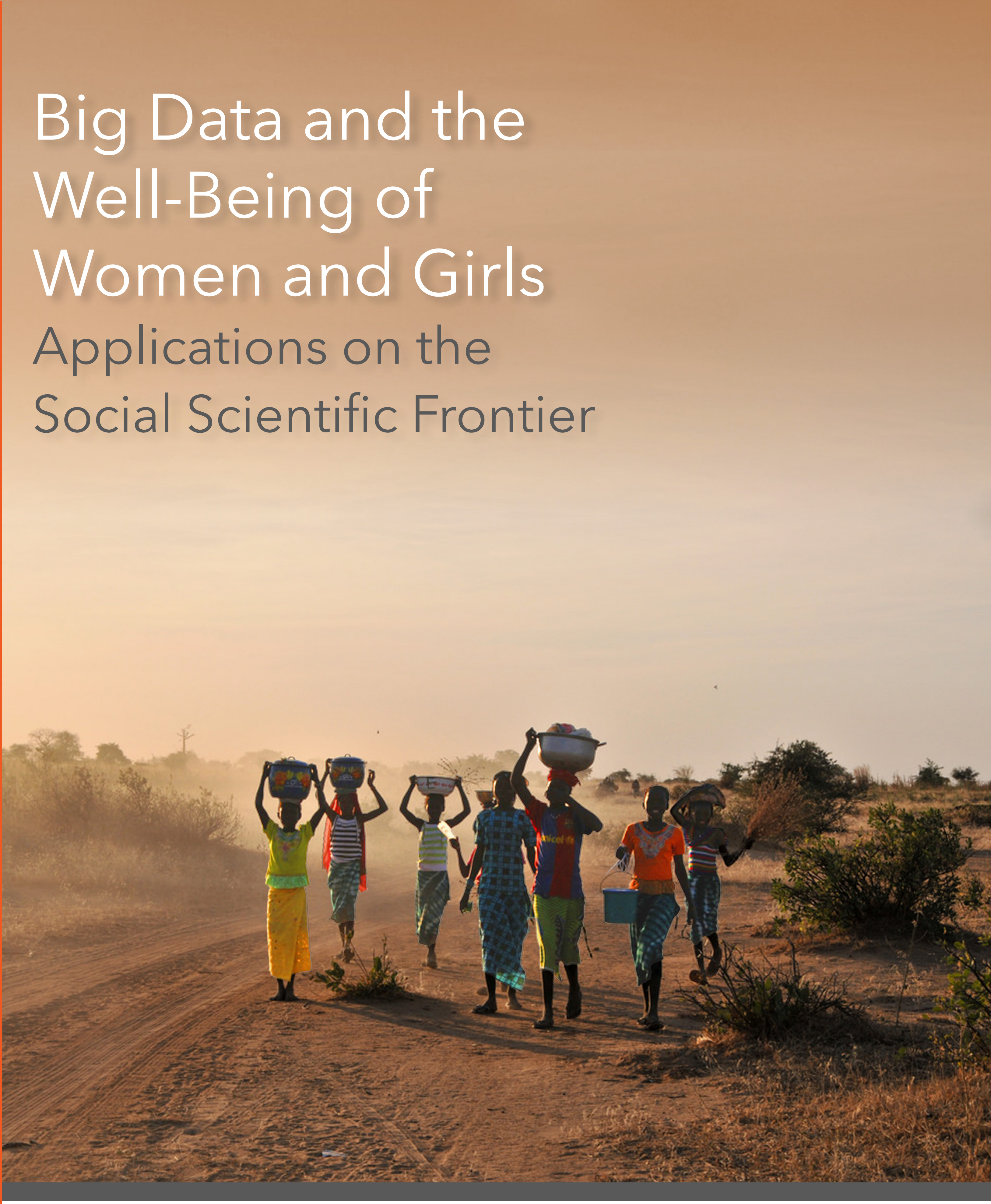
# Big Data and the Well-Being of Women and Girls

## Applications on the Social Scientific Frontier



April 2017

data2x

This report was written by Bapu Vaitla (overall); Claudio Bosco, Victor Alegana, Tom Bird, Carla Pezzulo, Graeme Hornby, Alessandro Sorichetta, Jessica Steele, Cori Ruktanonchai, Nick Ruktanonchai, Erik Wetter, Linus Bengtsson, and Andrew J Tatem (Section II); Riccardo Di Clemente, Miguel Luengo-Oroz, and Marta C. González (Section III); United Nations Global Pulse and University of Leiden, lead authors René Nielsen, Thomas Baar, and Felicia Vacarelu (Section IV, part 1); Munmun de Choudhury, Sanket Sharma, Tomaz Logar, Wouter Eekhout, and René Nielsen (Section IV, part 2). We thank Anoush Tatevossian and Robert Kirkpatrick for their guidance and access to key datasets. Editing and design by Julia Van Horn. We are grateful to Rebecca Furst-Nichols, Mayra Buvinic, Emily Courey Pryor, and Alba Bautista for helpful comments and guidance.

This work was initiated by Data2X, a collaborative technical and advocacy platform dedicated to improving the quality, availability, and use of gender data in order to make a practical difference in the lives of women and girls worldwide. Data2X works with UN agencies, governments, civil society, academics, and the private sector to close gender data gaps, promote expanded and unbiased gender data collection, and use gender data to improve policies, strategies, and decision-making. Hosted at the United Nations Foundation, Data2X receives funding from the William and Flora Hewlett Foundation and the Bill & Melinda Gates Foundation.

Cover photo © Elizabeth Whelan

# CONTENTS

# EXECUTIVE SUMMARY

Conventional forms of data—household surveys, national economic accounts, institutional records, and so on—struggle to capture detailed information on the lives of women and girls. The many forms of big data, from geospatial information to digital transaction logs to records of internet activity, can help close the global gender data gap. This report profiles several big data projects that quantify the economic, social, and health status of women and girls.

The first project, described in Section II ("Geospatial Data"), uses satellite imagery to greatly improve the spatial resolution of existing data on girls' stunting, women's literacy, and access to modern contraception in Bangladesh, Haiti, Kenya, Nigeria, and Tanzania. This project develops modeling techniques that use publicly available high-resolution geospatial data to infer similarly high-resolution patterns of social and health phenomena across entire countries. The approach takes advantage of the fact that many types of social and health data are correlated with geospatial phenomena. These relationships can predict social and health outcomes in areas where surveys have not been performed but correlated geospatial data is available. This project generated a series of highly detailed maps that clearly illustrate landscapes of gender inequality (see Figure A).

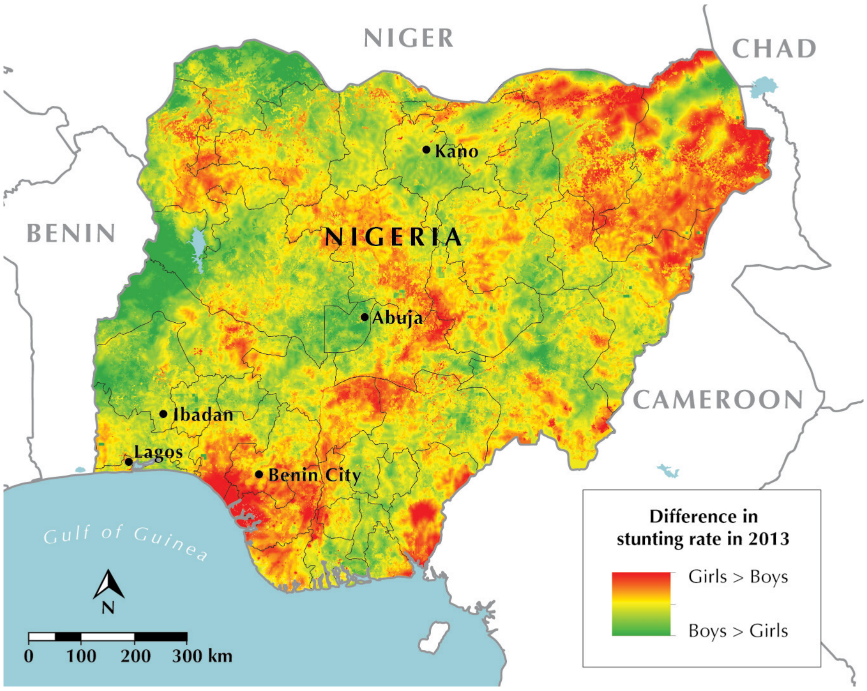


Figure A. Differences in stunting between girls and boys, Nigeria, 2013. Red areas are where girls' stunting is higher than boys' stunting, green where girls' stunting is lower than boys' stunting.

The second project, profiled in Section III ("Digital Exhaust"), utilizes anonymized credit card and cell phone data to describe patterns of women's expenditure and mobility in a major Latin American metropolis. The credit card data includes 10 weeks of transactions from 150,000 users, with associated age, sex, and location information; for a subset of these credit card users, cell phone data is also available. The two types of information together create portraits of economic lifestyles—patterns of behavior that illustrate the needs and priorities

