FLOWMINDER.ORG

Towards high-resolution sex-disaggregated dynamic mapping

Final report

Flowminder Foundation
September 2019

Towards high-resolution sex-disaggregated dynamic mapping

Final Report

Claudio Bosco
Flowminder Foundation, Stockholm, Sweden.
WorldPop, Department of Geography and Environment, University of Southampton, UK

Samantha Watson
Flowminder Foundation, Stockholm, Sweden.
WorldPop, Department of Geography and Environment, University of Southampton, UK

Alina Game
Flowminder Foundation, Stockholm, Sweden.
WorldPop, Department of Geography and Environment, University of Southampton, UK

Chris Brooks
Flowminder Foundation, Stockholm, Sweden.
WorldPop, Department of Geography and Environment, University of Southampton, UK

Daniele de Rigo
Maieutike Research Initiative, Milano, Italy

Sarchil Qader
WorldPop, Department of Geography and Environment, University of Southampton, UK

Joshua Greenhalgh
Flowminder Foundation, Stockholm, Sweden.

Kristine Nilsen
WorldPop, Department of Geography and Environment, University of Southampton, UK

Amy Ninneman
Flowminder Foundation, Stockholm, Sweden
WorldPop, Department of Geography and Environment, University of Southampton, UK

Richard Wood
Flowminder Foundation, Stockholm, Sweden

Linus Bengtsson
Flowminder Foundation, Stockholm, Sweden
WorldPop, Department of Geography and Environment, University of Southampton, UK

The study was made possible by the generous support of Data2X at the UN Foundation

Abstract

While the social status and position of women and men, girls and boys in Nepal - as elsewhere - is cut through by geography, social class, race, ethnicity, and age (life-stage), historically women and girls have been disproportionately subject to gender-based disadvantages, both legally enshrined and institutionalised as social norms and expectations (Matinga et al., 2019).

In recent years, the Government of Nepal has sought to address major sites of gender-based disadvantage, introducing a series of legal and regulatory provisions to strengthen women's position in society and advance gender equality. The 2015 Constitution mandated that women occupy a third of parliamentary seats, and introduced a raft of new rights previously withheld from women. Newly available rights include: rights to inheritance (lineage), to reproductive and maternal health provision, and equal rights in property and family matters (Government of Nepal, 2015). There followed a series of measures to address gender-based inequalities in educational attainment and in legally recognised use-rights over land (at a time when under 20% of women had land registered in their name (IOM, 2016).

Despite these recent moves to diminish gender-based inequalities, women and girls in Nepal - as elsewhere - continue to be disproportionately subject to gender-based disadvantages, both legally enshrined and institutionalised as social norms and expectations (Care, 2015).

Against this backdrop, this study investigated the potential for novel digital data sources to support gender-equitable development across Nepal.

The study was organised around two work packages. In the first, we combined nationally representative, geo-located survey data with satellite imagery and mobile phone data, to model and map spatial variations and gender-based inequalities for three, key development indicators (literacy, agriculture-based-occupations, and births in health facilities) across Nepal.

The results obtained for work package one demonstrate the power of modern and robust statistical methods to exploit geolocated survey data in new and innovative ways, so permitting the geographical scale of survey estimates to be greatly refined. We discuss the data requirements underpinning good model performance, contrasting, for example, the weaker results obtained for male literacy rates with results for the best-performing indicators.

Notwithstanding the potential for results to be improved through the inclusion of additional information, we suggest that the showcased techniques can (potentially) be applied to a wide variety of development indicators. We outline the practical relevance of the study outputs for the design, implementation, and monitoring of gender-equitable development in Nepal.

The second work package sought to leverage de-identified mobile phone data to produce robust, frequently updatable, information on gendered mobility and migration patterns, trajectories, and dynamics within Nepal. This entailed the development of methods to predict

gender for a 'population' of mobile phone subscribers. As part of this workstream, we administered a primary survey to validate gender for a representative sample of subscribers.

To our knowledge, this study is the first time that a rigorous assessment of SIM-card (Subscriber Identification Module-card) sharing has been undertaken and incorporated into model architectures for demographic prediction. The study findings indicate that it is common for individuals to use one another's SIM-cards, despite (overall) high rates of individual mobile phone ownership in Nepal. Our results suggest that the 'single-SIM/single subscriber' assumption (which has, to date, underpinned demographic prediction models) is untenable in the study setting.

The uncertainty introduced by widespread SIM sharing in this setting is higher than traditionally allowed for by 'classic methods'. The extent to which the pattern observed for Nepal holds in different settings is an empirical question. Ultimately, it may be necessary to reassess the performance of 'classic methods' to predict demographics from CDR data in light of previously undetected sources of uncertainty. This will depend on further research to assess the extent of (unacceptable) uncertainty posed by SIM use and sharing in different settings.

Seeking to compensate for the uncertainty introduced by reported widespread SIM-sharing, we applied state-of-the-art semantic array programming - a robust, modular modelling approach - to model women's and men's mobility and migration patterns.

While the model results are encouraging, indicating that analysis of individual CDR data can enhance our understanding of the spatial variation and temporal dynamics of sex and gender-based inequalities, more work is needed to unravel the implications of SIM sharing for gender (and more broadly, demographic) prediction models. We make a number of recommendations in this regard.

Table of Contents

List of Acronyms and Abbreviations

ANN	Artificial Neural Network
AUC	Average area Under the Curve
BGS	Bayesian Geostatistical
CAPI	Computer Aided Personal Interviewing
CBS	Central Bureau of Statistics
CCI	Climate Change Initiative
CDR	Call Detail Records
CEDA	Centre for Environmental Data Analysis
CIESIN	Columbia University Centre for International Earth Science Information Network
DEM	Digital Elevation Model
DHS	Demographic and Health Surveys – the DHS Program assists developing countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs.
D-TM	Data-Transformation Model – a D-TM is a conceptual unit which transforms a set of input data and model parameters into a corresponding set of output data. In this context, data are intended as array-based aggregations of (potentially uneven) elements.
EA	Enumeration Area – is the operational geographic units for the collection of census data
EOC	Earth Observation Centre
EPR	Ethnic Power Relations
ESA	European Space Agency
ET	Total Evapotranspiration
EVI	Enhanced Vegetation Index – is an 'optimised' vegetation index designed to enhance the vegetation signal with improved sensitivity in high biomass regions and improved vegetation
GBT	Gradient Boosting Trees
GBV	Gender-Based Violence

GeoSemAP	Geospatial Semantic Array Programming - geospatial application of the SemAP paradigm, where the conceptual units (D-TMs) of the modelling workflow are a composition of geospatial transformations and array-based D-TMs.
GFSAD	Global Food Security Support Analysis Data
GHSL	Global Human Settlements Layer – is a dataset containing new global spatial information, evidence-based analytics and knowledge describing the human presence on the planet
GIS	Geographic Information Systems
GNI	Gross National Income
GPP	Gross Primary Productivity
GRUMP	Global Rural Urban Mapping Project
GUF	Global Urban Footprint
HDX	Humanitarian Data Exchange
INLA	Integrated Nested Laplace Approximations – is a package (in R) that exploit the approach of Integrated Nested Laplace Approximations to do approximate Bayesian inference for latent Gaussian models.
MAE	Mean Absolute Error
MIR	Middle Infra-red reflectance
ML	Machine Learning
MNO	Mobile Network Operators
MODIS	Moderate Resolution Imaging Spectrometer
MSE	Mean Squared Error
NDHS	Nepal Demographic and Health Survey
NDVI	Normalised Difference Vegetation Index
NPP	Net Primary Productivity
OSM	Open Street Map – is a project that creates and distributes free geographic data for the world.
PCA	Principal Component Analysis
PET	Potential Evapotranspiration

PSU	Primary Sampling Unit - refers to sampling units that are selected in the first (primary) stage of a multi-stage sample ultimately aimed at selecting individual elements
QA	Quality Assurance
RMSE	Root Mean Square Error
RS	Remote Sensing
SDG	Sustainable Development Goal
SemAP	Semantic Array Programming - a computational modelling approach to compactly process arrays of data preserving the consistency of their underpinning semantics. SemAP is based on the modularisation of the modelling workflow into conceptual units (modules) of data-transformation (see D-TM), and on the systematic use of array-based semantic constraints.
SIEVE	Selective Improvement by Evolutionary Variance Extinction - Training architecture for nonlinear computational models, such as artificial neural networks.
SIM	Subscriber Identification Module
SPDE	Stochastic Partial Differential Equations
SRTM	Shuttle Radar Topography Mission - is an international research effort that obtained digital elevation models on a near-global scale from 56° S to 60° N
USGS	United States Geological Survey
VI	Vegetation Index
VIF	Variance Inflation Factor
VIIRS	Visible Infrared-Imaging Radiometer Suite
WDPA	World Database on Protected Areas
WHO	World Health Organization

Background and Objectives

An equitable and efficient allocation of international aid relies on knowing where resources are needed most. Unfortunately, detailed, reliable and timely information on the spatial distribution and characteristics of intended aid recipients in many low income countries are rarely available, impacting the ability of aid agencies to effectively and equitably distribute resources to those most in need.

To meet and assess progress towards the global Sustainable Development Goals (SDGs), it is crucial to improve the understanding of geographic variation in population wellbeing indicators such as health status, wealth and access to resources.

Sustainable development will, however, not be possible if significant life opportunities are denied to women and girls. Women and girls should be provided with equal access to education, economic resources and political participation, and have equal opportunities in all fields and at all levels. Unfortunately, in Nepal, frequently women do not have equal access to resources, and it is critical to understand the impact on women's welfare.

Indices of need can be derived from demographic measures obtained from individual-based surveys, the majority of which now record anonymised spatial data on the locations of surveys. In many cases, census data may provide the relevant information, and can be used to accurately depict the status of a population, sometimes at the level of enumeration areas or cities. However, enumeration area-level census data can often be outdated or unreliable, or simply hard to obtain.

An alternative approach to the overall measurement that census data provide is to use a random subsampling to obtain a representative sample of the population. Geolocated household survey data from the Demographic and Health Surveys (DHS) (http://dhsprogram.com), for example, have been used extensively to provide broad-scale estimates of factors such child mortality, nutrition and literacy. Such surveys can be used to enrich census-based data or, where census data are outdated, unavailable or unreliable, infer values at unobserved locations using predictive modelling.

One of the main aims of this study is to leverage the large-scale spatiotemporal data collected by Mobile Network Operators (MNOs) on mobile phone users in Nepal, to improve understanding of sex-disaggregated demographic and vulnerability characteristics, as well as their dynamics. To reach this objective we worked on developing innovative modelling techniques focused on splitting the CDR database by gender (based on the differences in men's and women's observable SIM use episodes), in order to support the mapping of women's and men's mobility, migration patterns and dynamics.

Another objective of this project is to explore the feasibility of producing high-resolution maps of a diverse set of sex-disaggregated indicators of relevance to female welfare. High resolution maps of sex-disaggregated vulnerability characteristics are important for evidence-based policy making. These maps will, for instance, support better targeted interventions aimed at

increasing female resilience and education. A single map provides a snapshot of the spatial distribution of given conditions or challenges. We aim at providing a series of updatable maps: as mapping becomes dynamic, changes and progress over time and across space can be tracked and monitored.

The applied statistical methodology was built upon a combination of GPS-located household survey data, satellite-derived covariate datasets and individual level Call Detail Record (CDR) data, producing high-resolution (1km^2) gridded datasets of key indicators together with associated uncertainty.

To ascertain that the data and maps are used and understood by policy makers, we are working closely with the Central Bureau of Statistics (CBS) in Nepal.

The context

Nepal is characterised by having a medium-low human development index (UNDP, 2018). It is ranked 149 out of 189 countries worldwide. Nepal has a mean annual population growth rate of 1.7% with a Gross National Income (GNI) per capita, calculated using the Atlas method, of $960 (https://databank.worldbank.org). A nationally representative, gender-balanced survey sample, the 2016 Nepal Demographic and Health Survey (MOH et al., 2017), provides a comprehensive dataset of a variety of indicators by gender. Available data show that in Nepal 36 percent of children under the age of five are stunted, with a similar level in male and female children. Men are more likely to have a secondary or higher education (71%) compared to women (50%). Educational attainment is higher in urban areas for women and men aged 15-49 (57% and 76% respectively) compared to those in rural areas (39% and 62% respectively). Literacy rates in this age group are also higher for men, with 89% being recorded as literate compared to the 69% of women.

