Towards High-Resolution Sex-Disaggregated Dynamic Mapping

FLOWMINDER FOUNDATION

Introduction

In this brief, we present results of a mixed-methods study to investigate how novel digital data sources can support gender-equitable development across Nepal. We implemented two bodies of work. First, we combined geolocated survey data, satellite imagery, and mobile phone data to map three key gendered indicators — literacy, agriculture-based occupations, and births in health facilities — at very high spatial resolution. Second, we sought to use de-identified mobile phone data to produce robust, frequently updatable information on gendered mobility and migration patterns within Nepal. This second body of work required us to predict gender among a population of mobile phone subscribers. Our results suggest that SIM sharing is an important complicating factor in predicting gender, and thus inferring individual well-being, from mobile operator data. Overall, we find that combining traditional survey data sources with various forms of digital data holds great promise for a spatially and temporally rich understanding of women's and girls' lives, although more validation work is needed on patterns of SIM use.

Our Approach

In this study, we worked with geo-tagged survey data collected for the 2016 Nepal Demographic and Health Survey (NDHS), focusing on seven indicators: educational attainment, literacy, labor market participation, agriculture-based occupations, attitudes to gender-based violence, births in health facilities, and child stunting. Because of cost and logistical considerations, demographic and health surveys are typically not designed to permit highly local inferences. However, the spatial resolution of survey data can be improved with the complementary use of geospatial and mobile phone data.

We collated a variety of geospatial information from open-source platforms, including datasets containing physical (topography, climate, land cover, vegetation, biomass, evapotranspiration), social (population, ethnicity), and built environment (urbanization, human settlements) variables. We also partnered with one of Nepal's largest mobile network operators, Ncell[1], to analyze a subscriber database of 15 million registered SIMs, with information between January-December 2016 on time, duration, location, and parties of each call, as well as daily financial credit 'top-up' totals and counts of recharge type.

While subscriber records may include a gender tag, this information is not validated, and is often entirely missing. We thus built a model to predict gender based on daily and home locations of subscribers, calling patterns, and top-up behavior. To support this model — specifically, to validate existing gender tags — we conducted a survey of a subset of 5,180 subscribers.

Finally, using machine learning and Bayesian geostatistical methods, we built models analyzing the geospatial and mobile phone variables as predictors of the survey indicators of interest. We used these observed relationships to create high-resolution (1km2) gridded datasets of the indicators, with associated uncertainty. The sex-disaggregated maps are updatable, and thus able to track progress on development targets over time, as well as assess short-term fluctuations in well-being associated with economic and environmental shocks.

---

1. This study would not have been possible without the support of NCELL Axiata.

Results

Figure 1. The proportion of variance explained by the applied models.

Proportion of variance explained

(Female — orange, Male — teal)
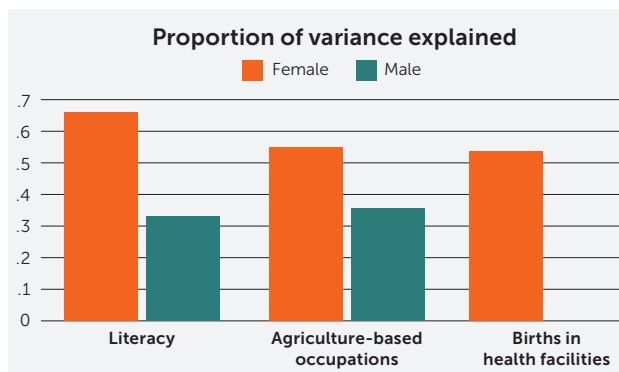
The modeling results are generally encouraging: analysis of individual CDR data can enhance our understanding of the spatial variation and temporal dynamics of gender inequality. We found that female literacy, agriculture-based occupations, and births in health facilities were especially amenable to our approach, with the best-performing models explaining around 60% of variance (Figure 1). The models predicting literacy and farm-based livelihoods performed much better for women than men. The reasons for this are not clear but are likely related to the fact that social and economic institutions — especially a strong historical male-child preference in South Asia — play an intervening role in the relationship between geospatial and well-being variables. Educational attainment, child stunting, and gender-based violence were only weakly associated with our chosen set of predictors. However, these indicators may perform better in other contexts, as suggested by our previous work.[1]