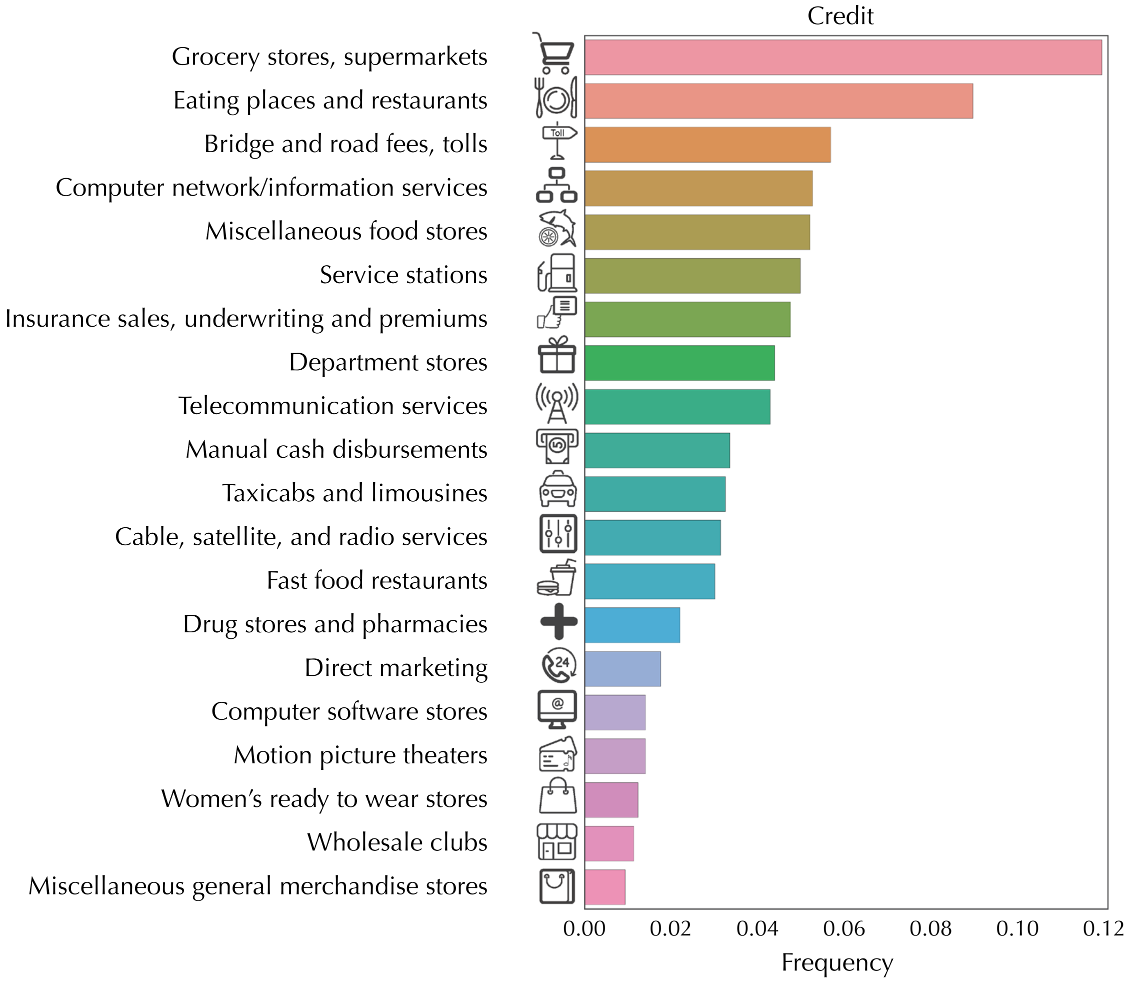


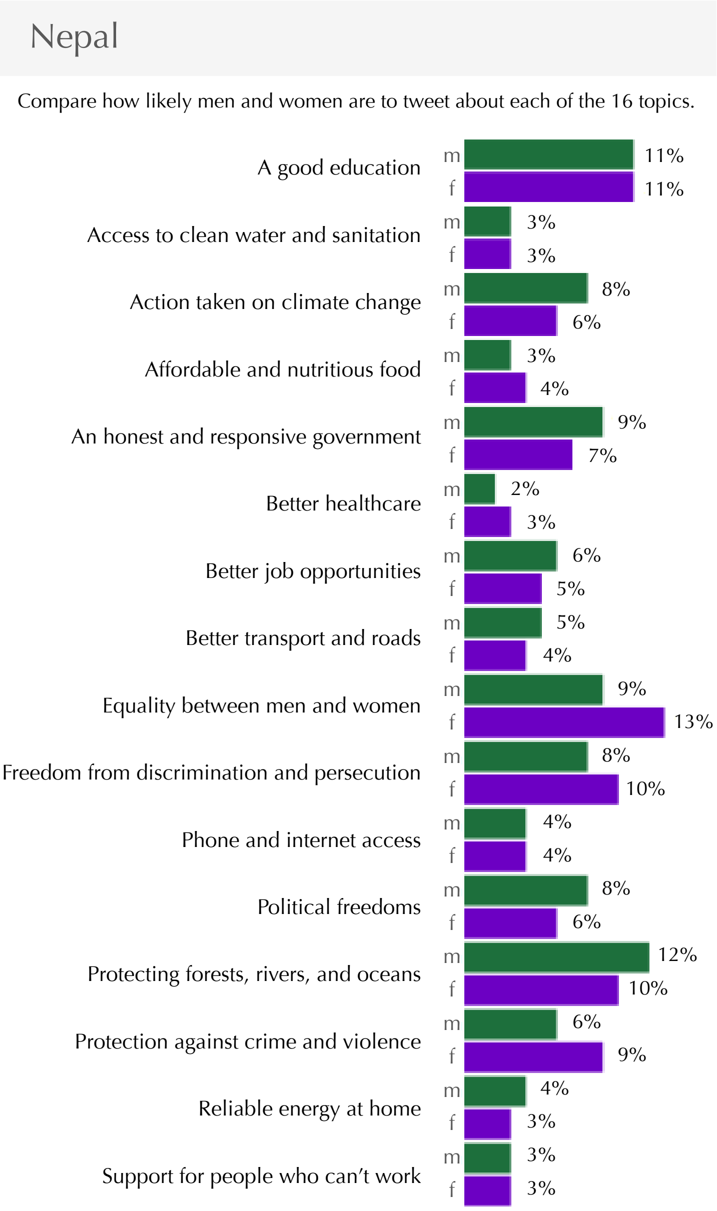
Figure B. Frequency of women's transactions in different expenditure categories, as assessed by credit card data.

of women (see Figure B). Over a longer timeframe, such data could also reveal signals about how women are coping with a wide range of environmental and economic shocks and stressors.

The third and fourth projects, profiled in Section IV ("Internet Activity") concentrate on the expression of ideas and emotions on the social media platform Twitter. The third project develops and prototypes a tool for automatically identifying the sex of Twitter users, and then uses this method to quantify the concerns of women on a wide range of global development issues. The algorithm created in this project automates the process of looking up user's names and pictures from Twitter profiles. Using open source software, the tool analyzes users' names from a built-in database that contains sex information. If name alone is insufficient to infer sex, the tool analyses profile photos using face recognition software. The tool was tested on more than 50 million Twitter accounts across the world to understand the differing priorities of women and men on topics related to sustainable development (see Figure C).

The final project locates signals of depression in a large database of publicly available tweets from women and girls in India, South Africa, the United Kingdom, and the United States. The project uses machine learning techniques to identify genuine self-disclosures of mental illness from nearly 1.5 million social media posts and half a million Twitter users. The method accurately identifies mental illness in 96% of cases. The project also compares modes of linguistic expression and topical content across female and male users. Overall, the findings reveal significant differences in how different sexes express mental health concerns on Twitter. The work suggests two major applications for monitoring and treatment. At the individual level, signals of mental illness could provoke response, either from the user's community or through automated means from the social media platform itself (for example, offering counseling resources). At the population level, mental health trends can be monitored in near real-time, which may be especially useful following recessions, natural disasters, and other shocks.

This report illustrates the potential of big data in filling the global gender data gap. The rise of big data, however, does not mean that traditional sources of data will become less important. On the contrary, the successful implementation of big data approaches requires investment in proven methods of social scientific research, especially for validation and bias correction of big datasets. More broadly, the invisibility of women and girls in national and international data systems is a political, not solely a technical, problem. In the best case, the current "data revolution" will be reimagined as a step towards better "data governance": a process through which novel types of information catalyze the creation of new partnerships to advocate for scientific, policy, and political reforms that include women and girls in all spheres of social and economic life.


Nepal

Compare how likely men and women are to tweet about each of the 16 topics.

Figure C. Trending topics among Twitter users in Nepal, May 2012-July 2015, disaggregated by sex.

# I INTRODUCTION

The term "big data" encompasses diverse types of information, from satellite imagery to cell phone records to internet activity. These forms of data differ in many ways, but all have digital origins, record observations at high frequency, and are massive in size. Such characteristics are invaluable in studying human well-being as it changes over time.

Traditional data systems struggle to quantify trajectories of physical and mental health among a population, especially during and following economic recessions, natural disasters, and other unpredictable shocks. The problem is exacerbated—and present even during periods of relative economic stability—with respect to women and girls, who often work in the informal sector or at home, suffer social constraints on their mobility, and are marginalized in both private and public decision-making. Household surveys, national economic accounts, institutional records, and so on often do not successfully capture the lives of women and girls, especially at the kind of frequency needed to assess changes in economic and health status.

This report profiles groundbreaking approaches to using various kinds of big data to fill the global gender data gap. For each of three major big data categories—geospatial data, digital exhaust, and records of internet activity—we present exemplary research initiatives conducted over the past two years:[a,b]

- In Section II ("Geospatial Data"), researchers at the Flowminder Foundation and WorldPop project use satellite imagery to improve the spatial resolution of existing data on women and girls obtained from Demographic and Health Surveys (DHS) in Bangladesh, Haiti, Kenya, Nigeria, and Tanzania;
- In Section III, ("Digital Exhaust"), researchers at the Massachusetts Institute of Technology (MIT), working with a colleague at United Nations Global Pulse (UNGP), utilize credit card and cell phone data to discern patterns of women's expenditure and mobility in a major Latin American metropolis;
- In Section IV, ("Internet Activity"), we look at two projects concentrating on the expression of ideas and emotions on the social media platform Twitter. In the first, researchers at UNGP and the University of Leiden create an algorithm for automatically identifying the sex of Twitter users, and then use this method to quantify the concerns of women across a wide range of global development issues. In the second project, researchers at Georgia Tech University, supported by colleagues at the University of Leiden and UNGP, locate signals of depression

[a] More detailed reports on each of these projects are available at http://data2x.org/resources

[b] Note that the first-person plural "we" is used throughout this report to refer in different sections to different groups of researchers. The relevant researchers for each section are listed on the inside front cover.

in a large database of publicly available tweets from women and girls in India, South Africa, the United Kingdom, and the United States.

In all cases, the projects yielded important new insights into the lives of women and girls. The sections that follow describe each in detail.

Elizabeth Whelan

# II GEOSPATIAL DATA

## High-resolution Mapping of Sex-Disaggregated Indicators

The big data conversation usually centers on novel forms of data, ignoring a valuable source of information that has been available in the public domain for decades: geospatial data. In recent years, the amount of freely accessible geospatial data, especially satellite imagery, has greatly expanded, spurred by increased investment from government agencies and private businesses. This data is increasingly fine-grained in both time and space: satellite imagery, for example, is now able to record rapid changes in both biophysical phenomena (for example, vegetation, soil cover, and water flows) and human infrastructure (for example, settlements, roads, and light intensity).

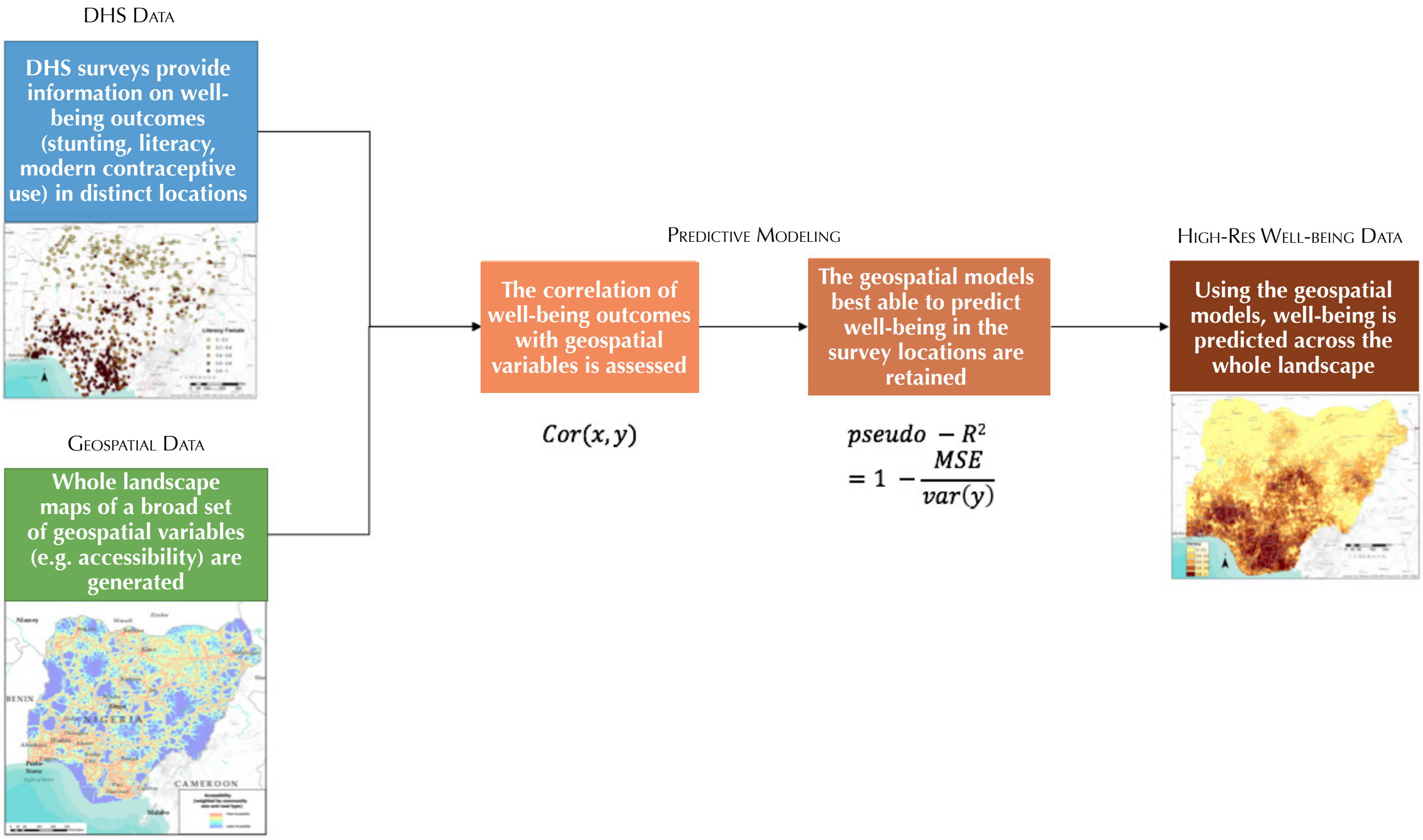
Equally high-resolution data on social and health indicators is critically needed, but still lacking. Human well-being varies considerably within countries, and development indicators assessed at national scales conceal these inequalities. Importantly, the status of women and girls in economically marginalized or geographically isolated communities is often unknown. Although four out of every five countries in the world regularly produce sex-disaggregated statistics at national or provincial scale, this data is not spatially refined enough to support local policymaking or program targeting.

To address this problem, we developed modeling techniques that use high-resolution geospatial data to infer similarly high-resolution patterns of social and health phenomena. This approach takes advantage of the fact that many types of social and health data—for example, child stunting, literacy, and access to modern contraception, the indicators we focus on in this case study—are correlated with geospatial phenomena that can be mapped in great detail across entire countries using satellite imagery. These relationships are then used to predict social and health outcomes in areas where surveys have not been performed.[1] The result is maps that provide entire landscapes of information on indicators of interest. The workflow is illustrated in Figure 1, and the methods more fully explicated in the following pages. We focus especially on outcomes related to girls and women, that is, girls' stunting, women's literacy, and contraceptive access; results for boys' stunting and men's literacy are presented in the accompanying technical report by Bosco et al. (2016).

### METHODS

The DHS program has been a leader in collecting and disseminating survey data on key development indicators in low- and middle-income countries. Large-sample household

Figure 1. Workflow, geospatial modeling of well-being outcomes.