There are also gender-based differences in work and employment patterns within the country. The number of men aged 15-49 who are currently in employment stands at 78%. The comparable figure for women is 57%. This pattern is reversed for agricultural employment however. The percentage of currently working women engaged in the agricultural sector is 70%, compared to 33% of currently working men. Furthermore, women in agriculture are less likely than men to receive payment for their work (29% versus 47% for men).

An indicator of particular interest in the Nepal DHS 2016 is domestic violence, due to concerns of increasing risk of gender based violence. The percentage of women aged 15-49 years old who have experienced any physical violence (in the last 12 months) is at 22%, with 9% of records reporting physical violence “often” or “sometimes” within the past year.

Material and methods

1. Primary Survey Data

1.1 Rationale

A major aim of the study is to assess the scope for gender to be predicted for a population of mobile phone subscribers based on 'demographic traces' detectable in CDR data. While the MNO subscriber database records gender for the subscriber, this information is both incomplete (missing in up to 15% of cases), and un-validated. A broader challenge concerns the in-practice distinction between SIM subscriber and SIM user, i.e. the extent to which the subscriber-of-record and user of a SIM can be assumed to be identical. Another important consideration is the extent to which SIM usage events can be assigned to a single, gendered, individual. While powerful methods are available to predict individual characteristics from CDR traces alone, they rely on strong assumptions.

Primary survey data, gathered for a probabilistically sampled subset of the subscriber base, offer a means to 'anchor' prediction models in validated data. Work in this arena is part of a nascent but growing field to integrate novel and traditional datasets, so leveraging the advantages of both[1]. The overall objective of the primary survey is to support the development of modelling tools to predict SIM user's gender from routine mobile phone data. Two interrelated survey aims follow:

i. Validate gender for a representative subsample of SIM records

ii. Test the assumption that activity observed for a specified SIM wholly or mainly corresponds to a single individual

Survey linkage' performs a second important function, providing data to enable bias and uncertainty to be quantified and compensated for in the final model architecture. This is important, since the generation of CDR 'profiles' depends on a number of bracketed assumptions around the wider generalisability of profiles for the 'population' of (relatively active) mobile phone users to the national population (or other population of interest).

1.2 Sample design

A telephone-mode survey was planned with a probabilistically sampled subset (n = 4,268) of SIM users drawn from a database maintained by the partner MNO. At the time of the sample selection, the subscriber database held information on approximately 15 million SIMs, and their registered subscribers.

[1]Recent work has focussed on deriving gender (Jahani et al 2017), employment status (Almaatouq et al 2015), education (Sundsøy 2016a), occupation (Sundsøy et al 2016a), household wealth (Šćepanović et al 2015), and individual income (Blumenstock et al 2015, Sundsøy et al 2016b, 2016c).

We utilised a stratified single-stage without replacement (stsrswor) design (95% CI; e= 0.015), exploiting features of the subscriber database sampling frame and CDR dataset to produce a geographically balanced sample across Nepal's seven provinces and three ecological zones (Terai, Hilly, Himal), and allow for independent estimates for rural and urban areas (figure 1.1).

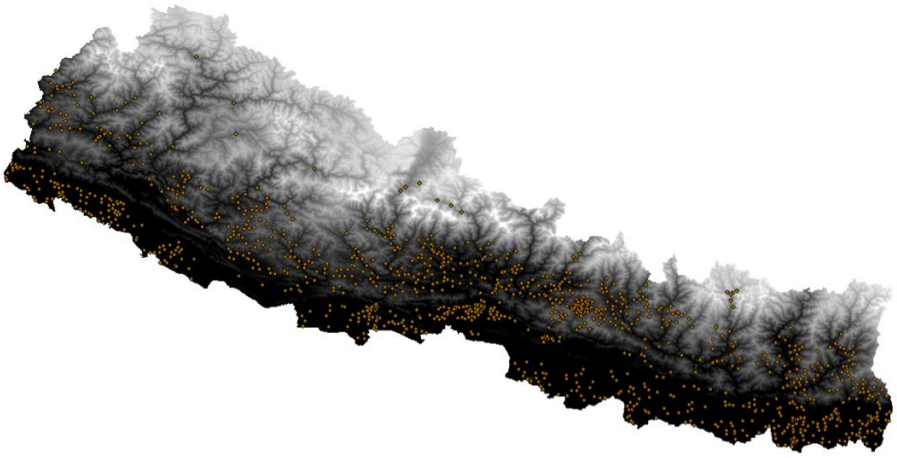

Figure 1.1: Sample distribution, primary mobile-phone-mode survey

The method for 'home location' estimation, to enable a geographically balanced sample, is detailed in the subsequent section. Rural residents[2] and subscribers registered as female were oversampled relative to urban and male subscribers[3] to address under-representation of these groups in the sampling frame (the CDR data and subscriber database). Figure 1.2 provides an overview of the sample selection process.

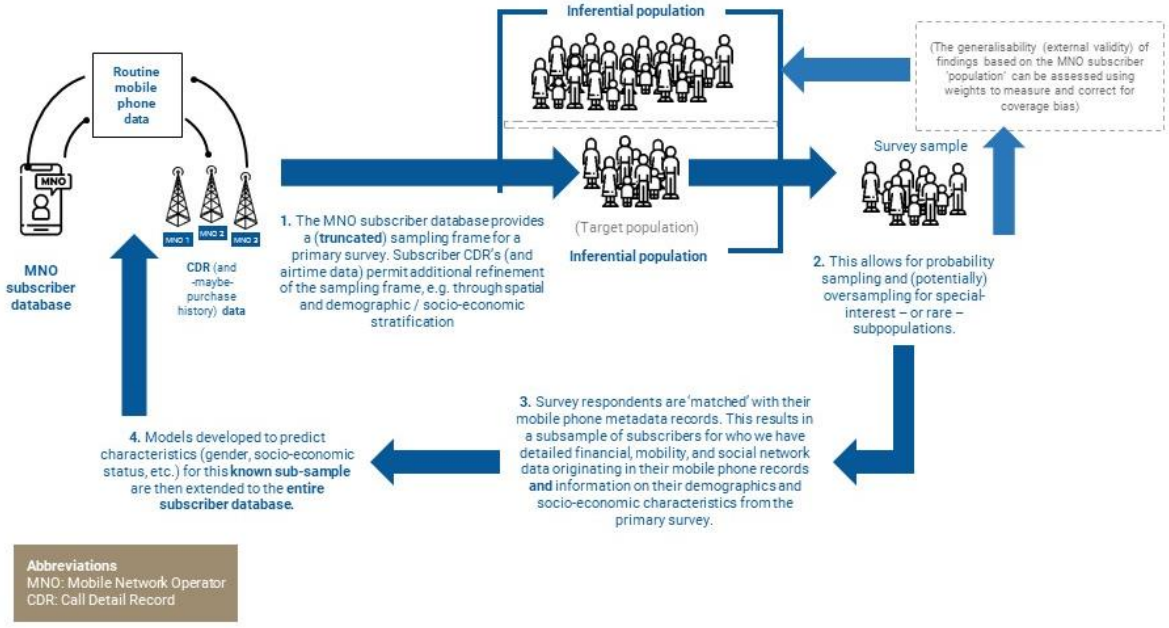

Figure 1.2: Schematic, sampling a statistically representative subset of subscribers

[2]Analysis of CDR home location demonstrated the MNO subscriber database to be skewed towards urban residency (75% urban to 25% rural). We oversampled rural home locations at a 50:50 sampling rate.

[3]The MNO subscriber database recorded 25% of subscribers as female and 60% as male. The field for sex was vacant for 15% of records. We merged subscribers with vacant sex fields with the male listing prior to sampling (since proportionately more likely to be men) to avoid systematic exclusion of subscribers with no recorded sex. Subscribers recorded as female were oversampled with equal (i.e. 50:50) sampling based on recorded sex.

The initial sample size was inflated to insulate results from non-response (planning was based on a conservative total contact rate of 40%, adjusted on the basis of piloting). Favouring simplicity, we initiated a standard 'sample replicates' approach, with selected sample units randomly assigned to a series of 'batches' for consecutive release until the required sample size was obtained[4]. A protocol for repeat calls was instituted, requiring three contact attempts to be made to survey each sample unit, with calls made on different days and at different times of the day, prior to declaration as non-response[5].

As detailed below, non-trivial anomalies became apparent in the survey data during cleaning. The anomalies were indicative of response and / or data-entry error, though the source(s) of the error not be definitively established. Confidence in the reliability and validity of the data was undermined, particularly with regard to data on mobile phone sharing. As a result, we undertook a second survey round with a probabilistically selected subset of the original respondents, to conduct reverse-record checks. The objective was to assess the magnitude and characteristics of misreporting. A sampling rate of 0.25 (n = 1,294) was applied to the original sample of respondents, retaining the stratified single-stage without replacement (stsrswor) design.

1.3 Instrument development

The survey instrument was developed by Flowminder's senior survey statistician, with input from the project PI. Modules included: individual socio-demographics, household composition and characteristics (including a condensed asset-based wealth index), and mobile phone usage and sharing. These modules were prefaced by a participant information script, an informed consent script and record of consent, and eligibility and identification fields. The survey was designed to take no more than 12 minutes to complete. Item and response code wording were aligned with validated 2016 Nepal Demographic and Health Survey instruments where feasible (bearing in mind the two surveys' different modes).

The module on mobile phone usage was developed following a review of prior research. The utility of past work to predict demographics from behaviour traces in de-identified CDR data was limited in this regard. The extent, characteristics, and implications of SIM sharing remain an (acknowledged) 'black box' issue for the field (Blumenstock et al 2015).

In developing the mobile phone usage module, we reviewed the peer-reviewed and grey literature on technological uptake in the Global South and the 'digital divide'. Surveys measuring mobile phone sharing were identified for Egypt (Samuel et al., 2005), Namibia (Stork, 2005), Botswana (Sebusang et al., 2005), Ethiopia (Lopez, 2000), Tanzania (Goodman, 2005), South Africa (Goodman, 2005), India (Rangaswamy and Singh, 2009), and the Philippines (Lopez, 2000). The literature review revealed substantial variation in questionnaires designed to measure ostensibly the same construct of interest (mobile phone sharing). In the absence of standardised, validated questionnaires - or question banks -, idiosyncratic tools have tended to be developed on a case-by-case basis. The identified studies exclude (discussion of) statistical

[4] See Vallient et al 2012: 177 for details of this approach and for comparison with alternative methods

[5] 'Hard refusals' were exempt from repeat calls.

assessment of the reliability and validity of the developed questionnaires, further limiting their scope to support development of standardised tools.

A parallel research stream investigating SIM sharing practices with the objective of improved representativeness for single frame mobile-phone-mode and dual frame telephone-mode surveys proved more promising. Here, the concern is with the development of respondent selection procedures, and the proper computation of inclusion probabilities and design weights, for the mobile-phone-mode sample (Busse and Fuchs 2013, 2014; Ghandour et al., 2019).

Research on inclusion probabilities for mobile-phone-mode surveys has focused on the US and Europe, and demonstrates that SIM sharing occurs even in contexts where individual mobile phone ownership is ubiquitous (Busse and Fuchs 2012a, 2012b). There is evidence, too, that SIM sharing may be more prevalent among subpopulations with above median incomes, suggesting that, at least in some settings, economic constraint is not the deciding factor (Busse and Fuchs, 2013). Published efforts to statistically validate questionnaires were limited to a few select studies within this literature.

The absence of standardised definitions, operationalisation, and measurement hampers comparisons between different surveys and in different settings, while also increasing scope for questionnaire-related error to inflate total survey error[6]. The complexity of the 'mobile phone sharing' construct amplifies the difficulties. Busse and Fuchs (2013), reflecting on their own and others' attempts to measure mobile phone sharing, call for new, "less burdensome and less difficult sharing questions". In developing the questionnaire, we prioritised simplicity and parsimony in our definition and operationalisation of 'mobile phone sharing', mindful that numeracy and literacy are subject to wide variations in the study setting.

Given the absence of standardised, validated instruments, we also imported insights from ethnographic and qualitative research, where mobile phone sharing has been studied in terms of (fictive) kin relations, social networks and connectivity, gift giving, and as artefacts implicated in the reproduction/subversion/ (re)enforcement of social norms. Within this literature, gender and generation or life-stage emerge as prominent sites of differentiation in mobile phone access and usage. There is evidence too, that sharing within close kin and friendship networks can be so ubiquitous as to be taken-for-granted, such that respondents do not consider joint use of mobile phones by their spouse or child to constitute 'sharing' (Wright-Steenson & Donner 2009). This literature provides further evidence on the heterodox drivers of sharing, noted above. A wide variety of mobile phone sharing practices are identified in which economic constraints are either absent, or partial factors (Wright-Steenson & Donner 2009). It must be noted here that the social dynamics of mobile phone use in Nepal have been little studied to date (research by Matinga et al., 2019) on the gender contours of mobile phone uptake and usage in two districts in rural Nepal, is a notable exception).

