and organizations in Ghana, 2) awareness and knowledge sharing among a global set of development practitioners, and 3) practical access and use of the data and tools needed to replicate or build upon our work.

In-country efforts were spearheaded by CDD-Ghana and focused on identifying and engaging with organizations whose work intersected with our project's focus with regards to use of AI, geospatial data, gender bias and equitability, or use of household survey data. Early foundational work focused on identifying relevant organizations, spanning across private industry, academic institutions, government agencies, NGOs, and advocacy groups.

The culmination of the in-country engagement was a workshop hosted by CDD-Ghana in Accra which brought together approximately 40 in-person participants and was live streamed on Zoom and Youtube to over 30 virtual participants and has since been viewed by many others. Presentations on our project and findings as well as on the forthcoming 2022 DHS round in Ghana were moderated and discussed by a panel of Ghanaian experts on AI, gender, and population statistics. In subsequent Q&A sessions participants were deeply engaged in discussions around the influence of biased gender norms within the country, concerns around the use of AI more broadly, and how Ghana can embrace AI to improve lives. Notable concerns raised included whether a nation struggling with more basic issues is ready for AI, and whether AI risks making conditions worse (e.g., people losing jobs to AI).

While the overall sentiment towards AI was optimistic even in the face of these concerns, one participant emphasized that "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." A number of Ghanaian media organizations released articles on the workshop, reiterating some of these core concerns, and a popular morning radio show interviewed AidData and CDD-Ghana researchers on the topic, extending the reach of the workshop's contents further.

To reach a broader audience of international development practitioners, researchers, and decision makers, we have generated materials which will be disseminated across a range of platforms and formats, and will leverage AidData's existing online presence and networks. Outputs include a blog post; social media posts across LinkedIn, Facebook, & Twitter; and inclusion in AidData and William & Mary newsletters. All project materials, including the full length technical report and this executive summary, will be made publicly available through a dedicated project page on AidData's website. In addition, the geospatial data and code used to train models and conduct the analysis are publicly available on Github to support replication and future use.

Future Work

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 based on head of household or specific asset ownership is appropriate, and, more broadly, whether future surveys can be improved to assess gender-specific wealth and support gender-specific AI applications in general. The self-reported head of household responses alone have the potential to simplify or obscure complex household dynamics based on culture and other factors, while classifying household gender based purely on asset ownership is difficult and imprecise. Additional survey responses on who provides the household income, who purchases or controls assets, and makes associated decisions may provide a more realistic view of household gender dynamics.

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? While our application leverages primarily globally available and satellite derived geospatial features, a range of additional data on employment, health, and more can be available on a country by country basis. Future research might also explore the possibility of utilizing wealth indices other than those produced by the DHS. As indicated by the preliminary analysis of a gender-specific DHS WI and the IWI, alternative wealth indices may be more suited to gender equitable applications. These areas of exploration and open questions provide meaningful and discrete avenues for others to build upon in the future.

Related Project Materials

In addition to this Executive Summary, all project materials and associated outputs are publicly available and can be accessed through the project page at aiddata.org/projects/equitable-ai. This includes the full length Technical Report, the local context report produced by CDD-Ghana (a summary of which is available as an appendix in the Technical Report), the GitHub repository, a recording of the in-country workshop, and other associated links such as to join the Equitable AI Community of Practice group on LinkedIn.