DHS surveys provide information on well-being outcomes (stunting, literacy, modern contraceptive use) in distinct locations

Whole landscape maps of a broad set of geospatial variables (e.g. accessibility) are generated

The correlation of well-being outcomes with geospatial variables is assessed

$$Cor(x,y)$$

The geospatial models best able to predict well-being in the survey locations are retained

$$pseudo - R^2 = 1 - \frac{MSE}{var(y)}$$

Using the geospatial models, well-being is predicted across the whole landscape

data collection of this type, however, is costly, and so surveys are normally designed to be representative at the national or the largest subnational administrative level (typically called states or provinces). These areas often contain millions of people, and statistical assessments at such scales obscure substantial lower-level heterogeneity in social, economic, and health status. However, recent DHS surveys—and, increasingly, other household surveys—provide GPS coordinates for observations or clusters of observations, which enables us to utilize our geospatial modeling approach to improve the spatial resolution of DHS indicators.[2]

In this study, we focus on three countries in Sub-Saharan Africa (Kenya, Nigeria, and Tanzania), one country in South Asia (Bangladesh), and one country from the Western Hemisphere (Haiti); all have a low or medium human development index.[3] We use DHS data from the last several years on child stunting, literacy, and the use of modern contraception (hereafter collectively referred to as "well-being outcomes"), the first two of which are disaggregated by sex; only girls' and women's results are presented in this report.[c,4]

We chose geospatial variables, summarized in Table 1, by combing existing publicly available libraries for those variables that had previously shown correlation with the outcomes.[d,5] We then analyzed the relationship of these variables with stunting, literacy, and contraceptive access at each recorded survey location. The final step used these observed relationships to infer, using high-resolution landscape maps of each geospatial variable, outcomes in all non-survey locations. A continuous landscape of girls' stunting, women's literacy, and access to contraception was thus generated for each country.

[c] Children under age five whose height is considerably (two standard deviations) below the median of the World Health Organization's reference population are considered stunted. People ages 15-49 who attended at least secondary school or could read part of a sentence during the DHS interview are defined as literate. The current use of any modern method of contraception is asked of all women ages 15-49, but in Bangladesh only of ever-married women.

[d] Selecting the optimal subset of geospatial variables is critical for maximizing the ultimate predictive accuracy of a model: too few informative variables and the model will not explain much; too many and the resulting model may explain the observed data extremely well but perform badly when applied to new datasets.

# RESULTS

We first present results of the geospatial variable selection exercise, summarizing overall model performance and then listing the most strongly correlated set of geospatial variables for each indicator in each country. For selected indicators, we show maps comparing DHS survey results with the landscape of values generated by geospatial variables.

Table 1. Geospatial variables used in this study. See Bosco et al. (2016) for extended descriptions and sources.

| Geospatial Variate | Description |
|---|---|
| Accessibility | Likely travel times between two points, a function of distarce and infrastructure |
| Aridity evapotranspiration | Weather station-based interpolation of moisture availability |
| Births | WoldPop-derived number of live births |
| Crop suitability | Rainfed crop suitability given crop/technology mix |
| Distance to conflicts | Nigeria only, between years 2010-13 |
| Distance to health facility | calculated from Open Street Map datasets |
| Distance to roads | Calculated from Open Street Map datasets |
| Distance to schools | Caulculated from Open Street Map datasets |
| Economic productivity | Gross domestic product, calculated with economic data and geospatial correlates of economic activity |
| Elevation | Elevation above sea level |
| Ethnicity | Estimated distribution of ethnic groups |
| Land surface | Land biophysical properties estimated by reflectance |
| Livestock density | Modeled spatial distribution of livestock |
| Nightlights | Light intensity, denoting population density and electrification |
| Population density | Density inferred from settlement and land use patterns |
| Pregnancies | WorldPop-derived number of pregnancies |
| Protected areas | Geospatial conservation on databases |
| Temperature/rainfall | Global climate layers |
| Urban/rural settlements | Estimated distance to settlements, country-specific datasets |
| Vegetation/land cover | Plant cover estimated by surface reflectance |

First, we note that model performance varied greatly across indicators and countries (Figure 2). Models for girls' stunting, for example, were inadequate for all countries except Nigeria. Geospatial variables were generally informative in building models for women's literacy. For modern contraceptive use, models performed strongly in Tanzania and Nigeria. The results suggest that geospatial modeling requires careful investigation of a broad set of variables—even broader than the set explored here—and some outcomes in some countries

Figure 2. Explanatory power of geospatial models, by country and indicator. Country/indicator models with no information shown (stunting in Tanzania and Haiti, literacy in Haiti, contraception in Kenya and Bangladesh) were not modeled, due to lack of sufficient survey indicator data. Boys' stunting and male literacy is not shown.

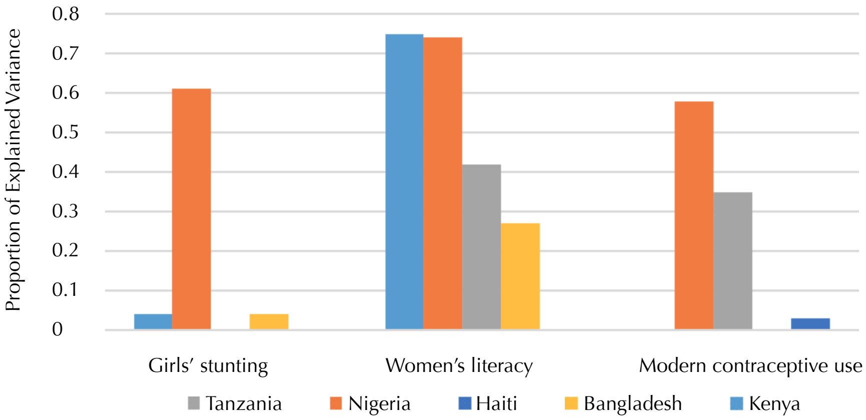


Figure 3. Geospatial correlates of girls' and women's well-being outcomes in the six best-performing models. Shaded box indicates that the variable was included in the final model.

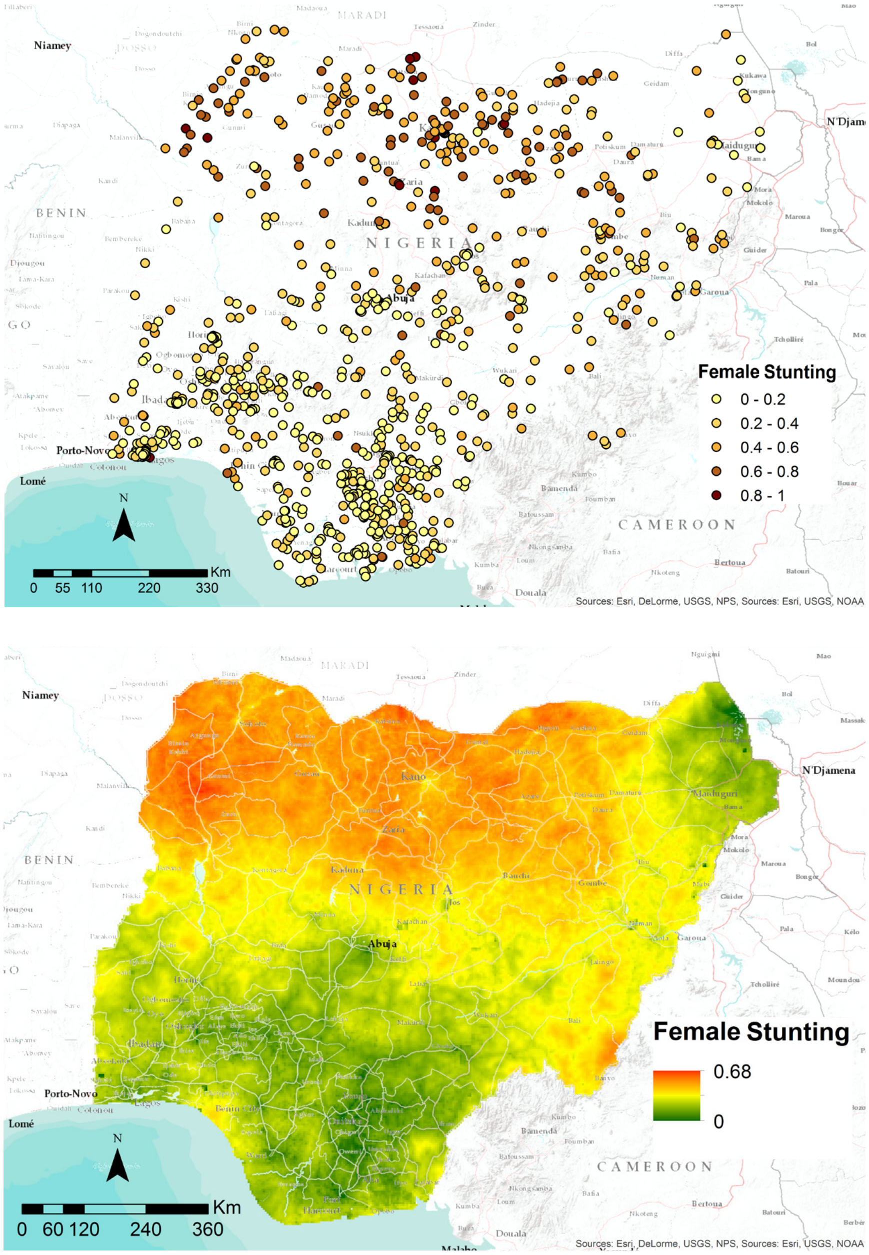
| | Girls' stunting | Female literacy | | | Modern contraceptive use | |
|---|---|---|---|---|---|---|
| | Nigeria | Kenya | Nigeria | Tanzania | Nigeria | Tanzania |
| Accessibility | ■ | ■ | ■ | ■ | ■ | ■ |
| Aridity, evapotranspiration | | ■ | | ■ | ■ | ■ |
| Births | | | | | | |
| Crop suitability | ■ | | | | ■ | |
| Distance to conflicts | ■ | | ■ | | | |
| Distance to health facility | | | | | | |
| Distance to roads | ■ | | | ■ | ■ | ■ |
| Distance to schools | | | | | | |
| Distance to waterways | ■ | | | | ■ | |
| Economic productivity | ■ | | | | ■ | |
| Elevation | ■ | ■ | ■ | | ■ | ■ |
| Ethnicity | | | | | | |
| Land surface | ■ | ■ | ■ | ■ | | ■ |
| Livestock density | ■ | | ■ | | ■ | |
| Nightlights | | ■ | | ■ | ■ | |
| Population density | ■ | | | | ■ | ■ |
| Pregnancies | | | | | | |
| Protected areas | | | | | | |
| Temperature/rainfall | ■ | ■ | | ■ | | ■ |
| Urban/rural settlements | | ■ | ■ | ■ | | ■ |
| Vegetation/land cover | | | | | ■ | ■ |

may not be correlated well to any set of geospatial variables. In the present work, literacy appears to have strong geospatial correlates almost universally, while the performance of girls' stunting and contraceptive use models depends on context.

We also see that the optimum subset of geospatial variables also differs by country and indicator, as shown in Figure 3 for the six best performing models. Accessibility, a general indicator of transport infrastructure quality, is the only geospatial variable that is statistically significant in all six models; elevation and land surface are usually important, and aridity/evapotranspiration, distance to roads, temperature/rainfall, and the distance to urban and rural settlements are significant in most cases. The key message, however, is that the set of optimum geospatial variables depends on context; even the same indicator will have different geospatial correlates in different locations.

We now turn to the core results, presented in a series of maps. Figure 4 makes clear the overall value of the approach. The top panel shows stunting rates for girls in the original DHS survey locations from the Nigeria 2013 dataset. The data appears as a scatter of points distributed unevenly throughout the country; between the survey locations are large areas of space in which stunting prevalence is not known. The bottom panel then shows the girls' stunting landscape in 2013 as generated by the best-performing geospatial model (which includes those variables shaded in the "Girls' stunting/ Nigeria" column of Figure 3). We see a continuous gradient over the entire expanse of the country; not only broad geographic patterns but also differences within sub-regions of Nigeria become more evident.

Figure 4. DHS survey data for girls' stunting in Nigeria in 2013 (top panel) and geospatially predicted landscape of girls' stunting in the same year (bottom panel).



Geospatial modeling also unveils inequalities between girls and boys across the landscape. Figure 5 shows differences in 2013 stunting rates across sexes in Nigeria; positive values (colored in orange/red) indicate areas where girls have higher rates of stunting, and negative values (colored in green) where boys have higher rates (separate results for absolute levels of boys' stunting are available in Bosco et al. 2016). Notably, the areas with higher absolute levels of girls' stunting, as shown in the right panel of Figure 4, are not necessarily the areas of greatest inequality; the northeast, central, and southern urban regions of Nigeria appear to exhibit the largest disadvantage for girls. Overall, the map provides a fine-grained picture of inequality across the entire landscape.

Figure 5. Differences in stunting rate between girls and boys, Nigeria, 2013.