The broader literature on telephone-mode surveys cautions against the use of lengthy respondent selection procedures, on the basis that the increased respondent burden entailed

[6]Studies in which floor or ceiling effects are discernible, or where variance in responses is limited by a departure from standard item response anchors indicate the presence of questionnaire-related error.

can considerably reduce the response rate. This consideration was balanced by the need for the respondent to be knowledgeable about mobile phone usage and sharing. A single respondent selection question was included to determine if the person who answered the call, the 'gatekeeper', was a or the 'main user'[7]. Where s/he answered affirmatively the survey proceeded with the original respondent. Otherwise, a request was made to speak to a 'main user' if available. When a main user was unavailable, a call back was arranged for when s/he was expected to be present. Analysis of contact and response rates (Table 1.1, below) indicates that the inclusion of respondent selection items did not adversely impact contact and response rates.

We adopted an expansive definition of 'mobile phone sharing', consistent with the survey objective to ascertain gender for all persons using the SIM. Respondents were first asked to report the frequency (never, exceptionally, rarely, sometimes, often) of their own usage: 'How often do you use this mobile phone number to make or receive calls or send or receive text messages?'. Respondents were then asked: 'Does anyone other than you ever make or receive phone calls or send or receive text messages using this mobile phone number?' Where the respondent answered in the affirmative, s/he was asked to list each sharer. Response items were coded as gendered kin (e.g. wife / husband) and non-kin relations (e.g. employer neighbour, friend).

Where non-kin relationships were specified, a follow-up item recorded the sharer(s) gender. Sharing frequency (never, exceptionally, rarely, sometimes, often) was recorded for each person listed. In closing the mobile phone use module, all respondents (i.e. both those reporting sole use and those reporting shared use) were asked: 'If this mobile phone number rings when you are not in the vicinity, will the call be answered by another person?[8] (never, exceptionally, rarely, sometimes, often). Relationships were recorded as above in affirmative instances. The purpose of this latter question was to capture non-purposive 'sharing' episodes, which may not be regarded as 'others' usage' by respondents, but which are observable as CDR 'events'.
The choice of the five-item response scale to measure usage and sharing frequency is consistent with best practice for response anchoring. It is also consistent with measures used by the few surveys on mobile phone usage that have published results of statistical validation (Busse and Fuchs, 2012a, 2012b, 2013). Pilot data were analysed to assess variance in responses and confirm the absence of floor or ceiling effects. The selection of the 'ever' reference period similarly proceeded following testing and piloting of a variety of reference periods (including, 'in the past 24 hours', 'in the last seven days', 'in the last four weeks', and 'thinking about a normal day…').

The questionnaire was independently translated and back translated both prior to and subsequent to piloting (following revisions to item and response code wording based on analysis of pilot data). In addition to Nepali, the questionnaire was translated and administered

[7]We gave due consideration to instituting respondent selection mechanisms where a 'gatekeeper' reported that s/he was a main user and that the phone number was used in equal shares among a group of people, in order to reflect unequal inclusion probabilities. The decision was taken to address unequal probabilities of selection post-hoc, on the basis of the information collected on 'sharers' and to minimise respondent burden at the selection phase.

[8]This phrasing is adopted (in slightly adapted form) from a questionnaire developed by Busse and Fuchs (2012, 2013) to assess the extent and characteristics of 'mobile phone sharing' in Germany.

in Maithili, Bhojpuri, Tharu, Tamang, and Newari languages, to minimise the opportunity for systematic exclusion of minority language speakers.

1.4 Implementation

The study's partner MNO stipulated that the survey be implemented by its preferred market research firm, using CAPI (computer aided personal interviewing) software developed in-house. These conditions reflect the data's commercial sensitivity (section 2.4, below, details our technical and organisational data protection and privacy safeguards). The implementing firm had many years' experience conducting market research for the MNO, including conducting telephone-mode surveys with the MNO's customer base. Favouring quota sampling, the firm had only limited familiarity with representative survey design and implementation. For this reason, simplicity in the survey's design and implementation was favoured as far as a probabilistic sample design would allow.

The telephone-mode survey was administered by 16 enumerators based in a call-centre in Kathmandu. While a majority of the enumerators were existing employees of the implementing firm, three additional staff were hired to extend coverage of minority languages. Supervision was provided by management staff at the implementing firm. The MNO provided logistical support prior to and during the survey, as well as day-to-day survey coordination.

Flowminder's senior survey statistician delivered six days of training for enumerators and supervisors, and oversaw three days of survey piloting at the implementing firm's Kathmandu office in the fortnight prior to the survey launch. During this visit, procedures for survey implementation, and for daily data transfers were initiated and tested to enable the survey statistician to conduct monitoring, quality assurance, evaluation, and course-correction activities remotely over the course of the survey.

The survey took place over a period of six weeks, from the 13thof December 2017 to the 25thof January 2018. It was scheduled to fit with the implementing firm's existing commitments. Calls were conducted between 08:30 and 18:00, seven days a week.

Data were stripped of direct identifying features (including names and phone numbers) before being uploaded to the server for analysis. The procedure for daily uploads to permit remote monitoring and course correction was rendered unworkable by a series of server stoppages. The survey data were received in two batches in late January 2018, upon conclusion of the survey. In the absence of daily progress review, sample replicates continued to be released for several days after the desired sample size had been achieved. As a result, the final sample (n = 5,180) was inflated relative to the specified sample size (n = 4,268).

As noted above, preliminary analysis of the survey data identified non-trivial anomalies and internal inconsistencies indicative of response error and/or data entry error. A major source of concern was the scope for data entry error to occur owing to the need for the enumerator to manually enter the phone number dialled each time a survey unit was contacted (a result of the limited functionality of the CAPI software, preventing survey record linkage with the CDR). As

part of a broader root-cause analysis, a validation exercise was undertaken with 5% of the original respondents (n = 248). The validation survey took place over four days in early February 2018. The survey tool included a sub-set of six questions, repeated verbatim from the original survey and covering items with limited or predictable time variance (respondent sex, age, usual address, education level, and marital status). On the basis of further anomalies detected between the original survey responses and the validation survey, a follow-up round of data collection was planned with a 25% subsample (n = 1,294) of the original survey respondents.

The 25% cap was instituted in response to budgetary and logistical considerations. In light of the anomalies detected in the original survey data, the MNO agreed to the use of third-party CAPI software (with functionalities to limit scope for systematic data-entry error, including pre-entry of phone numbers to automate phone number assignment and de-identification, as well as hard-coding of survey routing and flags), and the presence of an experienced independent survey coordinator, tasked with directly managing and supervising survey implementation and conducting a parallel validation exercise for a 10% sub-set of survey units.

The timing of the follow-up survey was determined by the implementing firm's availability, given existing, competing, commitments. Following a series of postponements, data collection commenced 12 months after the initiation of the original survey exercise (December 2018 to January 2019), following five days of enumerator training and two days of piloting delivered by Flowminder's senior survey statistician. The questionnaire for the second round was streamlined to include only the socio-demographic and SIM sharing modules (the participant information script, informed consent script and record of consent, and eligibility and identification sections were retained). The second survey round was free from the anomalies and internal inconsistencies identified in the original data. Comparison of the two datasets could not conclusively establish the origins of the anomalies identified in the first round of data collection, however, meaning that analysis of SIM sharing was limited to the 2018/2019 subset of data.

1.5 Results

The survey obtained a 44% contact rate, 1,731 contacts were made from a total of 3,952 dialled calls. The response rate for contacted units was 84% (n = 1,461). Item missing data for key variables (including respondent gender) reduced the usable sample size to 1,280.

Table 1.1: Contact and response rates of the survey

Contacts	n	%
Total dialled	3,952	100%
Phone did not ring	1,637	41.4%
Phone rang but no one answered	584	14.8%
Total non-contact	2,221	56.2%
Contact rate	**1,731**	**43.8%**

Response	n	%
Total contacted	1,731	100%
Unstable / inaudible connection	93	*5.4%*
Respondent unavailable	41	*2.4%*
Age ineligible (<16 years of age)	21	*1.2%*
Consent withheld	93	*5.4%*
Language barrier	3	*0.2%*
Other	19	*1.1%*
Total unit non-response	270	*15.6%*
Response rate	**1,461**	***84.4%***

The sampling strategy produced a well-balanced sample, compensating for the under-representation of women in the sampling frame, via equal (i.e. 50:50) sampling of records across all strata. Subscribers with vacant gender fields (15% of records) were merged with the male listing prior to sampling (since proportionately more likely to be men) to avoid systematic exclusion of subscribers with no gender recorded. Oversampling produced a female:male ratio of 43:57[9].

The deviation from the 50:50 sampling rate is a result of the discrepancy between the registered and true user(s) of the SIM as well as the ubiquity of SIM sharing (discussed below). The survey data demonstrate the limited scope for administrative data held in the subscriber database to support demographic prediction models. Overall, the subscriber database-recorded-sex differed from self-reported main user sex for 32% of survey units (and was missing from the former in a further 7% of observations). 28% of female respondents who reported that they were a or the main user of the mobile phone number were recorded as male in the subscriber database (a further 5% had vacant sex fields). The equivalent figures for male respondents were 36% (recorded as female) and 8% (vacant sex fields).

Geographical stratification produced a rural urban ratio of 42:58 (based on self-reported 'usual residence')[10]. This compares with a rural urban ratio of 25:75 for the sampling frame (based on CDR analysis to assign 'home location'). The deviation from the sampling rate (50:50) can be attributed to the different definitions employed (usual residency vs mobile phone CDR activity). Timing (and time-lag) may also play a role (usual home location was calculated based on the tower locations of CDR events in the twelve months preceding the first survey round).

[9]This represents a slight deterioration of the gender ratio for the first survey round (46:54, n = 2,382 women and 2,796 men)

[10]A deterioration from the first survey round's rural:urban ratio of 46:53.

Table 1.2 presents the (unweighted) sample characteristics for the 2018/2019 survey data

Table 1.2: Summary statistics (unweighted 2018 -2019 data)

	Women (n = 555)					Men (n = 725)					p
	n	mean	SD	min	max	n	mean	SD	min	max	
Age (years)	555	32.6	11.9	16	89	*725*	*35.1*	*14.2*	*16*	*91*	*0.003*
Education (years)	555	6.3	4.6	0	11	724	8.1	3.7	0	11	0.000
Marital status (married)	555	0.79	-	0	1	725	0.70	-	0	1	0.000
Residence (rural)	540	0.45	-	0	1	*722*	*0.40*	-	*0*	*1*	*0.060*
Household size	554	5.1	2.5	1	24	*724*	*5.6*	*2.9*	*1*	*19*	0.001
Mobile phone sharing: Use frequency - self	527	3.5	0.6	0	4	*683*	*3.7*	*0.6*	0	4	0.000
Mobile phone sharing: Use frequency - others	527	1.3	0.5	0	4	*683*	*1.2*	*1.0*	0	4	0.285
Mobile phone sharing: Number of users	555	1.8	1.0	1	6	*725*	*1.7*	*0.9*	1	7	0.388

Shared SIM use was reported by 47% of respondents. 12% of respondents reported that one or more people of the same gender 'often', 'sometimes', or 'rarely' used the SIM to make or receive calls or send or receive messages. 35% of respondents reported mixed-gender use of the SIM 'often', 'sometimes', or 'rarely'. The remaining 53% of respondents reported either that s/he alone used the SIM (52.6%) or that others used the SIM only in exceptional circumstances (0.6%). There was no statistically significant difference in women's and men's reported tendency to share SIMs. 48% of women reported shared use compared with 46% of men (p = 0.453).

For both women and men, SIM sharing tended to be highly concentrated within the family. Just 8% of women and 8% of men who disclosed SIM sharing reported that a non-family member(s) used the SIM 'often', 'sometimes', or 'rarely'. Spouses and children dominated reports of sharing. 45% of women and 48% of men who disclosed sharing reported that their spouse used the SIM 'often', 'sometimes', or 'rarely'. 42% of women and 45% of men reported that their child(ren) used the SIM 'often', 'sometimes', or 'rarely'. Sharing between siblings (27% of women and 18% of men reported use by a sibling) was also relatively common.