We also found that SIM sharing within households greatly complicates inferences from mobile operator data, despite the very high overall rates of individual mobile phone ownership in Nepal. Most demographic prediction models have assumed that a single CDR record corresponds to a single individual (the 'single-SIM/single-subscriber-user' assumption). Our results suggest that this assumption should be strongly questioned, particularly in Nepal but possibly also in other contexts. Almost one-half of survey respondents indicated shared SIM use, with no difference between men and women in the likelihood of sharing. Most of the sharing (92%) occurred with the family. In addition, nearly one-third of gender tags in the MNO database did not correctly identify the gender of the main user of the SIM.
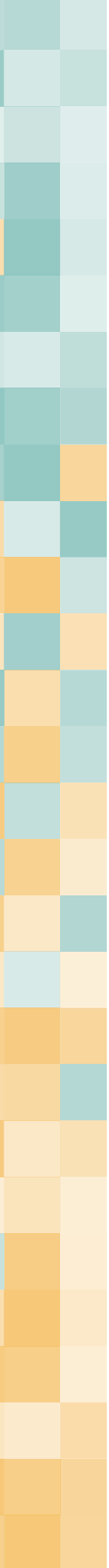
Implications

The combination of geo-located survey, geospatial, and mobile phone data holds great promise for creating well-being monitoring systems with high resolution in both space and time. Figure 2 shows how the portrait of female participation in agriculture-based occupations is enriched by our modeling approach. Such high-resolution data systems are necessary to monitor progress towards the Sustainable Development Goals, and more generally allocate resources to the places and at the times when they are most needed.

However, the use of digital data for spatial modeling must confront several obstacles. First, the underlying survey data is typically limited in

Figure 2. (A) Map of female engagement in farm-based livelihoods at province level (source: authors analysis of weighted NDHS survey data). (B) Map of the population of female engagement in in farm-based livelihoods at 1km2 resolution
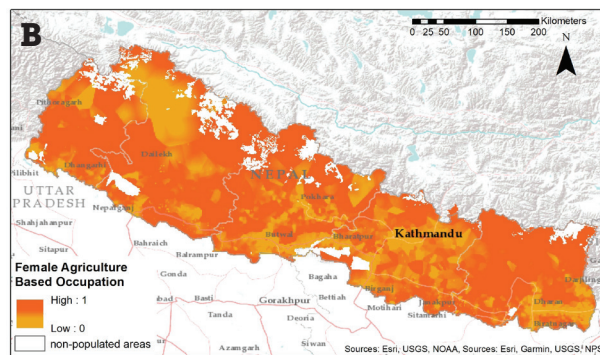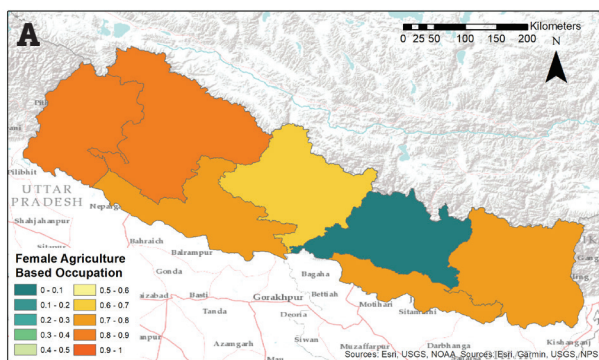
1. See, for example, our research featured in the Data2X report "Big Data and the Well-being of Women and Girls."

sample size, which inhibits the full exploitation of the model architectures. This reinforces the critical point that traditional and new forms of data are complementary, not competing. Second, because of the large number of geospatial covariates and cell phone features available, stronger theory about which combinations of variables are most likely to predict a given indicator would be valuable. Finally, the use of CDR data for improving our understanding of gendered phenomena depends on accurately identifying user gender. Even when tags are available, they may be inaccurate for a variety of reasons, including SIM sharing. Improving our understanding of SIM sharing, which is likely to vary greatly by context — and even in a given context, may change over time as economies evolve — is thus vital. When tags are not available, gender prediction models may help, but again requires careful validation research on calling and sharing behaviors.

These limitations notwithstanding, the potential of these new types of digital data is clear. The data used in this study — geospatial and mobile phone information — is readily available in massive quantity at low cost, and these datasets will continue to increase in size and resolution in the coming years. Building capacity in public sector agencies is an important step to fully realizing this potential. In Nepal, the Flowminder Foundation is working closely with the Central Bureau of Statistics (CBS) to use the modeling approaches outlined in this brief to create a well-being monitoring system embedded within the agency.