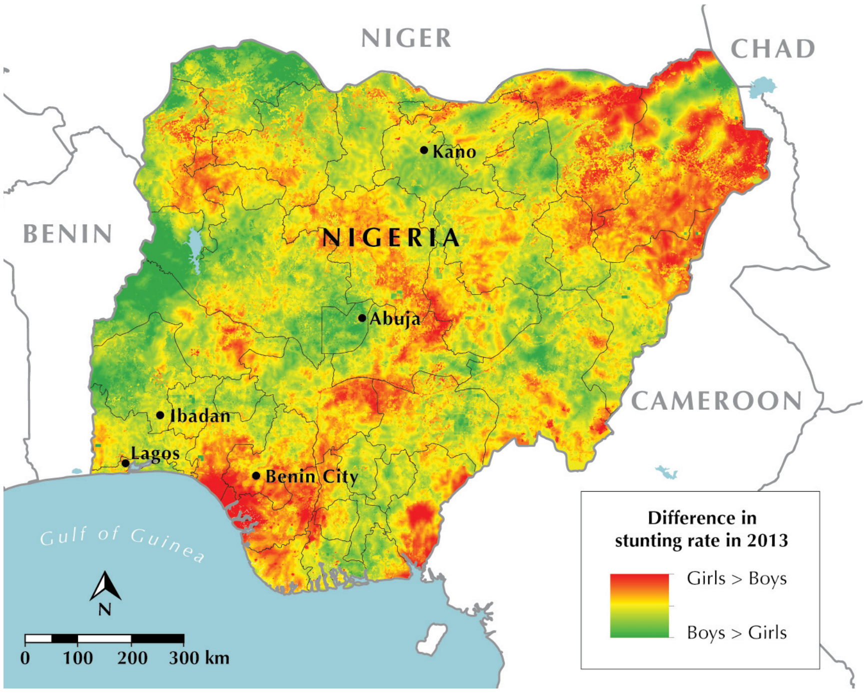


Figure 6 and Figure 7 show similar results for women's literacy in Kenya and modern contraceptive use in Tanzania. Once again, we see that geospatial modeling can transform a limited number of survey data points distributed unevenly across the country into a continuous landscape of information.

This approach does face challenges. Some of the geospatial models we attempted were unable to accurately predict well-being outcomes, as shown earlier by Figure 2. It is possible that a wider set of geospatial covariates would have improved modeling performance. For many variables and locations an exploratory approach is necessary, as little theoretical guidance linking geospatial phenomena with well-being outcomes is available. In addition, the exact nature of the relationships between geospatial phenomena and well-being outcomes—linearity vs. non-linearity, for example—is also unclear.

Overall, however, geospatial modeling of women's and girls' social and health status shows great promise. Some of the maps produced in the present study, especially the maps of all three indicators in Nigeria and the map of women's literacy in Kenya, have sufficiently

Figure 6. Women's literacy in Kenya.

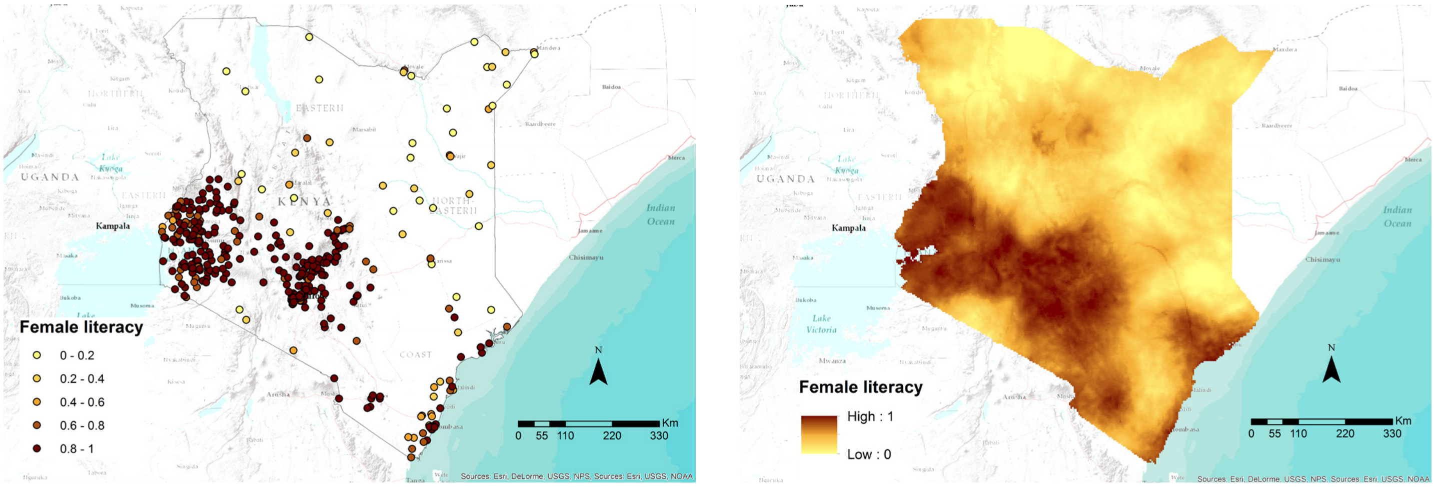
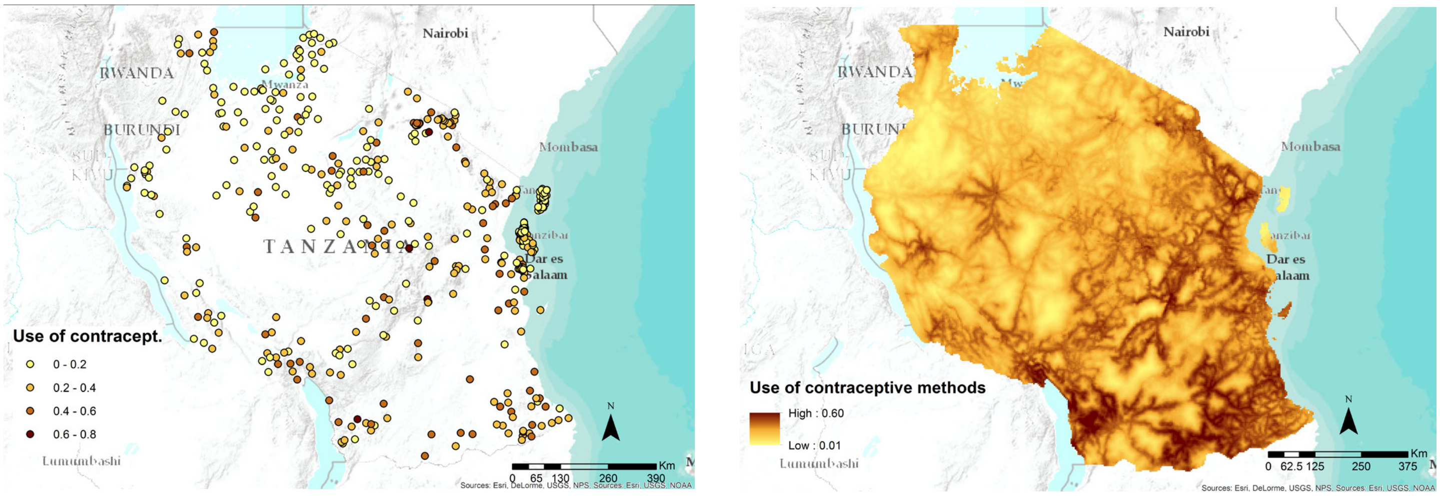


Figure 7. Modern contraceptive use in Tanzania.



low uncertainty to be utilized by policymakers seeking to target interventions at local administrative levels. We have in this section presented results on three indicators from five countries, but the modeling architecture can be extended to other indicators and countries for which DHS has information, as well as to other household surveys containing georeferenced data. The recent expansion of publicly available high-resolution satellite imagery offers a rich bounty of data for exploring geospatial correlations in the many countries where traditional data systems are insufficient to capture the status of women and girls.

# III DIGITAL EXHAUST

## Analyzing Economic Activity with Credit Card and Cell Phone Information

Digital technologies are ubiquitous, and their use leaves traces—records of the goods and services we consume, the places we go, and the people with whom we interact.[6] If information on the sex of the technology user is available, these types of "data exhaust" can offer insight in near-real time about the lives of women and girls.

The following pages describe a project that uses credit card and cell phone data to analyze patterns of economic activity among tens of thousands of women living in one of the most populated cities of Latin America.[e] We use credit card records (CCRs) to examine the expenditure priorities and patterns of mobility of different sexes, income levels, and ages.[7] Call detail records (CDRs), meanwhile, store information about the time, duration, and location of mobile phone calls, as well as the anonymized IDs of the people receiving calls. Past research has used CDRs to analyze social interactions, the laws of human mobility, and the economic welfare of users.[8]

For this project, we obtained over 10 weeks of anonymized individual credit card transactions from 150,000 users, with associated age, sex, and location information. The CCRs include data on the broad types of goods and services purchased, expenditure amounts, and the chronological sequence of transactions. For 10% of these credit card users, call detail record (CDR) data is also available. We used CCRs and CDRs together to describe economic lifestyles—patterns of behavior that illustrate the needs and priorities of individuals. A detailed analysis is done specifically for women in the sample.

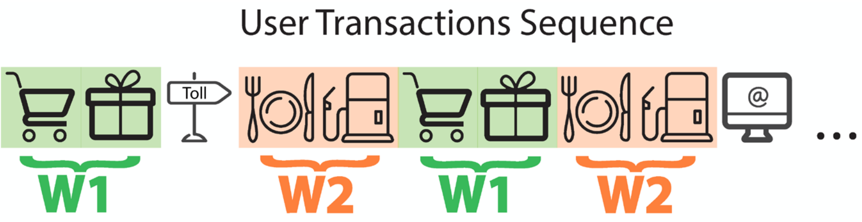
### METHODS

Detecting differences in economic lifestyles is complicated by the fact that only a few categories of purchases dominate spending: most people, regardless of sex, wealth level, or age, spend most of their income on food, transport, and communication, as the "Results" section below discusses. This project thus delves deeper, looking not only at transaction type but also the order in which individuals made purchases, as well as their patterns of geographical mobility. Certain sequences of transactions may be repeated in an individual's purchase history; for example, in panel A of Figure 8, sequence W1 captures grocery store expenditures (the shopping cart icon) followed by department store purchases (the gift box icon), while sequence W2 represents restaurant expenditures (the plate and silverware icon) followed by fuel purchases (the gas pump icon). Both W1 and W2 are repeated twice in this

[e] For contractual and privacy reasons, we cannot make details of the dataset publicly available. Upon request, the authors can provide anonymized data used for a subset of the analyses described below; the code to replicate methods is available upon request.

short transaction history.[f] Such patterns—more than ten thousand of which were detected in the CCR dataset—are the basis for inferring economic lifestyles. As each user's sequences are analyzed, and data on mobility from geocoded transactions added, similarities begin to emerge; people cluster together into distinct economic lifestyles.

Figure 8. Examples of repeated sequences within a transaction history.

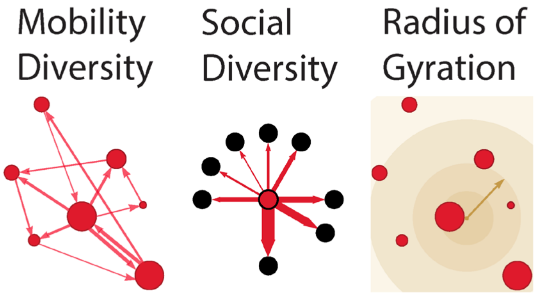


Mobile phone data further enriches our understanding of patterns of economic and social behavior, helping to delineate economic lifestyles even more distinctly. In this project, we focus on three types of information about individuals obtained from CDRs: mobility diversity, social diversity, and the radius of gyration (Figure 9).[9] Mobility diversity—how evenly an individual splits travel time across the various locations he/she visits—can be constructed using location information from CDRs, gathered by the towers through which cell phone signals pass. Social diversity quantifies how evenly an individual splits airtime across all people in his/her calling network. Finally, the radius of gyration defines the physical area where the user is most likely to be found.

## RESULTS

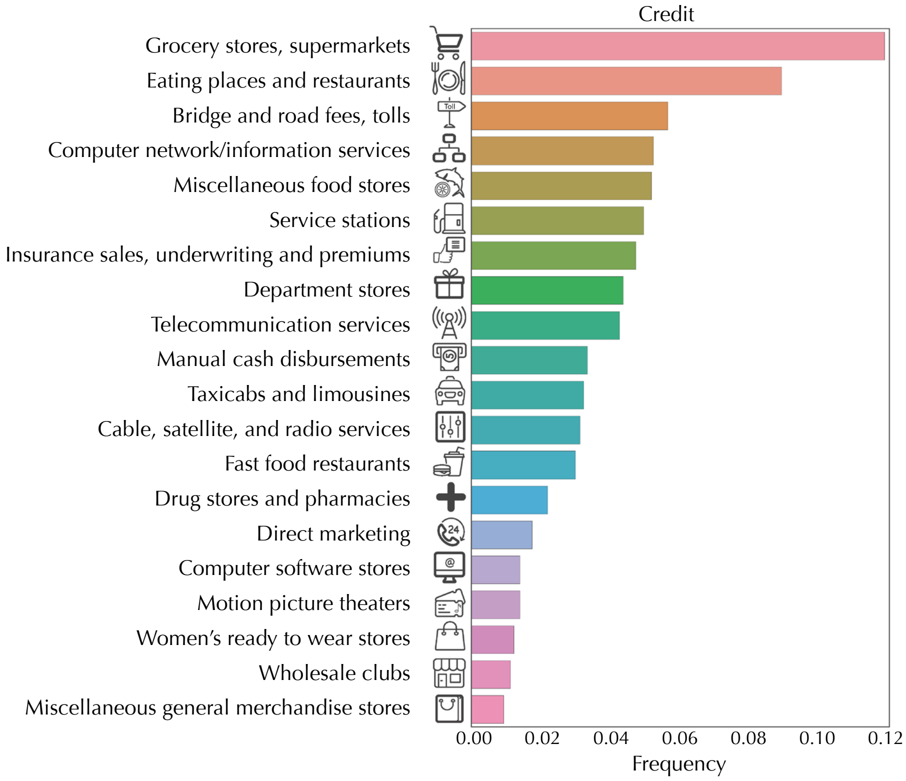
We find that expenditures on food are the most important transactions for women, with over a quarter of transactions in grocery stores/supermarkets, eating places/restaurants, and miscellaneous food stores (Figure 10). A closer look shows expenditure patterns across sexes, ages, and income levels (Figure 11); notice that expenditure across sexes shows strong differences in some categories, while differences across income levels are minor. Women have more transactions than men with respect to grocery stores/supermarkets, insurance-related expenses, and department stores, while the opposite is true for restaurants and transport-related

Figure 9. Features of individual behavior obtained from a combination of call detail records and credit card records.