Table 1.3 presents the results of regression analysis of gender-based variation in sim sharing. The stratified survey design is controlled for in the analysis, and design weights, calculated on the basis of secondary survey data, are applied. Reported standard errors are based on Jackknife estimation.

Table 1.3: Ordinal logistic regression, SIM usage and sharing frequency

	Frequency of sim usage - self			Frequency of sim usage - others		
	OR	Jackknife SE	95% CI	OR	Jackknife SE	95% CI
Gender (women)	0.141**	0.128	[0.024 0.833]	0.313	0.296	[0.049 2.000]
Age (years)	0.980**	0.010	[0.961 1.000]	1.006	0.008	[0.990 1.021]
Education (years)	1.083**	0.036	[1.016 1.156]	1.017	0.034	[0.953 1.085]
Marital status (married)	0.989	0.325	[0.519 1.884]	1.592	0.485	[0.875 2.895]
Household size	0.978	0.041	[0.900 1.062]	1.019	0.043	[0.938 1.106]
Usual place of residence (rural)	1.465*	0.295	[0.988 2.174]	0.894	0.153	[0.640 1.250]
Woman*Age (years)	1.023	0.016	[0.993 1.055]	1.008	0.015	[0.978 1.038]
Woman*Education (years)	1.054	0.049	[0.961 1.155]	1.014	0.048	[0.924 1.112]
Woman*Marital status (married)	1.672	0.894	[0.586 4.772]	1.523	0.753	[0.577 4.018]
Woman*Household size	0.996	0.066	[0.875 1.134]	1.100	0.075	[0.962 1.257]
Woman*Usual place of residence (rural)	0.706	0.204	[0.401 1.244]	1.426	0.388	[0.836 2.432]
/cut1	-5.635***	0.867	[-7.336 -3.934]	1.037*	0.580	[-0.101 2.175]
/cut2	-4.620***	0.721	[-6.035 -3.205]	1.060*	0.580	[-0.079 2.199]
/cut3	-3.496***	0.625	[-4.722 -2.271]	1.333**	0.585	[0.186 2.480]
/cut4	-1.065*	0.569	[-2.181 0.052]	5.296***	0.732	[3.859 6.733]
n = 1,189	F(11, 1178) = 5.83			F(11, 1178) = 1.81		
	Prob > F = 0.000			Prob > F = 0.0481		

The analysis demonstrates statistically significant gender-based variation in frequency of own sim usage (with women's use significantly less frequent than mens), controlling for age, education, marital status, household size, and rural residency. There is, however, no statistically significant difference in the frequency with which women and men report others' use of the phone number.

2. Mobile network operator data

Large-scale spatio-temporal data collected by the study's partner MNO supported both of the project work packages. A series of CDR feature sets were generated from the 'raw' CDR data to produce:

1. User profiles hypothesised to be predictive of gender
2. Geospatial layers hypothesised to be predictive of human presence

2.1. Data Pipeline Management

A high-level overview of the Flowminder's secure end-to-end analytics process is presented in Figure 2.1.

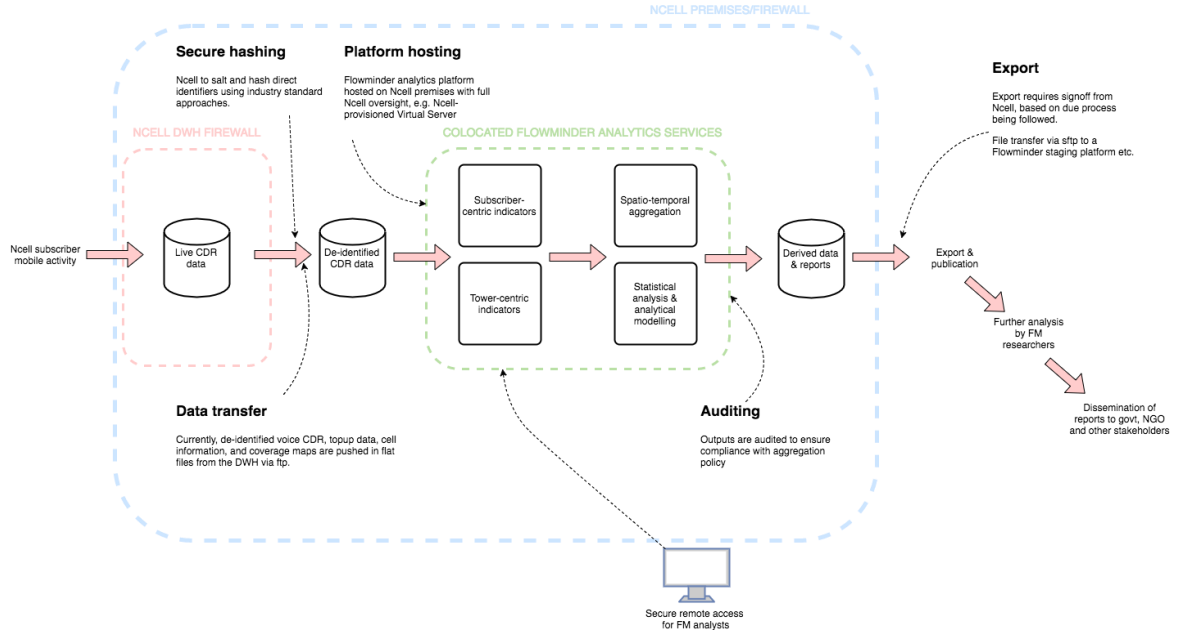

Figure 2.1. Flowminder analytics end-to-end process

Flowminder jointly manage a remote-access analytics platform hosted on Ncell premises in Kathmandu. The platform provides a secure environment for the processing of pseudonymised Call Detail Records (CDRs), together with supporting datasets, under an agreed set of data governance and collaboration principles.

2.2. Data Extraction and Preparation

The following MNO datasets were prepared with the assistance of Ncell staff, and quality checked, prior to ingestion into the Flowminder analytics platform.

Pseudonymised voice call records: containing the time, duration, location, and parties of each voice call.

Pseudonymised 'top-up' records: containing daily financial credit 'top-up' totals and counts per recharge type per subscriber.

Cell attribute tables: containing the geolocation of individual cells in the network.

Subscriber gender-of record: containing, where available, the gender of the subscriber as recorded on the Registration Form submitted on purchase of a SIM card. The registration form[11]

[11]Available here: https://www.ncell.axiata.com/Upload/forms/Individual%20Subscription%20Form.pdf

provides for three responses to 'gender' (with response codes corresponding to sex): 'male', 'female', and 'other'.

2.3. Supporting Primary Survey Sample Selection

The CDR data was first processed to support the selection of a representative subset of subscribers for the primary survey:

1. Active subscribers were determined as those making or receiving at least one call during the month of November 2017. This strategy was employed to minimise the inclusion of inactive subscribers.

2. Active subscribers' 'home municipality' (Administrative Level 2) was determined over the entire year of 2016. The home municipality was defined as the modal location of the last call of the day made or received by a subscriber over the course of 2016.

3. Ncell subscriber database records were used to split the remaining subscribers into male and female listings. Subscribers where gender was recorded as missing (15%) were merged with the male listing, as discussed above.

4. Active female and male subscribers were grouped by home municipality, based on:

    a. Province, as presented in Figure 2.2
    b. Ecological zone, as presented in Figure 2.3a
    c. Urban - Rural designation, as presented in Figure 2.3b

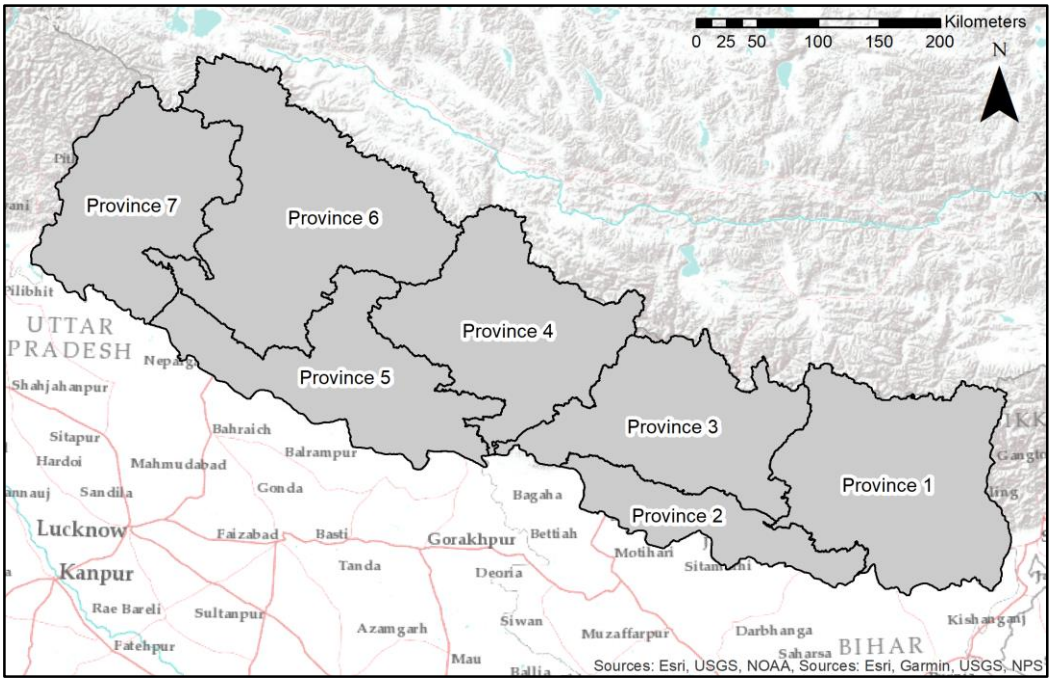

Figure 2.2. Map of Nepal at province level.

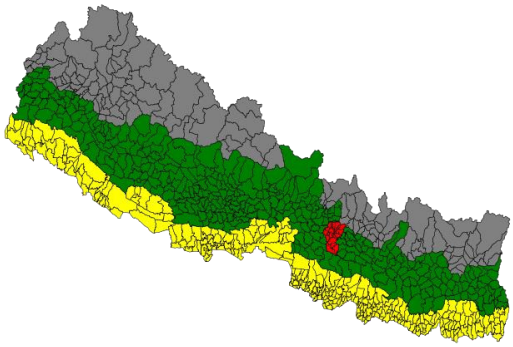

Figure 2.3a. Nepal ecological zone designation by municipality. Mountain [grey], Hill [green], Terai [yellow], Kathmandu Valley [red].

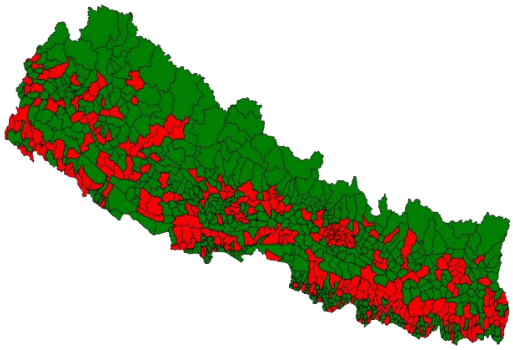

Figure 2.3b. Nepal urban - rural designation by municipality. Urban [red], Rural [green].

2.4. Supporting Privacy-Preserving Data Linkage

The development of CDR-based gender-prediction models relies upon individual-level linkage of primary survey data and CDR data for the subset of subscribers participating in the telephone-mode survey. This requires the institution of a range of technical, organisational, and protocol-based safeguards to maintain individual's privacy and insure their data is protected.

Processing of the project data is distributed across a triad of stakeholders, namely SoftTech (primary survey implementation), Ncell (CDR) and Flowminder (analytics). A privacy-preserving data linkage protocol was implemented, whereby SoftTech only have access to the survey data for the period of the survey, Ncell only have access to CDR data, and Flowminder only have access to linked de-identified phone survey records and CDR data. The protocol consists of key-matching tokens being provided by Flowminder to Ncell alongside the pseudonymised phone numbers of the survey participants. By propagating the tokens alongside the data records exchanged between the three parties, Flowminder could successfully match the de-identified survey and CDR records of the survey participants:

Provided by Ncell to Flowminder	HASHED MSISDN	HASHED IMEI	CELLID	DURATION	TAC	...
Provided by Flowminder to Ncell	HASHED MSISDN	MSISDN TOKEN				
Provided by Ncell to Soft Tech	MSISDN	MSISDN TOKEN				
Provided by SoftTech to Flowminder		MSISDN TOKEN	SEX	PHONE SHARING	SOCIO-DEMOGRAPHICS	...