[f] Sequences can be as short as two transactions or much longer. Longer sequences will occur less frequently in the transaction history.

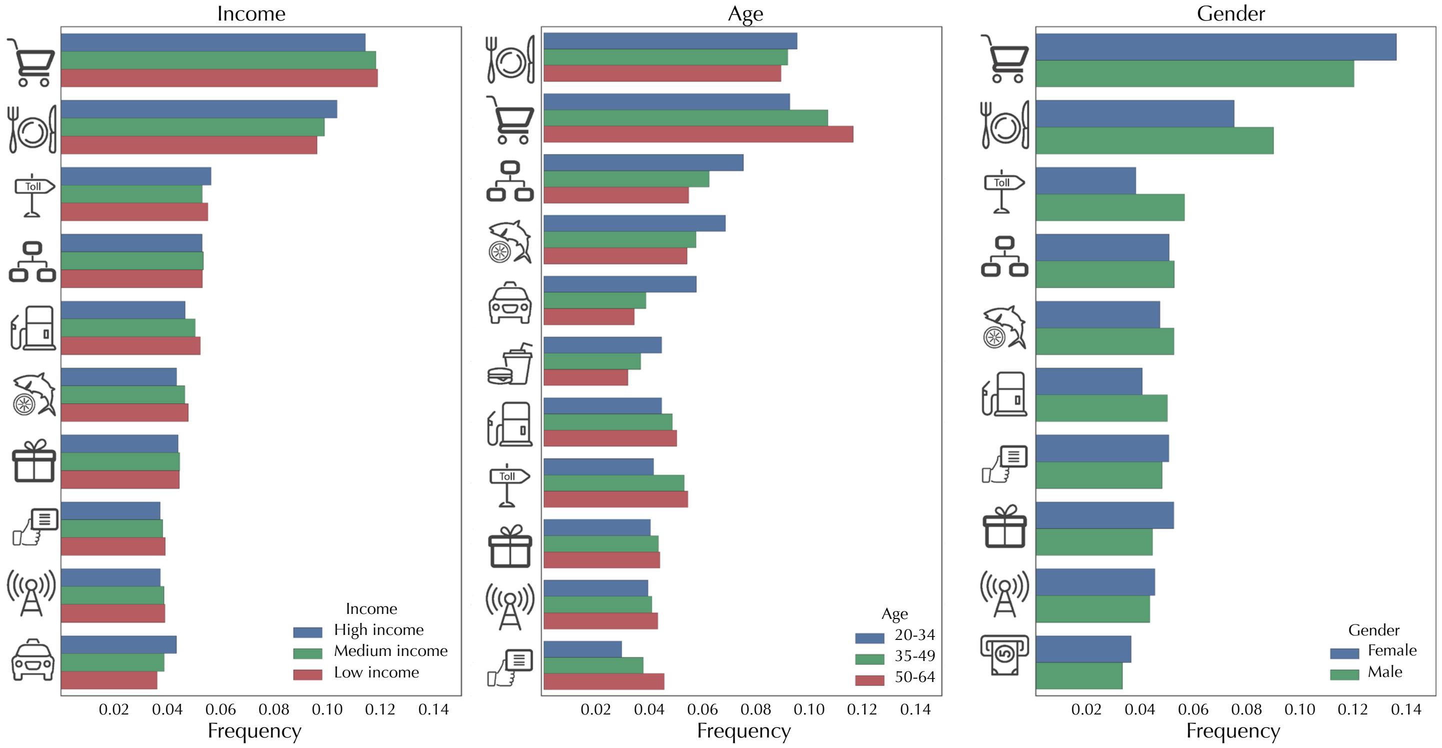
Figure 10. Frequency of women's transactions in each expenditure category. For example, 12% of all transactions in the CCR dataset were from grocery stores and supermarkets.



expenses. In general, women report less total expenditure per capita than men, indicating that they either have less access to economic resources in general and/or use credit cards less frequently. Such patterns are likely to vary based on the nature of the economy and the prevailing economic circumstances; this analysis will be especially relevant in the wake of economic and environmental shocks, when little real-time, sex-disaggregated data is available.

Using the combination of credit card and cell phone data, we identified seven economic lifestyle clusters among women in the dataset (Figure 12). One of the

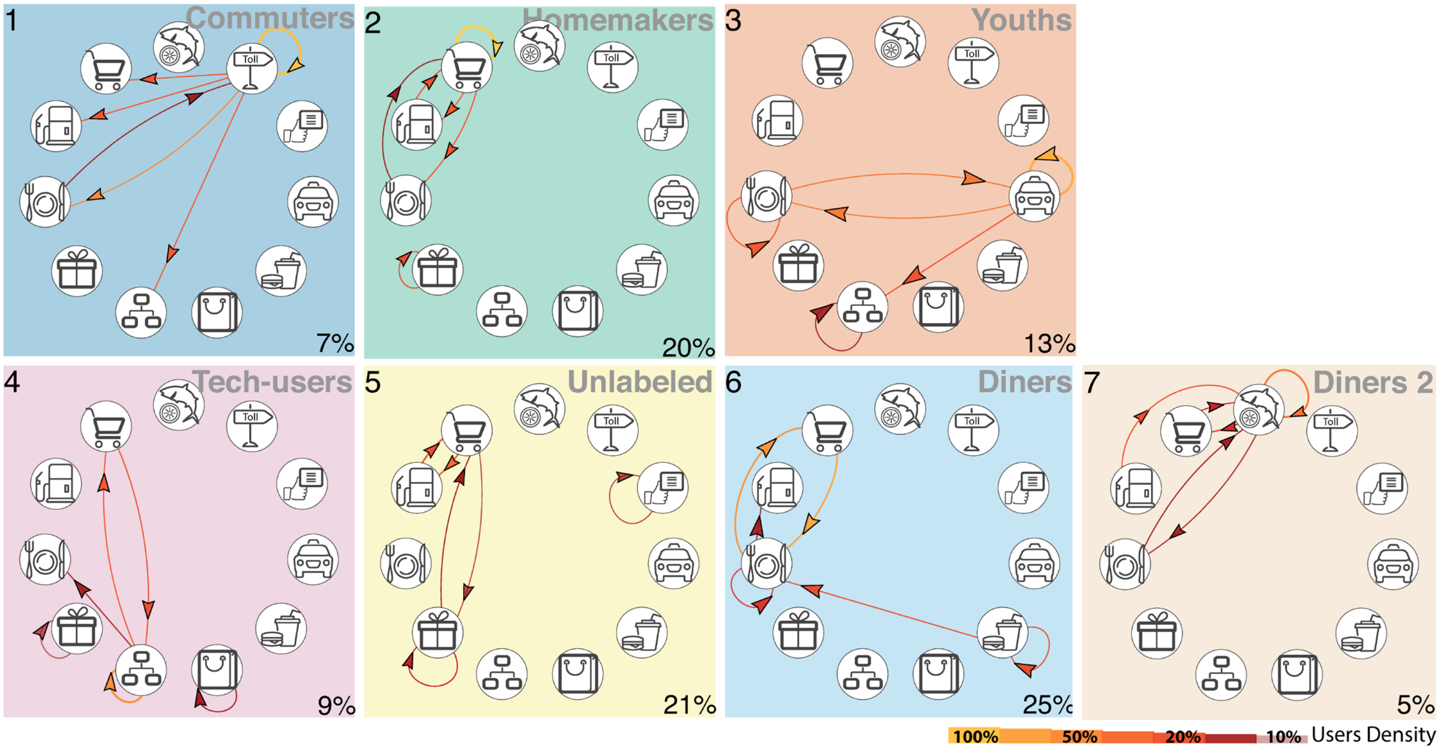
Figure 11. Frequency of transactions by income, age, and gender in selected categories.



clusters, however, does not exhibit a strong pattern of sequences, and is best left unlabeled (cluster 5 in the figure). The transaction sequences within each of the other clusters is dominated by a single type of expenditure, and we use this type to help label the clusters as commuters, homemakers, youth tech-users, and diners (of which there are two types, as discussed further below).

*Commuters*' primary transaction is toll fees, and their mobility metrics suggest that they travel long distances frequently. The core transaction of *homemakers* is grocery stores; they are less mobile, have less social network diversity, and spend less using credit cards. Women are overrepresented in this group, suggesting that women in this urban area perform traditional domestic roles. *Youth* have taxis as their primary expenditure and live close to the city center. *Tech-users* are of similar age as youths, but computer and information services are their most important transaction and they have greater diversity in their social contacts

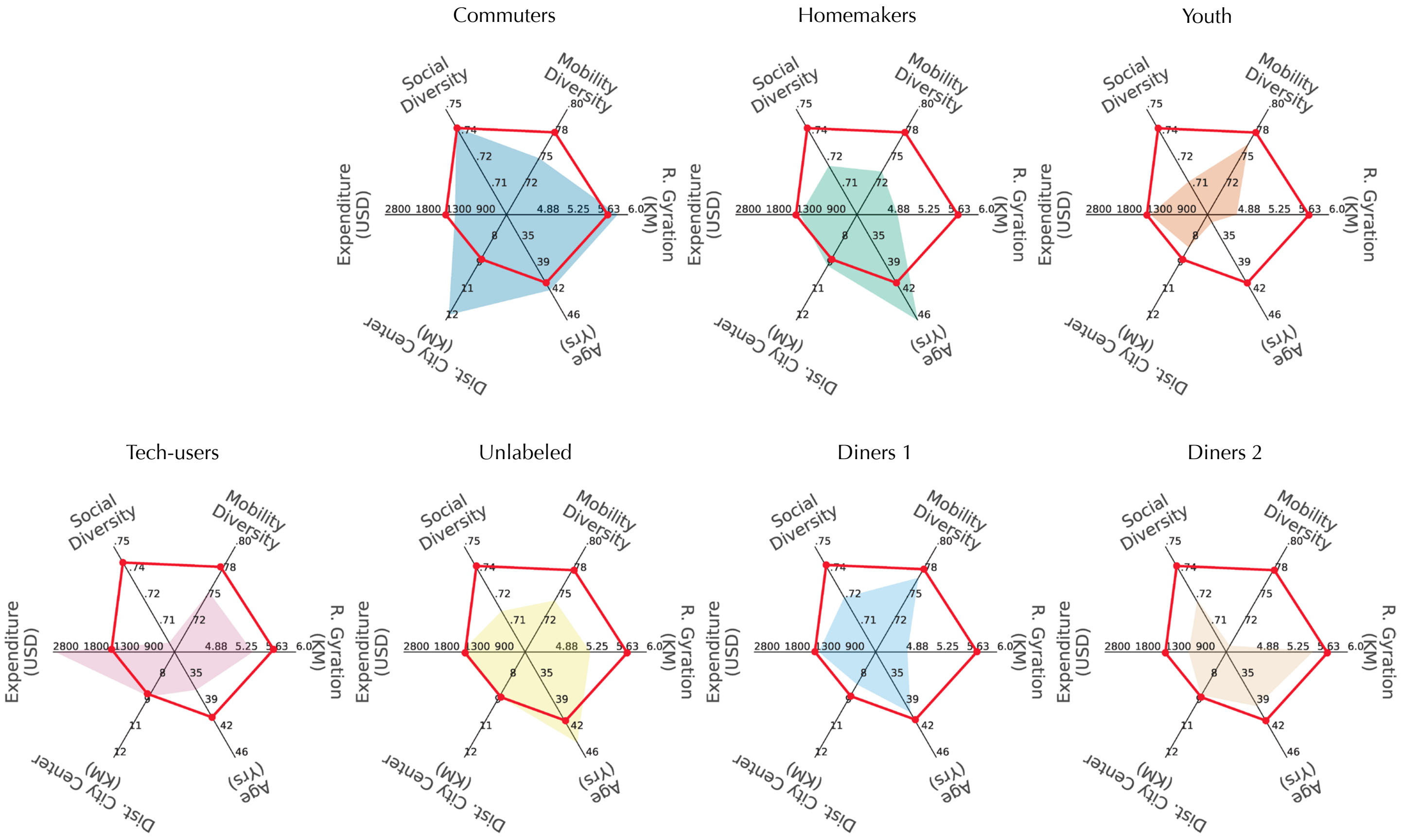
Figure 12. Economic lifestyles of women in the dataset. Arrows indicate frequent transaction sequences; color bar indicates how common that sequence is among the people in the category. For example, commuters are likely to pay tolls followed by expenditure on more tolls, restaurants, groceries, fuel, computer and information services, and telecommunication services.



and mobility networks, as well as higher spending overall. The first *diners* group has high mobility diversity, high expenditures, and restaurants as the primary transaction; the second *diners* group has lower mobility diversity and lower expenditure, and miscellaneous food stores are the core expenditure.

The identification of economic lifestyle clusters is a vital input for policy formulation. Subgroups within a population have distinct social and economic needs. For example, commuters may be hit hardest by fuel price increases, and the creation of inexpensive and efficient public transport systems may be an important investment in urban areas, especially those where low-income residential areas and job opportunities are not in proximity. The analysis above shows also that groups like homemakers and youth have distinct patterns of expenditure from other segments of the population. Information of this kind facilitates

Figure 13. Characteristics of women users in each group (shaded polygons) vs. the entire sample of men (red line).



an analysis of the relative costs and benefits of policies to improve (say) access to affordable food, information services, or transport. The segregation of the diners illustrates that subgroups do not access food in the same way, and food policies—nutrition subsidies for at-risk groups, programs to incentivize low-cost grocery stores in "food desert" areas, and so on—must be tailored to the needs specific to each economic lifestyle.

We also found that women's economic lifestyle clusters differed in important ways from men's. The diagrams in Figure 13 show median scores for social diversity, mobility diversity, age, distance from the city center, and total expenditure. The scores for women in each economic lifestyle group are represented by the shaded polygons; the average scores for all men in the sample are represented by the red line.

There are clear differences between women in each group and men in the overall sample. Men have much greater mobility diversity than women commuters, and tend to live much closer to the city center, indicating that men may have better access to economic opportunities. Women homemakers tend to be much less social and have reduced mobility compared to men, again with implications on market and non-market activity. Female youth, tech-users (especially), and diners also have much less social diversity than men; in this urban area, men appear to have greater numbers of social connections, while women have smaller networks. Young women also have a much smaller radius of movement, again pointing to a constrained economic and social world. Female tech-users are the only lifestyle group for whom expenditure is significantly greater than that of men in general, indicating that tech may be a sector in which women are finding more remunerative job opportunities. Two distinct types of female diners are present: one group with relatively high mobility diversity and expenditure but low radius of gyration (e.g., many different restaurants in a localized area; Diners 1), and another with low mobility diversity and expenditure but higher radius of gyration (e.g., people needing to travel relatively long distances to find cheaper food sources; Diners 2).

The current project does have some limitations—most notably, a bias towards inclusion of relatively better-off individuals with access to credit cards and cell phones, although penetration of the latter into even isolated rural areas of the developing world is increasing. We did examine whether CCR users were representative of the general population, especially given that less than a quarter of the population in the research area use credit cards. The monthly expenditure of the CCR users was high relative to wages in most of the neighborhoods included in the project, confirming that (within neighborhoods) the sample is biased towards wealthier individuals—although, evaluating the sample as a whole, users from all income levels are well-represented.

Despite the potential for bias, the project does demonstrate that a combination of credit card and cell phone data can provide detailed insights into women's economic behavior. This research could also serve as a basis for further work applying the proposed methodologies to other developing countries where mobile money, in addition to credit cards, is commonly used. The results above are drawn from a ten-week period in which economic conditions were relatively stable. Over a longer timeframe, our approach could also reveal signals about how women are coping with a wide range of shocks and stressors: environmental disasters, recessions, macroeconomic policy shifts, and so on. For example, reduced mobility among a low-expenditure group could signal that poorer economic classes are unable to afford the transport costs necessary for commuting and accessing markets and government services. Such early warning information would be valuable in designing and managing effective social protection systems.

Elizabeth Whelan

# IV INTERNET ACTIVITY

## Sex-Disaggregation of Social Media Posts

Social media can help monitor public perceptions and measure global development priorities and impact. It can also provide insights into the differences and inequalities between people of different income, sex, age, race, ethnicity, migratory status, disability status, geographic location, and other characteristics.

Sex disaggregation, in particular, can play an important role in providing information about the disparities between women and men. However, data from open social media channels such as Twitter may not indicate a person's sex. In this project, United Nations Global Pulse (UNGP) collaborated with the University of Leiden to develop and test a tool to infer the sex of Twitter users. The tool automates the process of looking up public information from Twitter profiles, especially the user's name and profile picture. Using open source software, the tool analyzes users' names from a built-in database of predefined names, built from sources such as official statistics that contain sex information. For cases in which name alone may not be enough to discern sex, the tool analyzes profile photos using face recognition software. The tool was applied to more than 50 million Twitter accounts from around the world to understand the different concerns and priorities of women and men on topics related to sustainable development.

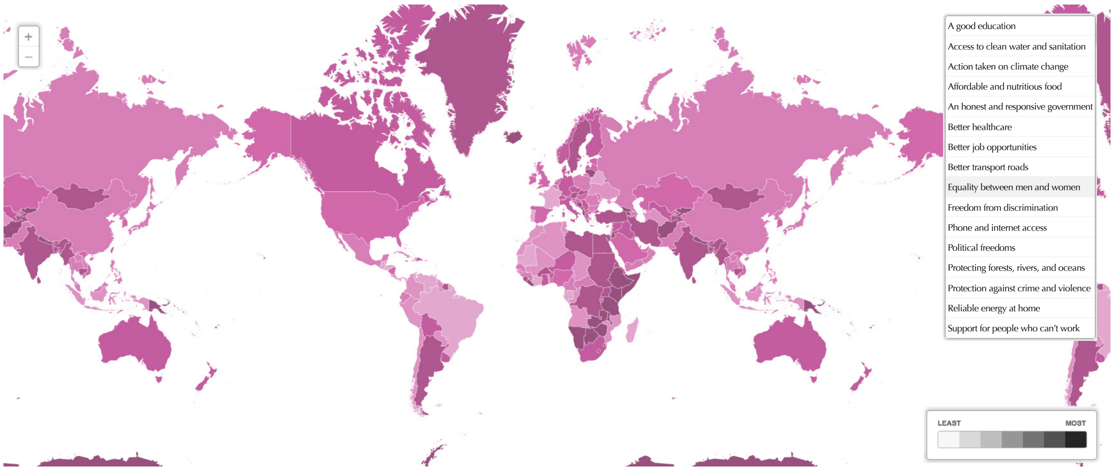
### METHODS

A key objective in the development of this tool was to ensure that the approach could be applicable at a global scale and across different languages. The tool disaggregates social media posts based on several automated classification methods in a "waterfall" approach—starting with the classifiers with the highest overall success rate and, when results are unknown or indecisive, moving on to another classifier.

The name classifier performed best in this study: a user's name was compared to a pre-existing "name dictionary" showing whether the name was more likely to indicate a woman or a man. These results could be further improved by using country-specific name dictionaries, although this would require a more complex process of first determining the home country of a specific user. Since exact location of the user is often omitted in tweets, the approach would adopt a separate script for classifying location, after which a country-specific dictionary could be used.[g] If a country-specific dictionary were absent or results were indecisive,

[g] See the technical report pertaining to the subsequent section ("Social Media Expression as Signaling Mental Health States") for another approach to country inference.

Figure 14. Global Pulse post-2015 tweets dashboard, conversations about equality between men and women.


+
−
A good education
Access to clean water and sanitation
Action taken on climate change
Affordable and nutritious food
An honest and responsive government
Better healthcare
Better job opportunities
Better transport roads
Equality between men and women
Freedom from discrimination
Phone and internet access
Political freedoms
Protecting forests, rivers, and oceans
Protection against crime and violence
Reliable energy at home
Support for people who can't work
LEAST
MOST

disaggregation would take place by relating a user's name either to language-specific dictionaries or an aggregated set of several dictionaries.[h]

In addition to name classification, image recognition of a user's profile picture demonstrated good results for sex disaggregation. For this project, the script classified a user's profile picture with the free-to-use tool Face++. However, if multiple persons are identified in the same photo, the results can be inconclusive. For the purposes of this prototype tool, the algorithm chose the face for which sex is most reliably identifiable.

The script for sex disaggregation of social media accounts is open-source and readily available.[i] For illustrative purposes, a demo version of the tool itself has been made available online for inferring the sex of a person based on their Twitter user name, first name, or an image URL.

## RESULTS

To test the accuracy of the waterfall method—first deploying the name classifier, followed by image recognition—a public website was created. The website allows users to manually determine whether a certain Twitter account is male or female. The accuracy of the hybrid classification approach (the waterfall combination of name and image recognition) was

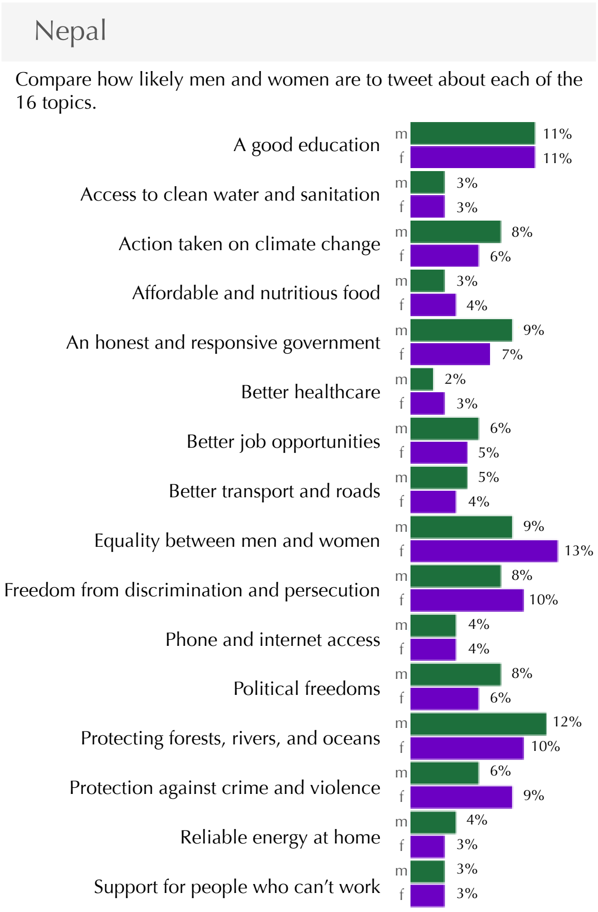
[h] The name classification process and script used in this tool built upon the code of "Gender Computer" developed by TU Eindhoven. The code of Gender Computer can be accessed here: https://github.com/tue-mdse/genderComputer. For the overall script, we updated several of the dictionaries with new names and included additional country specific dictionaries.

[i] GitHub repository: https://github.com/LU-C4i/gender_classifier

compared with crowdsourced verification mechanism, the results of which were assumed to be correct. The automated classification approach accurately determined sex in 74% of cases, a reasonable result for a tool in its initial stages of development.[j] As described above, future work using more context-specific dictionaries should improve accuracy.

The tool could be used for any study of tweets and other types of social media expression wherein the name and/or profile picture of users are available. For example, Global Pulse used the sex-disaggregation tool to improve an existing real-time online dashboard showing the volume of tweets around priority topics related to sustainable development, including gender equality (Figure 14). By filtering through 500 million daily tweets from over 50 million accounts for 25,000 keywords relevant to global development topics, this interactive dashboard showed which countries tweeted most about which topics between May 2012 and July 2015.[k]

Figure 15. Trending topics of Twitter users in Nepal, May 2012-July 2015, disaggregated by sex.



To further refine the dashboard, the gender classification script was run over the entire dataset. Once disaggregated by sex, the dashboard revealed new insights, highlighting the different concerns and priorities of women and men. For example, in Nepal, the sex-disaggregated data showed that women tweeted most on "equality between men and women" (Figure 15). In comparison, men discussed most about "protecting forests, rivers and oceans." In the second quarter of 2015, prompted by the earthquake that hit Nepal on April 25th, discussions were dominated by "support for people who cannot work"—a topic rarely mentioned previously—and "an honest and responsive government." The above topics were widely mentioned by both men and women.

This project faced several obstacles. Because of the anonymity standards of Twitter, identifying authentic user data is not always possible. Context-specific name databases and other tools—for example, profile and linguistic style choices—can help improve prediction of sex; improvement on the current accuracy rate of 74% is almost certainly possible. Overall, however, the approach developed here advances sex-disaggregated analysis of social media, and by doing so provides a window into large databases of ideas and opinions.

[j] With respect to privacy, the methodology uses publicly available data from Twitter profiles. Moreover, only publicly revealed gender markers such as the name and profile pictures of users were applied to building the tool. Users for whom the name and profile picture were insufficient to allow the classifier to detect sex were categorized as unknown.