2.5. Subscriber Profile Extraction

A set of subscriber profiles to support the gender prediction modelling was extracted from voice call and top-up data for 2016 using the following strategy:

1. Daily Locations of subscribers were extracted for each day of the year. Daily locations were defined as the location of the cell to which a subscriber was connected for their last call of the day.

2. Home locations of subscribers were calculated for each month. Home locations were defined as the modal daily location of a subscriber over a month.

3. The set of voice call features listed in Table A2.1 (appendix A) was extracted for each subscriber for each week of the year, and the mean and std were calculated

4. The set of topup features listed in Table A2.1 (appendix A) was extracted for each subscriber for each month of the year, and the mean and std were calculated

2.6. Geospatial CDR Feature Extraction

Additionally, a series of geospatial features were computed to support work package one (high spatial resolution estimation of gendered development indicators), features were derived from the voice call data for each month of 2016. The monthly aggregates were extracted using the following strategy:

1. Daily Locations of subscribers were extracted for each day of the month. Daily locations were defined as the location of the cell to which a subscriber was connected for their last call of the day.

2. Home locations of subscribers were calculated for each month. Home locations were defined as the modal daily location of a subscriber over the month.

3. The set of voice call features listed in Table A2.2 (appendix A) was extracted for each subscriber for each month of the year. Each feature was spatially aggregated using the subscriber Home Location.

4. The set of device features listed in Table A2.2 (appendix A) was extracted for each subscriber using the subscriber's most used device. Each feature was spatially aggregated using the subscriber's Home Location during each month of 2016.

3. Georeferenced DHS Indicators

3.1. Secondary survey data: 2016 Nepal Demographic & Health Survey

The indicators selected for high resolution mapping were estimated using data from the most recent round of the Nepal Demographic and Health Survey (NDHS) (MOH et al, 2017). The Demographic and Health Survey is an international quinquennial cross-sectional survey programme developed and coordinated by ICF International and funded by USAID. It is a multi-topic survey, with modules covering: demographics; household composition and characteristics; livelihoods, migration, and employment; reproductive, maternal, and child health; family planning; nutrition; communicable and non-communicable diseases; gender equity; and gender-based violence (GBV) targeting women. Additionally, and importantly for this study's aims, the 2016 NDHS provides:

1. Recent, geo-located, nationally representative household and individual survey data
2. Sex disaggregated data for key indicators
3. Data suitable for gender statistics (beyond sex-disaggregation)

DHS data are available for download upon submission and approval of a registered use-case. The most recent round of the DHS survey in Nepal was conducted in 2016. The survey was implemented face-to-face, using CAPI (computer aided personal interviewing), by the Nepali survey organisation New Era, with support from the Nepali Ministry of Health (MOH et al., 2017).

The 2016 NDHS employed a stratified, clustered design, with the preliminary sampling frame drawn from the 2011 Nepal Population and Housing Census. The 2016 NDHS was designed to be representative of the Nepali population at three sub-national levels, namely: ecological zone (Terai, Hilly, Himal), development region (N = 5), and province (N = 7). Design weights are available to account for the complex sample design for national and province level findings. Full details of the NDHS sample design are available in the survey documentation (MOH et al., 2017). For brevity, we limit discussion of the sample design to its pertinence for the present study. Briefly, the population was divided according to rural/urban residency within each of seven provinces, to create 14 strata. A total of 383 primary sampling units (PSUs) were selected within strata, with probability proportional to size (PPS). In rural areas, a two-stage design was adopted, with wards forming the PSUs. In urban areas, where wards are excessively large, a three-stage design was adopted, with enumeration areas (EAs), sampled within selected wards, forming the PSUs. Wards (rural) and EAs (urban) with 200 households or more, were further subdivided such that the PSU corresponds to a portion of the ward / EA rather than its entirety.

Following a manual listing phase, 30 households were selected within each PSU, resulting in a total sample of 11,490. Within each selected household, female household members and overnight guests aged 15 – 49 were eligible to complete the individual questionnaire. Within a 50% subset of selected households, male household members and overnight guests were eligible to complete the individual questionnaire. Response rates were high for both the

household (n = 11,040, 96% response rate) and individual questionnaires (12,862 women, 98% response rate, and 4,063 men, 96% response rate).

The 383 clusters selected as primary sampling units (PSUs) provide the data points used in the high-resolution geospatial analysis, presented below. PSUs (wards, EAs, or a segment thereof) were geolocated (i.e. assigned longitude and latitude coordinates). The GPS location data correspond to the cluster centroid. Household and individual observations are nested within clusters via unique cluster ids. In order to preserve respondents' anonymity, the DHS programme displaces cluster locations by as much as 5km for rural cluster and 2km for urban clusters. 1% of (undisclosed) rural locations are displaced by up to 10km, to prevent reidentification (Burgert et al., 2013; ICF, 2012. Cluster displacement is controlled for in the geospatial modelling, following published recommendations (Perez-Heydrich et al., 2013).

The study team undertook a review of the academic and grey literature and held a series of meetings with senior officials at the Nepal Central Bureau of Statistics (CBS) in order to refine the selection of indicators. Seven indicators were initially considered for inclusion in the study, based on their availability in the DHS data and their relevance for gender equitable development policy in Nepal:

1. Literacy (sex disaggregated)
2. Educational attainment (sex disaggregated)
3. Market labour participation (sex disaggregated)
4. Agriculture-based occupation (sex disaggregated)
5. Stunting in childhood (sex disaggregated)
6. Attitudes to gender-based violence (GBV) against women (sex disaggregated)
7. Births in health facilities

The initial group of indicators was generated in accordance with standard international definitions (Detailed definitions are provided in appendix A, table A3.1).

Figures 3.1 - 3.7 below, demonstrate the variation in the seven indicators by province and sex. The Province level estimates take account of the complex survey design and incorporate design weights.

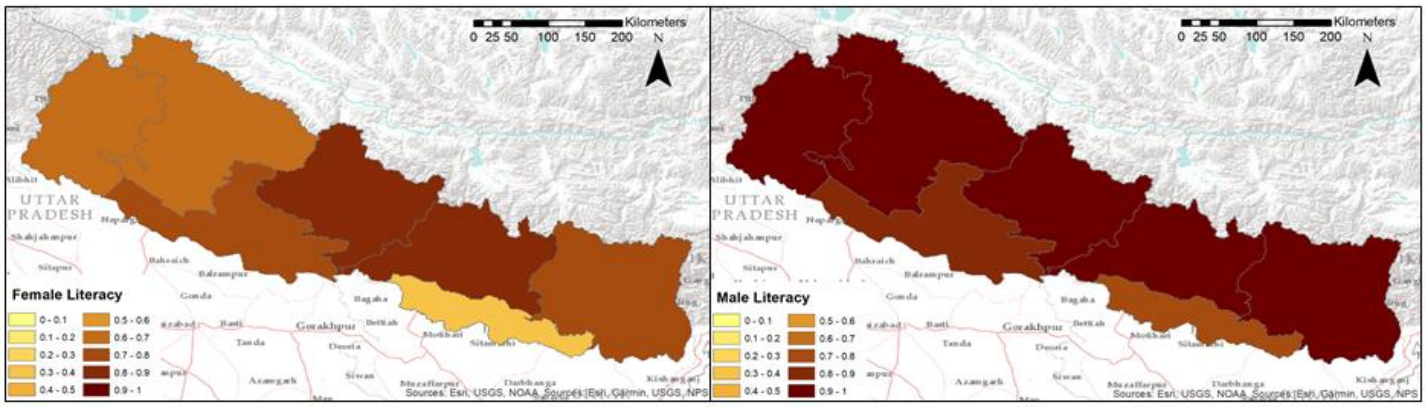

Figure 3.1. Maps of female literacy (left) and male literacy (right) by decile at province level. Both the maps derived from DHS survey data.

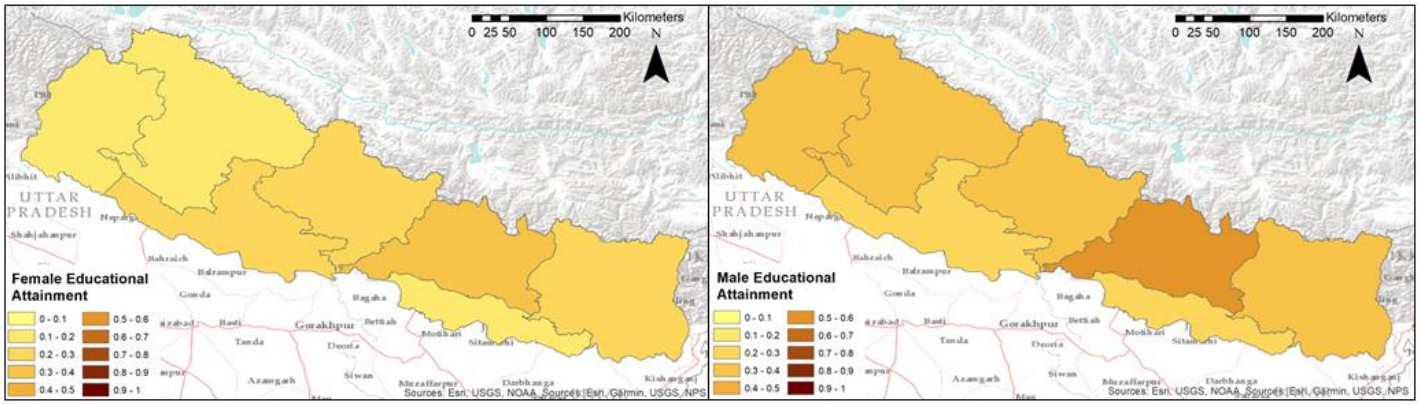

Figure 3.2. Maps of female educational attainment (left) and male educational attainment (right) by decile at province level. Both the maps derived from DHS survey data.

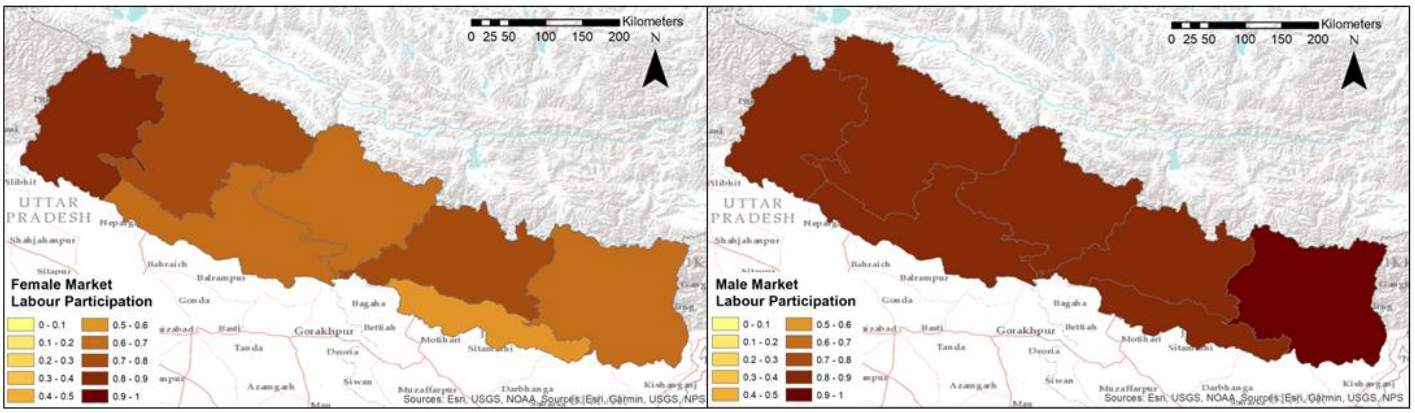

Figure 3.3. Maps of female market labour participation (left) and male market labour participation by decile (right) at province level. Both the maps derived from the DHS survey data.