[k] The project was initially developed by Global Pulse in collaboration with the UN Millennium Campaign and DataSift: http://post2015.unglobalpulse.net/

# Social Media Expression as Signaling Mental Health States

Several unique social and psychological characteristics are implicated in the mental health challenges experienced by women and girls.[10] In addition, poverty, inequality, and cultural expectations may heighten the risk of mental illness among women and girls.[11] Most of the publicly available data on mental health burden, however, comes from massive and infrequent exercises that rarely include sex-disaggregated information, especially in the developing world.[12] Methods of mental health assessment are also inconsistent across countries.[13] Overlooking sex-based differences can have drastic consequences, including misdiagnosis, inappropriate treatment, and constrained help-seeking.[14] More high-frequency data on mental illness and better understanding of the ways in which women and girls express their mental health concerns is thus needed.[15]

Research in recent years has proposed that social media data can help understand patterns of mental health in complement to more traditional assessments.[16] Here we present a gender-based, cross-cultural quantitative examination of mental health content shared on the social media platform Twitter. Using a dataset of half a million Twitter users and nearly 1.5 million posts from four countries, India, South Africa, the United Kingdom, and the United States, we employed machine learning techniques to identify genuine self-disclosures of mental illness from public social media posts. Comparison of these posts with content shared on online mental health support communities, as well as consultation with mental health professionals, suggests that the method accurately identifies mental illness in nearly all cases. We also compare modes of linguistic expression and topical content across female and male users. Overall, the findings reveal significant differences in how different sexes and cultures express mental health concerns on Twitter, and suggest that unobtrusively gathered social media data can serve as an important source of mental health information.

## METHODS

The various steps of the methodology are described in the paragraphs that follow. First, we filtered English-language Twitter posts from March 2015 to create a sample of mental illness disclosures ("MID users") containing any of the key phrases listed in Table 2. These phrases, denoting current experience with mental illness, were collated through reference to prior work as well as consultation with a practicing psychiatrist.[17] A control data sample of posts over the same period ("CTL users") was also created; none of these posts contained

Table 2. Key phrases to filter for signals of mental health concerns.

| | | |
|---|---|---|
| I want to die | I want to end my life | I want to suicide |
| I thought of suicide | I am depressed | I [*] diagnosed [*] depression |
| I attempted suicide | I [have/had] depression | Killing myself |
| I [*] thinking of suicide | I [*] diagnosed [*] mental illness | I tried to suicide |
| I [have/had] mental illness | Ending my life | |

the key phrases in Table 2. Sex and country information were then inferred for each post using an automated method.[1]

The key phrases in Table 2, however, may not indicate genuine disclosure of mental illness; for example, "when I have to wake up at 6am, I feel like killing myself" does not indicate suicidal intent. To eliminate such misleading posts, a machine learning method was used to compare the language of each Twitter user with the language of posts made by people who self-identify as suffering from mental illness on the Reddit sub-communities r/depression, r/mentalhealth, and r/SuicideWatch.[18] A similar process was used to validate the control dataset, but evaluated dissimilarity to the Reddit sample instead. A final qualitative validation exercise was also carried out: a licensed psychiatrist and two researchers experienced in mental health/social media research evaluated a subsample of 100 mental health disclosures, each from a different user. Overall, the machine learning approach in this project was 96% accurate in identifying genuine disclosures of mental health concerns.

Deeper analysis of the social media content followed. We developed linguistic measures (how users express themselves) and a topic model (what users are talking about) to quantify the differences between how female and male users disclosed their mental health concerns. Linguistic measures were divided into three categories—affective attributes, cognitive attributes, and linguistic style—and subtypes within these categories, drawn from previous psycholinguistic work (Table 3).[19]

With respect to affective attributes, the project considered positive affect (PA), negative affect (NA), and four other more specific measures of emotional expression: anger, anxiety, sadness, and swearing. Cognitive measures were divided into cognition and perception, which together evaluate cognitive complexity and emotional stability.[20] Finally, four measures of linguistic style were considered: lexical density, temporal measures, social/personal concerns, and interpersonal awareness/focus. These measures of linguistic style

[1] Because Twitter does not allow individuals to self-report their sex and location information is often inaccurate, sex and country inference is necessary. For sex inference, we matched the self-reported name string in Twitter profile names with name databases from government and other sources. For country inference, we corrected location names using standard techniques and matched locations to various large geographic databases.

indicate one's underlying psychological processes (lexical density), personality (temporal references), social support and connectivity (social/personal concerns), and awareness of one's surroundings and environment (interpersonal focus). Prior work suggests that these cues are valuable in understanding mental health, including in social media expression.[21]

Table 3. Types of linguistic measures used.

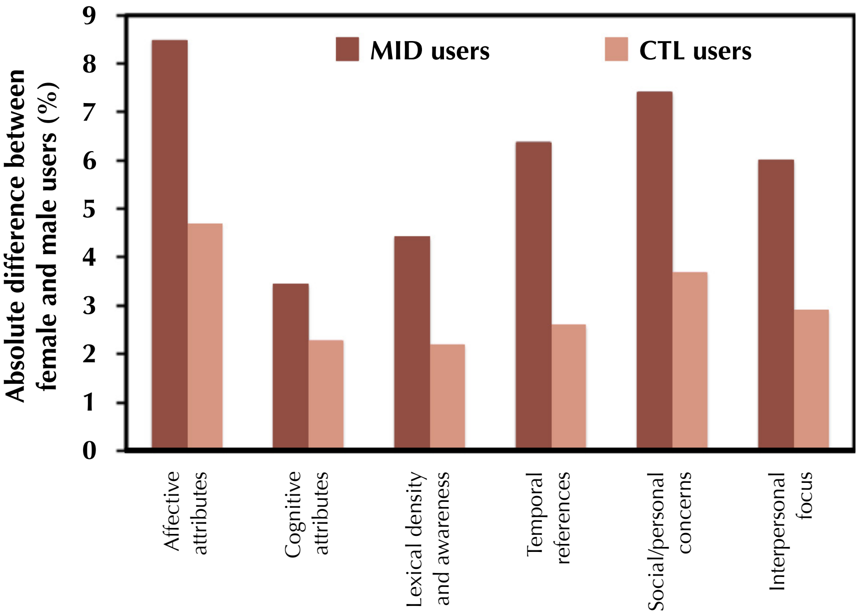
| Attribute Category | Subtype | Description/example |
|---|---|---|
| **Affective** | Positive affect | Expressions denoting positive moods (e.g. joy, energy, alertness) |
| | Negative affect | Expressions denoting negative moods (e.g. sadness, fear, lethargy) |
| | Anger | Expressions of anger |
| | Anxiety | Expressions of anxiety |
| | Sadness | Expressions of sadness |
| | Swearing | Use of swear words, denoting frustration, intensity of reaction |
| **Cognitive** | Cognition | Expression that reflects thought, possibly independent of external stimuli |
| | Perception | Expression that reflects sensory input (e.g. information received by seeing, hearing, feeling) |
| **Linguistic style** | Lexical density | Nouns, adjectives, adverbs, and verbs as a proportion of all words |
| | Temporal measures | Use of past, present, and future tenses |
| | Social/personal concerns | Words pertaining to social engagement or self-engagement (e.g. words about family, friends, social work, health, etc.) |
| | Interpersonal awareness/focus | Use of 1st person singular, 1st person plural, 2nd person, and 3rd person pronouns |

## RESULTS

As noted above, the machine learning approach accurately identified genuine mental health disclosures in nearly all social media posts we examined. We also observed considerable differences in the linguistic content and topical focus of Twitter posts of female and male users, as well as across cultures.[m] First, when affective and cognitive attributes are aggregated into single categories, we see that females generally show higher scores in all linguistic measures (Figure 16). This suggests a generally higher level of psycholinguistic expressiveness on social media by women and girls, a promising result for the objective of using such platforms to identify trends in mental health. Second, we see that the differences are even more pronounced in the MID (mental illness disclosure) user sample than in the CTL (control) sample.

[m] We focus on sex differences here in this summary. Please refer to De Choudhury et al. (2016) for a discussion of cultural differences.

Figure 16. Differences in linguistic measures between female and male users, disaggregated by mental illness disclosure (MID) and control sample (CTL). Positive values indicate higher scores for female users.



We can further break down the attribute categories. Female users in the MID sample show 15.4% higher sadness and 10.7% higher anxiety; prior literature indicates that expression of these emotions is associated with depression, mental instability, and feelings of helplessness, loneliness, and restlessness. However, female users also tend to use 7.1% more positive affect in their content, perhaps to demonstrate a positive outlook publicly despite the mental health challenges they are facing. Male users, on the other hand, express 2.6% more negative affect overall, including 5.3% higher anger and 9.5% more expressions with swearing. Females express fewer cognitive attributes on social media than do males. Lower usage of words that denote certainty, for example, may demonstrate heightened emotional instability. These differences in cognitive expression are not pronounced in the control sample, however, suggesting that experience of mental illness, not intrinsic differences between the sexes, is responsible for the observed gap.

> “#depression has invaded my peace and #anxiety has exhausted my thoughts. Pain isn’t always physical
>
> – female user

> “why am I even here... No one needs or wants me... I’m useless
>
> – female user

We turn now to social/personal concerns and interpersonal focus, both subtypes within the linguistic style category. Male MID users display an 8.1% lower sense of achievement than women and girls, a known signal of reduced self-esteem.[22] Female MID users, meanwhile, express 6.0% greater concern about their health and 2.7% greater concern about their body, which may indicate a greater self-awareness about their health or, alternatively, more fixation with social perceptions about their appearance. Another interesting finding is that male MID users exhibit lower use of words having to do with social concerns, friends, or family. Their female peers may be using such language more frequently in their Twitter posts to explicitly seek help from their social networks. The interpersonal focus metrics

> “Over the past 2 years I have been hit with physical and mental pain. The pain is real. It is still there.
>
> – female user

also reflect these patterns. Male MID users use first person pronouns 10.2% more than female MID users, but 3.0% less second person and 3.4% less third person pronouns, indicating that males tend to be less interactive. Once again, these differences are much less pronounced in the CTL sample. Mental illness appears to amplify differences between female and male expression on social media.

> “Hard to really feel sick with this support group. #family
>
> – female user

> “I miss having someone, a friend to talk to all night
>
> – female user

Our analysis of topics—what users were talking about—confirmed the patterns observed in the linguistic measures. We found that two topics were more likely to appear in male MID posts than in female MID posts. The first topic related to negative thoughts and hopelessness, and the second to detachment from the social realm and a hesitation to seek help. Female MID users expressed a positive outlook in coping with mental health challenges, as well as a desire for disclosure and help-seeking, to a much greater degree than their male peers. Women were also much more likely to share personal experiences around mental illness and engage in self-assessments.

> “Sometimes I wonder if anyone still looks out for me. I am a mess that nobody wants to clean up. I'm a wreck
>
> – male user

> “If I were going to kill myself, I wouldn't tell anyone. If I'm already invisible, why see me to favor your own self-righteousness?
>
> – male user

This work provides some of the first detailed insights on patterns of mental health among girls and women using public social media posts. We found that female users expressed higher sadness and anxiety, but lower anger and negative affect than male counterparts. These observations align with prior work in social psychology.[23] Female mental health disclosers in our dataset also expressed greater social and familial concerns than did males. The literature indicates that women tend to rely more on the social network of family and community, whereas men exhibit a relative orientation towards public stoicism.[24] The topic analysis confirmed this pattern. Although much work remains to link these differences to specific mental health conditions and severity of illness, this data suggests that such research would indeed be fruitful.

> “you're afraid to tell people how you feel because you fear rejection, so you bury it deep inside yourself where it only destroys you more
>
> – female user

> “I used to hurt myself because it was the only pain I could control.
>
> – female user

Adam Cohn

The present analysis has important limitations. The phrases used to filter for mental health concerns are not an exhaustive list of possible signals of depression, anxiety, or other states. In addition, our sample is not representative of the general population; it captures Twitter users, which are likely to be more affluent, more technologically skilled, and more willing to express themselves publicly about mental health issues than the population at large. Inferring overall mental health disorder prevalence rates from social media will clearly require validation surveys that precisely quantify bias.

Overall, however, we conclude that machine learning methods can filter through immense amounts of data available to identify signals of illness with a high level of accuracy. This suggests two major applications for monitoring and treatment. At the individual level, signals of mental illness could provoke response, either from the user's community or through automated means from the social media platform itself (for example, offering counseling resources). At the population level, given adjustments for biases in Internet use and other factors, mental health trends can be monitored in near real-time, which may be especially useful following acute events of social stress such as recessions, political crises, and natural disasters. Social media monitoring will not replace more formal approaches to mental health surveillance, but it can complement these other tools.

# V CONCLUSION: REIMAGINING THE REVOLUTION

Big data is a valuable resource in the fine-grained measurement of women's and girls' well-being. Flowminder's geospatial modeling work, based on satellite imagery, provides a high-resolution look at social and health outcomes as they vary over space; the same method could be used to create data systems that capture variation in welfare over time. Expenditure patterns inferred from credit card and cell phone expenditure records, as in the MIT-led work, provide a detailed look at economic activity across different social groups and over time. The social media-based projects achieve the same objective of high-spatial resolution, high-frequency measurement, but with a focus on emotions, thoughts, and ideas.

Overall, this report illustrates the potential of big data in filling the global gender data gap. In closing, however, we note that the rise of big data does not signify that traditional sources of data will become less important. On the contrary, the successful implementation of big data requires investment in proven methods of social scientific research, not least for the validation and bias correction of big datasets. For example, Flowminder's work requires DHS or other types of survey data as a starting point, as well as field biophysical data to calibrate the interpretation of satellite imagery. Inferring women's economic behavior from cell phone and credit card records demands ground-truthing work to assess how strongly, within a given culture and economy, these records reflect overall social and consumer behavior. Twitter users are a biased sample of society at large, and determining the magnitude and direction of that bias through surveys is critical if this information is to be useful in assessing population-level patterns.

More broadly, big data is not a panacea for all the challenges of development planning and research. The invisibility of women and girls in international and national data discourse is a political, not solely a technical, problem. New methods can indeed illuminate previously ignored aspects of the lives of women and girls, but it can also create a sense that technical advancements alone will compel investments in gender-sensitive data systems by national statistical agencies, civil society organizations, and international donors. They will not. In the worst-case scenario, they will have the opposite effect: the data deluge may shift policy focus towards the groups and regions for which the most information is available, not the people and places in greatest need. Even big data illuminates only small parts of the entire field of human experience.

In the best case, however, the current "data revolution" will be reimagined as a step towards building "data governance": a process through which novel types of data bring about not instant, perfect knowledge about global development processes, but rather catalyze the creation of new partnerships for the sharing and interpretation of information. In the best case, projects that use big datasets will be informed by thoughtful hypotheses advanced by women and girls themselves, take a pragmatic but tireless approach to data policy reform in a decision-making world still dominated by men, and work in concert with advocates for the inclusion of women and girls in all spheres of social and economic life. These are the kinds of projects we have profiled in this report, and they hold great promise for the future of social science and policymaking.

Elizabeth Whelan

# NOTES

[1] See Alegana et al. 2015 and Sedda et al. 2015 for similar past work

[2] ICF International 2012

[3] HDRO (UNDP) 2015

[4] KNBS 2010; NBS 2011; NPC 2014; NIPORT 2013; Cayemittes et al. 2013

[5] Alegana et al. 2015; Gething et al. 2015

[6] Lazer et al. 2009

[7] Yoshimura et al. 2009; Krumme et al. 2013; Giles 2012

[8] Toole et al. 2015; Gonzalez, Hidalgo, and Barabasi 2008; Jiang et al. 2016; Song et al. 2010; Blumenstock, Cadamuro, and On 2015; Lenormand et al. 2015; Louail et al. 2014; Çolak et al. 2016

[9] Eagle, Macy, and Claxton 2010; Gonzalez, Hidalgo, and Barabasi 2008; Pappalardo et al. 2015.

[10] Cauce et al. 2002; Taylor and Brown 1988

[11] Wang et al. 2000

[12] WHO 2001

[13] Spector 2002

[14] Taylor and Brown 1998; Spector 2002

[15] Ormel et al. 1994; Patel et al. 1999

[16] Coppersmith, Dredze, and Harman 2014; Coppersmith, Harman, and Dredze 2014; Culotta 2014; De Choudhury, Counts, and Horvitz 2013; De Choudhury et al. 2014; De Choudhury et al. 2013; Eichstaedt et al. 2015; Homan et al. 2014

[17] Coppersmith, Dredze, and Harman 2014; Coppersmith, Harman, and Dredze 2014

[18] Zhu, Ghahramani, and Lafferty 2003; De Choudhury et al. 2016

[19] Pennebaker, Francis, and Booth 2001; Chung and Pennebaker 2007; De Choudhury and De 2014

[20] Gross and Muñoz 1995

[21] Ramirez-Esparza 2008

[22] Chancellor 2016

[23] Lieberman and Goldstein 2006

[24] Guillemin, Bombardier, and Beaton 1993

## PHOTO CREDITS

Cover, pages 5, 19, and 30: © Elizabeth Whelan. All rights reserved. ElizabethWhelanPhotography.com

Page 28: Adam Cohn, "Indian Woman with Smartphone." Some rights reserved, CC BY-NC-ND 2.0 (Creative Commons Attibution-NonCommercial-NoDerivs 2.0 Generic) license. Cropped. AdamCohn.com

# REFERENCES

Alegana, Victor A., Peter M. Atkinson, Carla Pezzulo, Alessandro Sorichetta, D. Weiss, T. Bird, E. Erbach-Schoenberg, and Andrew J. Tatem. 2015. “Fine resolution mapping of population age-structures for health and development applications.” Journal of the Royal Society Interface 12, no. 105: 20150073.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. “Predicting poverty and wealth from mobile phone metadata.” Science 350, no. 6264: 1073-1076.

Cauce, Ana Mari, Melanie Domenech-Rodriguez, Matthew Paradise, Bryan N Cochran, Jennifer Munyi Shea, Debra Srebnik, and Nazli Baydar. 2002. “Cultural and contextual influences in mental health help seeking: a focus on ethnic minority youth.” Journal of Consulting and Clinical Psychology 70, no. 1: 44.

Cayemittes, Michel, Michelle Fatuma Busangu, Jean de Dieu Bizimana, Bernard Barrère, Blaise Sévère, Viviane Cayemittes et Emmanuel Charles. 2013. Enquête Mortalité, Morbidité et Utilisation des Services, Haïti, 2012. Calverton, Maryland, USA: MSPP, IHE et ICF International.

Chancellor, Stevie, Zhiyuan (Jerry) Lin, Erica Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. “Quantifying and predicting mental illness severity in online pro-eating disorder communities.” In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1171-1184. ACM.

Chung, Cindy, and James W Pennebaker. 2007. “The psychological functions of function words.” Social Communication: 343-359.

Çolak, Serdar, Antonio Lima, and Marta C. González. 2016. “Understanding congested travel in urban areas.” Nature Communications 7: 10793. doi:10.1038/ncomms10793.

Coppersmith, Glen, Craig Harman, and Mark Dredze. 2014. “Measuring post traumatic stress disorder in Twitter.” In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 579-582.

Coppersmith, Glen, Mark Dredze, and Craig Harman. 2014. “Quantifying mental health signals in Twitter.” In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: from Linguistic Signal to Clinical Reality, pp. 51-60. Baltimore, Maryland: Association of Computational Linguistics.

Culotta, Aron. 2014. “Estimating county health statistics with Twitter.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1335-1344. ACM.

De Choudhury, Munmun, and Sushovan De. 2014. “Mental health discourse on Reddit: Self-disclosure, social support, and anonymity.” In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM).

De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. “Discovering shifts to suicidal ideation from mental health content in social media.” In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2098-2110. ACM.

De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. "Predicting depression via social media." In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM).

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013. "Social media as a measurement tool of depression in populations." In Proceedings of the 5th Annual ACM Web Science Conference, pp. 47-56. ACM.

De Choudhury, Munmun, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. "Characterizing and predicting postpartum depression from Facebook data." In Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM.

Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. "Network diversity and economic development." Science 328, no. 5981: 1029-1031.

Eichstaedt, Johannes C., Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar, Martin E.P. Seligman. 2015. "Psychological language on Twitter predicts county-level heart disease mortality." Psychological Science 26, no. 2: 159-169. doi:10.1177/0956797614557867.

Gething, Peter, Andy Tatem, Tom Bird, and Clara R. Burgert-Brucker. 2015. "Creating Spatial Interpolation Surfaces with DHS Data." DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International.

Giles, Jim. 2012. "Making the links." Nature 488, no. 7412: 448-450.

Gonzalez, Marta C., Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. "Understanding individual human mobility patterns." Nature 453, no. 7196: 779-782.

Gross, James J., and Ricardo F Muñoz. 1995. "Emotion regulation and mental health." Clinical Psychology: Science and Practice 2, no. 2: 151-164.

Guillemin, Francis, Claire Bombardier, and Dorcas Beaton. 1993. "Cross-cultural adaptation of health related quality of life measures: literature review and proposed guidelines." Journal of Clinical Epidemiology 46, no. 12: 1417-1432.

Homan, Christopher M., Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. "Social structure and depression in TrevorSpace." In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 615-625. ACM.

Human Development Report Office (HDRO), United Nations Development Program (UNDP). 2015. Human Development Report 2015: Work for Human Development. New York: United Nations Development Program.

ICF International. 2012. Demographic and Health Survey Sampling and Household Listing Manual. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International.

Jiang, Shan, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. "The TimeGeo modeling framework for urban motility without travel surveys." Proceedings of the National Academy of Sciences: 201524261. doi:0.1073/pnas.1524261113.

Kenya National Bureau of Statistics (KNBS) and ICF Macro. 2010. Kenya Demographic and Health Survey 2008-09. Calverton, Maryland: KNBS and ICF Macro.

Krumme, Coco, Alejandro Llorente, Manuel Cebrian, and Esteban Moro. 2013. "The predictability of consumer visitation patterns." Scientific Reports 3: 1645.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Guttman, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Life in the network: the coming age of computational social science." Science 323, no. 5915: 721-723.

Lenormand, Maxime, Thomas Louail, Oliva G. Cantú-Ros, Miguel Picornell, Ricardo Herranz, Juan Murillo Arias, Marc Barthelemy, Maxi San Miguel, and José J. Ramasco. 2015. "Influence of sociodemographic characteristics on human mobility." Scientific Reports 5: 10075. doi:10.1038/srep10075.

Lieberman, Morton A., and Benjamin A Goldstein. 2006. "Not all negative emotions are equal: The role of emotional expression in online support groups for women with breast cancer." Psycho-Oncology 15, no. 2: 160-168.

Louail, Thomas, Maxime Lenormand, Oliva G. Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. 2014. "From mobile phone data to the spatial structure of cities." Scientific Reports 4: 5276. doi:10.1038/srep05276.

National Bureau of Statistics (NBS) [Tanzania] and ICF Macro. 2011. Tanzania Demographic and Health Survey 2010. Dar es Salaam, Tanzania: NBS and ICF Macro.

National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International. 2013. Bangladesh Demographic and Health Survey 2011. Dhaka, Bangladesh and Calverton, Maryland, USA: NIPORT, Mitra and Associates, and ICF International.

National Population Commission (NPC) [Nigeria] and ICF International. 2014. Nigeria Demographic and Health Survey 2013. Abuja, Nigeria, and Rockville, Maryland, USA: NPC and ICF International.

Ormel, Johan, Michael VonKor_, T Bedirhan Ustun, Stefano Pini, Ailsa Korten, and Tineke Oldehinkel. 1994. "Common mental disorders and disability across cultures: results from the WHO collaborative study on psychological problems in general health care." Journal of the American Medical Association 272, no. 22: 1741-1748.

Pappalardo, Luca, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. 2015. "Using big data to study the link between human mobility and socio-economic development." In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), pp. 871-878. IEEE Computer Society. doi:10.1109/BigData.2015.7363835

Pennebaker, James W., Martha E Francis, and Roger J Booth. 2001. “Linguistic inquiry and word count: LIWC 2001.” Mahway: Lawrence Erlbaum Associates, 71:2001.

Rachel E Spector. 2002. “Cultural diversity in health and illness.” Journal of Transcultural Nursing 13, no. 3: 197-199.

Ramirez-Esparza, Nairan, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. “The psychology of word use in depression forums in English and in Spanish: Texting two text analytic approaches.” In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM).

Sedda, Luigi, Andrew J. Tatem, David W. Morley, Peter M. Atkinson, Nicola A. Wardrop, Carla Pezzulo, Alessandro Sorichetta, Joanna Kuleszo, and David J. Rogers. 2015. “Poverty, health and satellite-derived vegetation indices: their inter-spatial relationship in West Africa.” International Health 7, no. 2: 99-106.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. “Limits of predictability in human mobility.” Science 327, no. 5968s: 1018-1021.

Taylor, Shelley E., and Jonathon D Brown. 1988. “Illusion and well-being: a social psychological perspective on mental health.” Psychological Bulletin 103, no. 2: 193.

Toole, Jameson L., Carlos Herrera-Yaqüe, Christian M. Schneider, and Marta C. González. 2015. “Coupling human mobility and social ties.” Journal of the Royal Society Interface 12, no. 105: 20141128. doi:10.1098/rsif.2014.1128.

Vikram Patel, Ricardo Araya, Mauricio de Lima, Ana Ludermir, and Charles Todd. 1999. “Women, poverty and common mental disorders in four restructuring societies.” Social Science & Medicine 49, no. 11: 1461-1471.

Wang, Xiangdong, Lan Gao, Naotaka Shinfuku, Huabiao Zhang, Chengzhi Zhao, and Yucun Shen. 2000. “Longitudinal study of earthquake-related PTSD in a randomly selected community sample in North China.” American Journal of Psychiatry 157, no. 8: 1260-1266. doi:10.1176/appi.ajp.157.8.1260.

World Health Organization. 2001. The World Health Report 2001: Mental health: new understanding, new hope. World Health Organization.

Yoshimura, Yuji, Stanislav Sobolevsky, Juan N. Bautista Hobin, Carlo Ratti, and Josep Blat. 2016. “Urban association rules: uncovering linked trips for shopping behavior.” Environment and Planning B: Planning and Design: 0265813516676487.

Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. 2003. “Semi-supervised learning using Gaussian fields and harmonic functions.” In Proceedings of the Twentieth International Conference on Machine Learning (ICML-03), pp. 912-919.

data2x

@Data2X

www.data2x.org