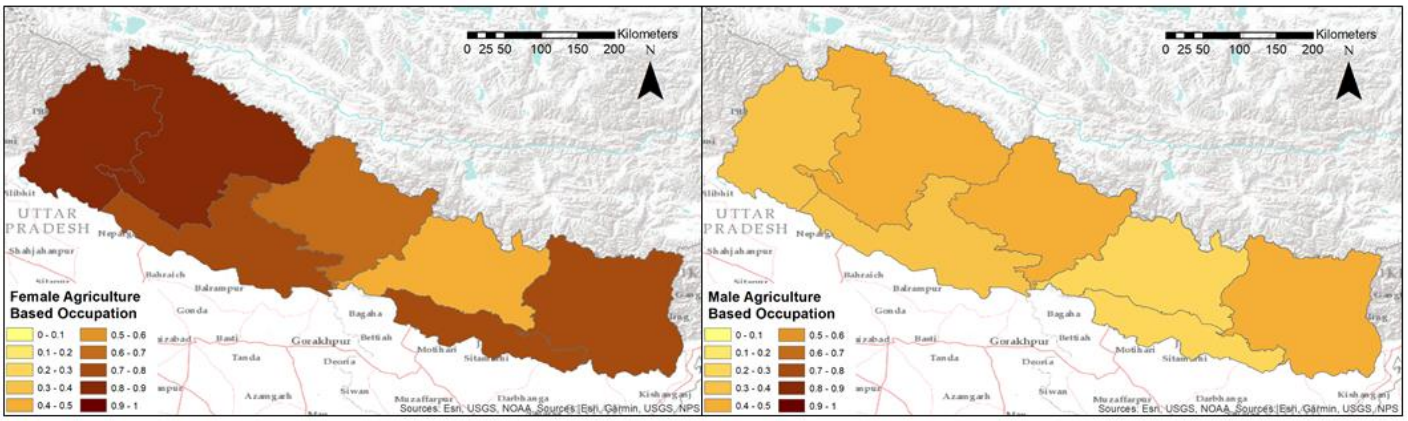

Figure 3.4. Maps of female engagement in agriculture (left) and male engagement in agriculture (right) by decile at province level. Source: Authors analysis of weighted NDHS 2016 survey data.

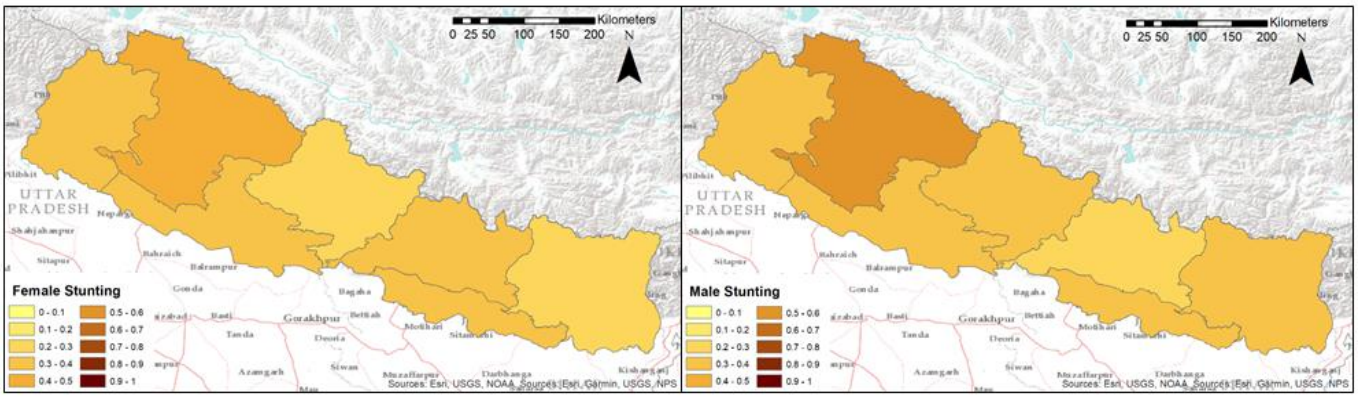

Figure 3.5. Maps of female stunting in childhood (left) and male stunting in childhood (right) by decile at province level. Source: Authors analysis of weighted NDHS 2016 survey data.

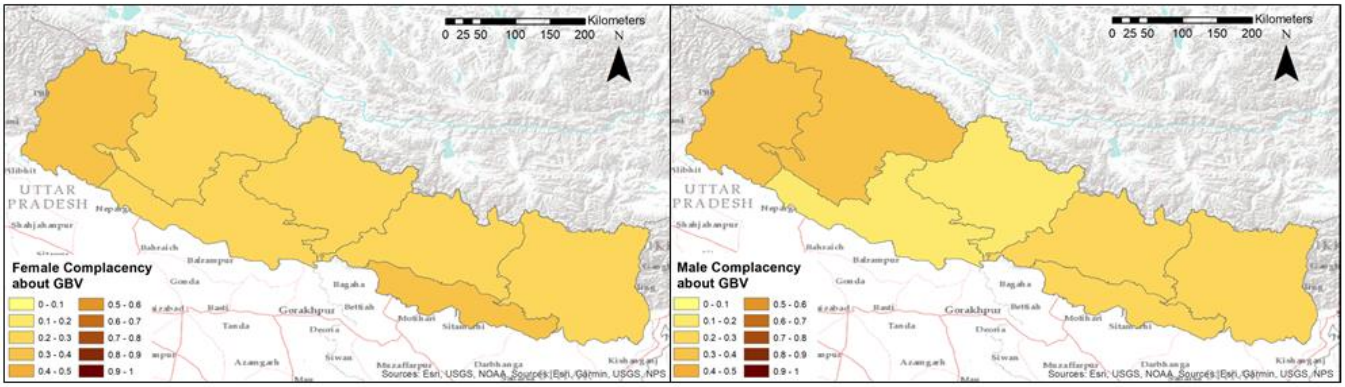

Figure 3.6. Maps of female complacency about GBV against women (left) and male complacency about GBV against women (right) by decile at province level. Authors analysis of weighted NDHS 2016 survey data.

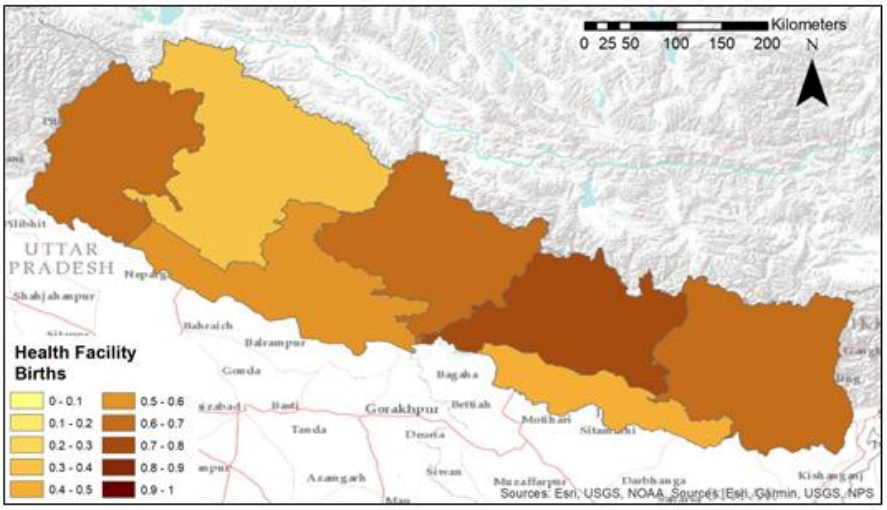

Figure 3.7. Map of health facility births by decile at province level. Authors analysis of weighted NDHS 2016 survey data.

Figures 3.1-3.7 show province level variations in the selected indicators for men and women. Province level variation is pronounced for some indicators. Where differences between

provinces are less marked (for example, women's and men's attitudes to GBV against women), they remain statistically significant ($p < 0.05$, Kruskal Wallis tests).

Standard inferential statistics can direct attention – and resources – to sub-national variations and inequalities at the province level. However, while design effects originating in the clustered sample design can be measured and controlled for, and within and between cluster variance decomposed, within cluster samples are not designed to support inferential analysis at lower geographical scales.

In contrast, spatial interpolation methods offer a means to exploit the available geolocated data at cluster level in order to refine the geographical scale of estimates. Three of the seven shortlisted indicators were selected for high resolution mapping. They are: Literacy, agriculture-based-occupation, and health facility births. Individual responses were aggregated to generate counts and proportions corresponding to each indicator for the 383 geolocated clusters (PSUs). Cluster level estimates are unweighted (design weights are not applicable at the cluster level).

4. Geospatial Covariate Layers

4.1. GIS Covariate Layers

A range of geospatial datasets were collated from open-source platforms and processed to support the high-resolution mapping of the selected development indicators. Datasets included physical (topography, climate, land-cover), social (population counts, ethnicity), and built-environment (urbanisation, human settlements) elements. A table listing the datasets and further details on those selected are outlined in Appendix A. Selection was driven by past research on correlation between geospatial data and the selected development indicators (Bosco et al., 2017).

The geospatial data were transformed (when required) into continuous raster format, with matching extents of Nepal, and consistent spatial resolution and coordinate systems. Due to the varying resolutions and extents of the source data, raster data underwent re-projecting and resampling using Geographic Information Systems (GIS) software ArcGIS. Re-sampling interpolation techniques varied depending on the original resolution and category of the datasets. This pre-processing was carried out to produce the final layers, each with a spatial resolution of 0.0083 decimal degrees (approximately 1km at the Equator), in geographic coordinate system WGS84. Original datasets were downloaded in either raster or vector format, where vector datasets (e.g. protected areas, ethnicity) were transformed into discrete rasters with the variable of interest being allocated a value of one. Categorical raster datasets were reclassified to contain one class of interest; a continuous dataset was produced from these by smoothing or calculating the distance from the feature of interest. For Openstreetmap data such as roads, schools, rivers and other location data (such a health facilities), distance was calculated from the locations to produce the continuous surface. More information on each dataset and the methods used to produce the covariates can be found in Appendix A. Figure 4.1, below, provides examples of the geospatial covariates included in the later analysis.

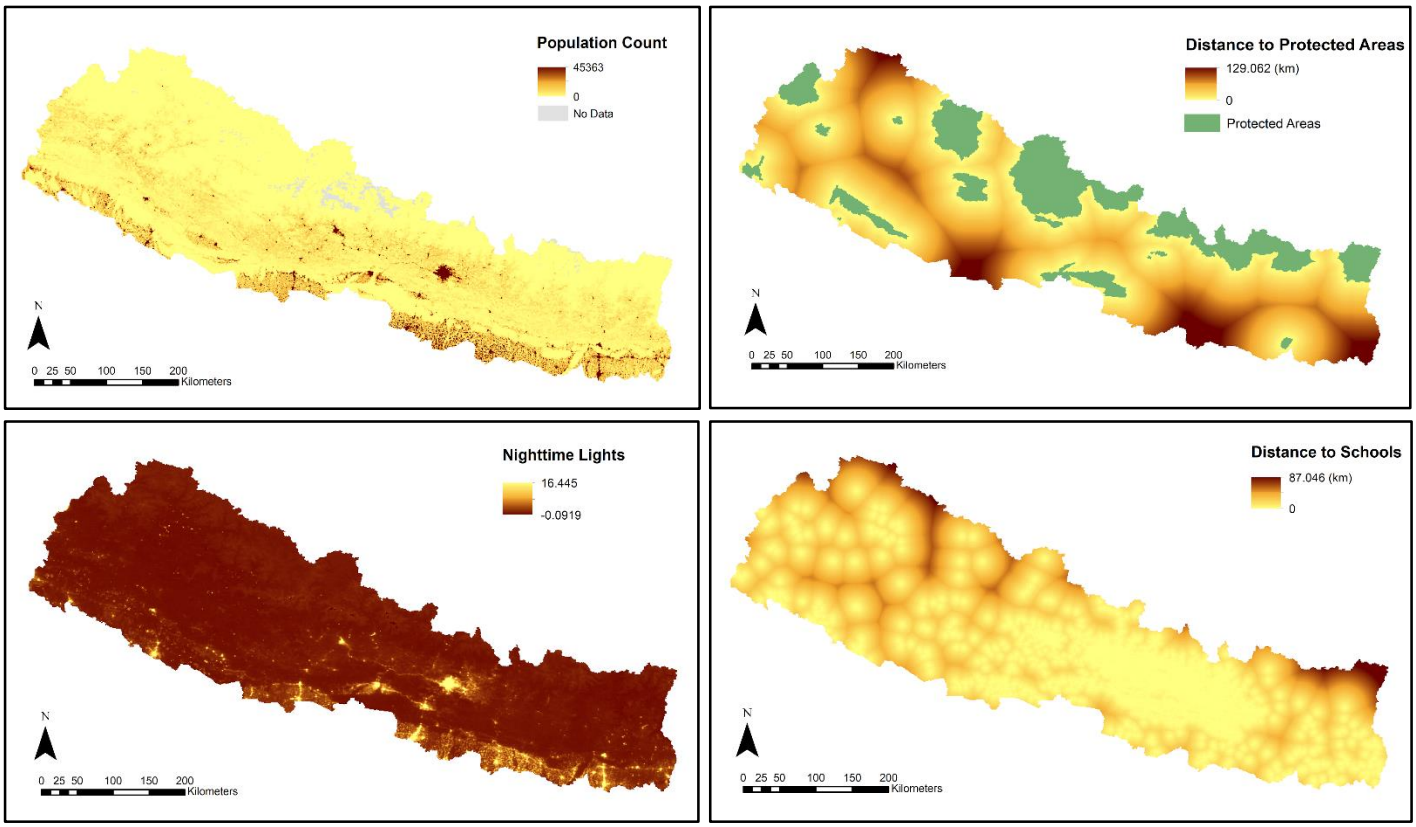

Figure 4.1. Geospatial covariates at 0.0083 decimal degrees resolution (approx. 1km at the Equator) showing (clockwise from top left) population count, distance to protected areas, distance to schools and nighttime lights.

4.2 Remote Sensing Covariate Layers

In addition to the GIS covariates detailed above, we created a set of seven covariates from remotely-sensed (RS) satellite datasets, downloaded from the United States Geological Survey (USGS). The datasets included three Moderate Resolution Imaging Spectrometer (MODIS) products:

1. The MOD13Q1 vegetation indices product, where two vegetation indices were extracted to provide a measure of live vegetation, namely, the Normalised Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), which has higher sensitivity in dense areas. The MIR (middle infrared surface reflectance) band was also extracted.
2. The MOD17A2H for the Gross Primary Productivity (GPP) and MOD17A3H for Net Primary Productivity (NPP), both measures of vegetation productivity.
3. The MOD16A2 for the total evapotranspiration (ET) and the total potential evapotranspiration (PET).

Further details on all the products and datasets can be found in Appendix A.

An R-script was created to download all of the data for each product between the years December 2009 - January 2017 (the NPP was available from December 2009- January 2014). Data was delivered as HDF-EOS format files, with separate tiles for the selected study area. Using R, these tiles were first transformed from HDF format to .tif files. The products were then

split into the seven indices selected for use along with each of their respective quality assurance layers. The information in the quality assurance (QA) layer and valid range values were used to remove contaminated pixels originating in sensor effects such as different orbits, adjacency, band quality, and MODLAND QA, and non-sensor effects such as cloud state and atmospheric noise (atmospherically corrected and clear cloud state). For each of the indices, summary statistics including the maximum, minimum, mean, median, sum (cumulative value) and sum per year (cumulative value per year) were computed for the specified period on the .tif tiles.

The tiles for each index and statistic were mosaicked together in ArcGIS to produce seamless coverage of Nepal. As the original datasets were in 250m or 500m resolution in a Sinusoidal coordinate system, they were reprojected to match the geospatial coordinates at a resolution of 0.00083 decimal degrees (approx. 100m at the equator) in GCS WGS 84. Nearest neighbour interpolation was applied during reprojecting and resampling. To produce a final layer, the data was clipped to match the extent of the other covariates. Rasters with large areas of no data values due to contamination were replaced with a value of 0. Figure 4.2, below, shows example RS covariates.

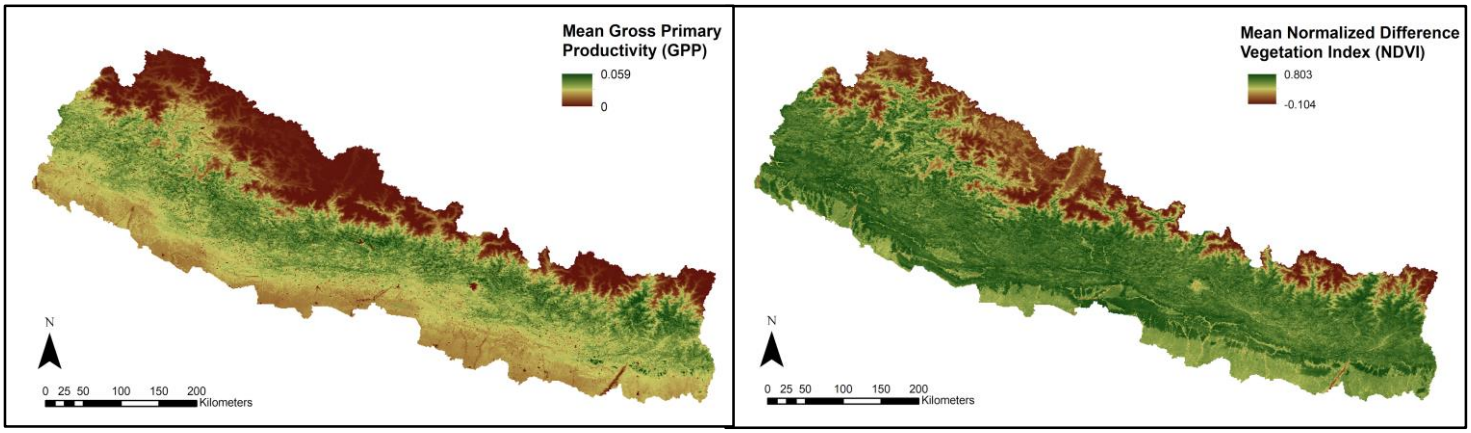

Figure 4.2. Remote sensing covariates showing the mean Gross Primary Productivity and mean Normalised Difference Vegetation Index.

4.3 CDR Covariate Layers

The CDR features described above, and presented in table A2.1, were produced as covariate layers for inclusion in the geospatial analysis. The CDR layers were produced to match the extent and spatial resolution (0.0083 decimal degrees) of the GIS and remote sensing (RS) layers. Using the latitude and longitude of each cell tower, a point dataset was created in ArcGIS to map the tower locations. These points were used as the input to create a set of voronoi polygons where the polygon size is determined based on the equidistance from each point location. The voronoi polygons provide an estimation of cell tower coverage, with each polygon representing coverage for the cell tower point within it. Due to the addition or removal of cell towers over time, a new set of voronoi polygons was calculated for each month of CDR data. Additionally, a second version of voronoi polygons was created with a buffer of 35km applied around each cell tower location, to provide an estimation for maximum coverage. This resulted in areas in the north of the country being omitted, as cell tower coverage is lower in areas that

have low population counts. To create continuous coverage, the buffers were dissolved and manually smoothed, Figure 4.3 presents both sets of voronoi polygons for January 2016.

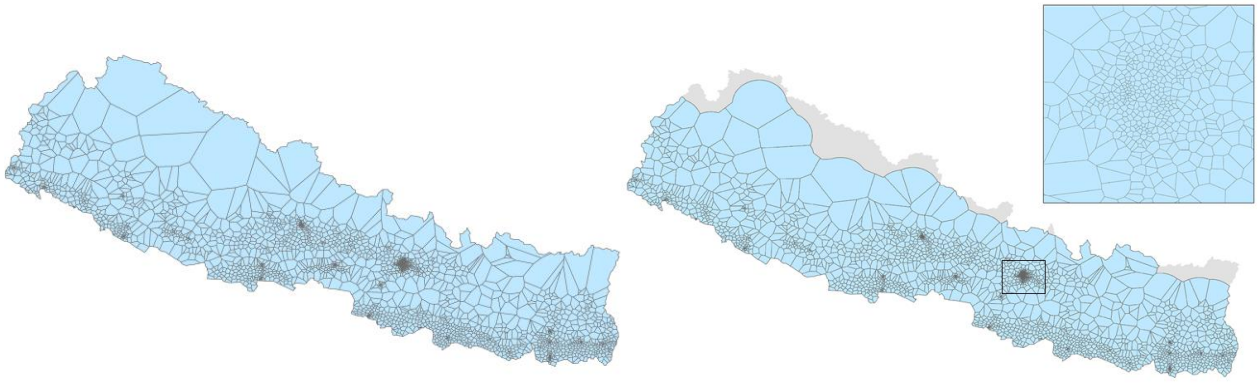

Figure 4.3. Voronoi polygons for Nepal (left) with voronoi polygons clipped at a 35 km radius (right) with inset of Kathmandu.

To map the CDR feature values, data for each month of the study reference period was joined to the voronoi polygons for each cell tower location. An ArcPy script was applied to convert the polygon layers to raster for each individual feature, with the same extent and in the same coordinate system (GCS WGS 1984) as the geospatial layers. The spatial resolution was selected as 0.00083 decimal degrees (approximately 100m at the equator), to preserve smaller voronoi polygons located in highly concentrated urban areas that would otherwise be covered by a 1km cell size. These layers were produced for each feature for every month of 2016, mapped to the voronoi polygons. Figure 4.4 shows the median of the incoming call counts mapped at 100m resolution for January 2016.

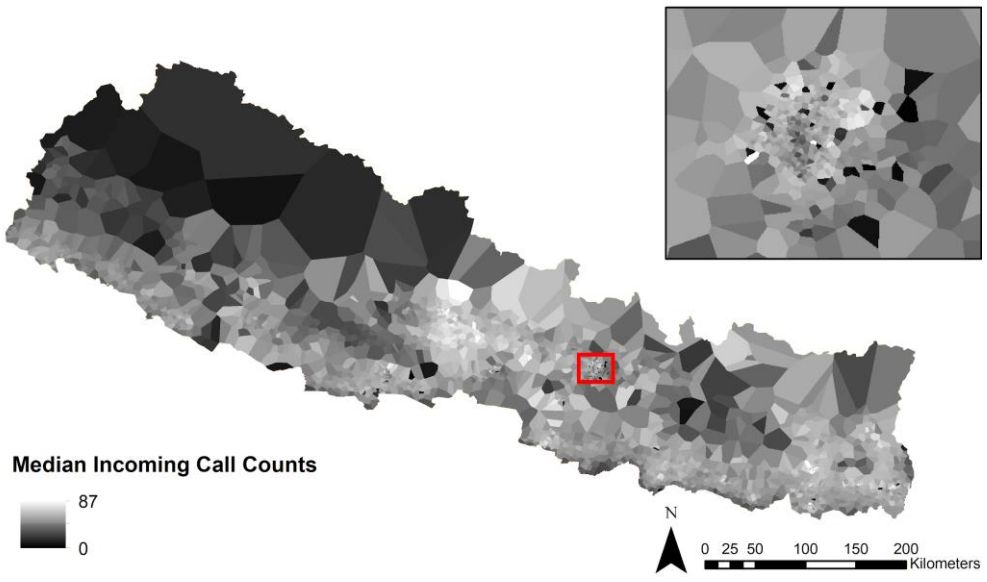

Figure 4.4. Median incoming call counts for January 2016 at 0.00083 decimal degrees (approx. 100m at the equator) mapped to the voronoi polygons derived from cell tower location.

The final set of CDR layers needed to be at a spatial resolution of 0.0083 decimal degrees (approximately 1km at the Equator) to match the geospatial covariates, the 100m raster layers

were aggregated to this cell size. The average of the selected CDR feature had to be calculated to account for multiple smaller voronoi polygons than would be included in a 1km grid cell, and those areas where a 1km grid cell would overlap with the boundary of neighbouring polygons. The values were aggregated by weighting the features by the number of subscribers at each tower location.

The number of subscribers per 100m grid cell was calculated and the result multiplied with each CDR feature 100m raster. Within every 0.0083 decimal degrees (approx. 1km) grid cell, the sum of this output was calculated, along with the sum of the number of subscribers. These were then divided to produce a CDR feature value for each pixel, which is based on the number of subscribers within the different voronoi polygons. This was calculated for all of the layers for each month, to produce a 0.0083 decimal degrees (approx. 1km) resolution set of layers for the twelve months (figure 4.5). An annual average was calculated whilst taking into account the number of days in each month, giving a final set of CDR covariate layers for each feature, for the year of 2016.

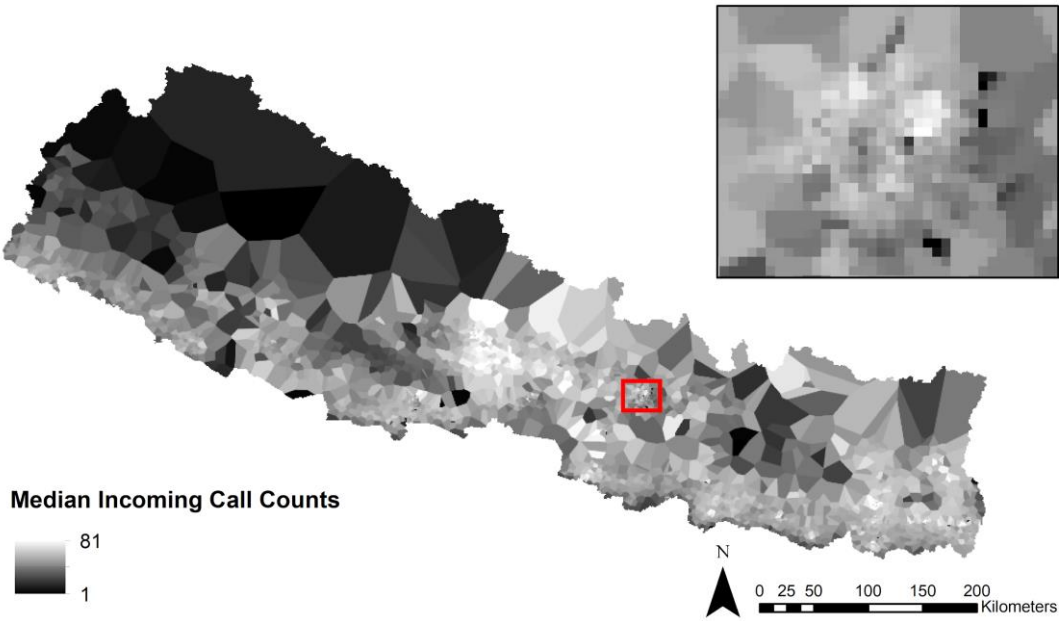

Figure 4.5. Median incoming call counts for January 2016 aggregated to 0.0083 decimal degrees (approx. 1km at the equator).

5. The Applied Methodology

5.1. the modelling architecture

Machine learning (ML) techniques and Bayesian Geostatistical (BGS) models were applied to build high resolution, sex-disaggregated maps for the selected development indicators (work package one) and to predict user gender for the CDR dataset (work package two)

A modelling architecture based on Artificial Neural Networks (ANNs) was used to split the Ncell database into users predicted to be women and those predicted to be men, and to provide the

necessary data to create sex-disaggregated mobility and migration maps. This work also aims to improve on previously applied disaggregation efforts[12], by investigating the impact of uncertainties introduced by SIM sharing on gender prediction models.

ANNs were also applied, this time in combination with BGS methods, to construct sex-disaggregated high spatial resolution maps of the selected development indicators (literacy, engagement in agriculture and births at health facilities in Nepal). Both methods utilised the geolocated surveys and gridded spatial covariate layers. The applied methodology allowed us to also estimate associated metrics of model uncertainty, highlighting the areas where predictions can be treated with greater confidence, versus areas where there is higher uncertainty due to conflicting data from surveys or poor explanatory power from the available covariate layers.

Each of the applied modelling architectures integrated Semantic Array Programming (SemAP) (de Rigo, 2015, 2012a, 2012b) and geospatial tools (ArcGis (ESRI, 2015) and QGIS (QGIS, 2018)) through the Geospatial Semantic Array Programming (GeoSemAP) paradigm (de Rigo et al., 2013, de Rigo, 2015).

Semantic array programming was developed to ease the integration of various conceptual modelling-units by formulating them as data-transformation modules or models (D-TMs). D-TM units do not force a user to master their internal details, since they exclusively exchange data, with broadly supported formats. SemAP is designed to ease the computational communication between local-contexts, different expertise, and disciplines in a simple way, while remaining compact and unambiguous. This is achieved by limiting the potential generality of the exchanged data by means of array-based semantic constraints (de Rigo, 2015; Bosco and Sander, 2015).

GeoSemAP exploits Semantic Array Programming and geospatial tools to split a complex D-TM into logical blocks. By applying mathematical and geospatial constraints, the reliability of these logical blocks can be more easily assessed. Constraints take the form of precondition, invariant and postcondition semantic checks and enable complex wide-scale transdisciplinary models to be described in terms of simpler GeoSemAP blocks (de Rigo, 2015).

In order to support the replication of methods and results, only free software tools and libraries (Stallman, 2009), and freely available datasets, were utilised in the geostatistical modelling techniques presented here. Similarly, the techniques used to apply all models and submodels within the modelling architecture developed for the study, are designed to be fully reproducible.

5.2. Artificial Neural networks

Artificial Neural Networks are based on an architecture inspired by the human brain. An ANN consists of a system of interconnected nodes. Information propagates through the nodes, transforming the inputs in intermediate derived signals to generate the final outputs, with the number and type of nodes that can be modified to perform the analyses.

[12](e.g. Jahani et al., 2017, Frias-Martinez, et al., 2010)

Internal nodes define the neural network hidden layers and are called neurons. Each of these nodes is a processing element that propagates weighted inputs received from other nodes (Pradhan and Lee, 2009).

The learning process that happens once the network is built, consists of iteratively adjusting weighted connections between neurons by comparing the modelling output with the calibration data. The output of a neural network, after a successful training, is a model that predicts a target value from input data (Lee, 2007).

Artificial Neural Networks can take advantage of complex relationships between covariates and output data because the relationship between nodes need not be linear, or even continuous. At the same time, data affected by large noise/signal ratios (or poor correlation with the desired target) may still be exploited by neural networks in their simplified near-linear relationships. Although ANNs can easily discover the principal components (e.g. linear) of covariation even with factors having a limited prediction capacity, the nonlinear components of relationship can also be exploited, when the unavoidable covariate-output noise allows these components to be numerically detected.

Multilayer feed-forward networks (multilayer perceptrons) form the basis of many ANN applications (Bosco et al., 2013, de Rigo et al., 2001, Secomandi, 2000), due to their universal approximation properties. Theoretically, a properly designed and trained perceptron is able to reproduce any relationship between the quantity to be modelled and the covariates (Kreinovich, 1991; Hornik et al., 1989).

In a feed-forward neural network, connections between the different units do not form a cycle or loop with information moving in only one direction (from input to output nodes and passing through the hidden layers (if any)). The simplicity of the theory, the ease of programming and the consistently good results are the main reason for using this type of Neural Network.

In this study we applied a feed-forward neural network to create the high resolution maps and to predict user's gender from the CDR dataset. We implemented a feed-forward multilayer perceptron by exploiting the Neural Network Package (Schmid, 2009) In MATLAB language, using GNU Octave (Eaton, 2008).

5.3. Bayesian Geostatistical Models

The Bayesian modelling approach encompasses a suite of statistical techniques which utilise the Bayesian method (Press 2002) to estimate the parameters of a posterior distribution. It is well-suited to modelling large datasets containing spatial information, owing to its flexibility, and its ability to accommodate correlation and hierarchical structure in the data, and to take account of uncertainty.

In the present study, we applied Bayesian geostatistical approaches to generate posterior estimates, exploiting the spatial and temporal covariance present in the data, and in the relationships with covariates (Banerjee et al., 2000). We used the Integrated Nested Laplace

Approximations (INLA) approach, available in the GNU R (R development core team 2014) package named R-INLA (Rue et al., 2009). INLA is a computationally-efficient approximation to a classic Markov-Chain Monte-Carlo approach. It differs from statistical inference for latent Gaussian Markov random field models, as described in Rue et al. (2009). The INLA approach approximates the Posterior marginals of the latent Gaussian field. Latent Gaussian models are a broad, flexible class of models that include (generalized) linear, mixed, spatial and spatio-temporal models. Combined with the Stochastic Partial Differential Equation approach (SPDE) (Lindgren et al., 2011), it can accomodate all kinds of geographically referenced data.

Bayesian geostatistical models are particularly suitable for the production of high resolution maps of development and health indicators (Bosco et al., 2018).

5.4. Mapping sex-disaggregated development indicators at high spatial resolution using survey data, tower level CDR data and environmental/socio-demographic covariates

Multiple approaches are available to create high-resolution surfaces using a combination of geolocated household surveys and geospatial covariate data. We tested and compared a variety of Bayesian geostatistical models and machine learning techniques (Artificial Neural networks, Random Forest (RF), Gradient Boosting Trees (GBT)).

Following an initial test phase to assess model performance on a subset of the selected sex-disaggregated indicators, Bayesian geostatistical models and Artificial Neural Networks were retained as the more promising methods.

For each of the selected sex-disaggregated indicators we produced two different high-resolution maps (each with associated uncertainty). The first set of maps utilise covariates derived from remote sensing data (e.g. NDVI, EVI, etc.) and GIS data (e.g. roads, rivers, urban/rural, etc.). The second set incorporate the CDR-derived covariates, aggregated at tower level, alongside the RS/GIS covariates. This permits us to assess the presence of additional informative power contributed by CDR data relative to RS/GIS data alone.

5.4.1 Selection of the indicators to be modelled

In order to maximise scope for the analysis to support development policy priorities and resource allocation indicator selection was informed by senior personnel at the Central Bureau of Statistics (CBS) in Nepal. As detailed above (section 3), short-listed indicators included: educational attainment (secondary schooling), literacy, market labour participation, farm-based livelihoods, attitudes to GBV against women, births in health facilities and childhood stunting.

Selection of the final set of indicators was guided by preliminary statistical analysis to detect the presence of (linear and nonlinear) correlation between the phenomenon under investigation and the available set of covariates.

Here, the availability of a sufficiently correlated set of explanatory covariates was the deciding factor. The model architecture, while sophisticated, cannot compensate for an absence of information. Two indicators, stunting in children and attitudes to GBV against women were omitted from further analysis due to lack of correlation between the geolocated survey data and the geospatial covariates. However, when correlation exists our models were always capable to exploit this signal for obtaining good modelling performance.

We tested for linear and nonlinear associations in the data. In addition to the Pearson correlation coefficient (for linear associations), we exploited a new measure of dependence (termed distance correlation) developed by Szekely, Rizzo, and Bakirov in 2007. Distance correlation provides a measure of association between two variables or vectors, but accommodates linear, nonlinear and nonmonotone dependence structures (Szekely et al., 2007).

Following this analysis, literacy, agriculture-based occupation, market-labour participation, and health facility births were retained for subsequent modelling.

5.4.2. Selection of geospatial covariate layers

To maximise the predictive accuracy of a model it is fundamental to select the optimal set of covariates. The inclusion of too many covariates can result in overfitting while the inclusion of too few informative covariates can cause a loss of explanatory power. A number of common and widely accepted (Murtaugh 2009) techniques are available to support statistically robust selection of the best performing covariates for inclusion within a modelling architecture.

We conducted a sensitivity analysis, using a jackknife approach, to select an appropriate set of GIS and RS covariates for each of the retained indicators. The jackknife technique (Tukey, 1958) consists of dropping one observation at a time from a set of data and recalculating the estimate anew each time. The full set of covariates is assessed at each stage of the selection process, with an iterative removal of the covariate contributing least to the model performance. Jackknife techniques can prevent model overfitting, while maximising possible explanatory power. The technique also minimises the risk of missing nonlinear patterns of correlation among multiple covariates.

Multicollinearity is another important consideration, with scope to impact on the stability and quality of the modelling results. We computed Variance Inflation Factors (VIF) for each explanatory variable (the VIF value increases in tandem with multicollinearity). Only those variables with a VIF equal to or less than three were retained (See Bosco et al. (2017) for a fuller discussion of VIF bounds for inclusion).

Starting with the highest-correlated pairwise variables, and applying the jackknife approach, we omitted the covariate that did not appear in the set associated with the lowest mean squared error (MSE). Once multicollinearity was under control, we continued applying the jackknife analysis in order to select the final set of explanatory variables, maximising the model prediction capacity. All of the explanatory variables were also normalized to have a mean of zero and unit

variance. This was done to limit the effects of outliers and different units of measure among the covariates.

Given the large number of covariates assembled from the RS/GIS and CDR data and the relatively small number of clusters (N = 383) within the DHS database, we applied unsupervised methods based on distance correlation to control for overfitting and strongly reduce the risk of chance correlation. This method was applied to select a subset of covariates with lower correlation with each other. While this approach represents a suboptimal selection process for the more informative covariates, it was adopted to maximise the probability of retaining a set of explanatory variables with good correlation, without feeding the model information on the dependent variable.

The unsupervised method was designed to select a group of up to 20 covariates, mostly uncorrelated with each other. To do that, we applied the dist_corr function of the Mastrave modelling library (de Rigo, 2012c) within the GNU Octave computing environment. We picked the pair of covariates with the highest distance correlation between them, eliminating one of those covariates at random. We proceeded iteratively until we reached the desired number of covariates. The application of this procedure ensures the retention of a set of covariates exhibiting maximum dissimilarity.

This unsupervised method for covariate selection has similarities with principal component analysis (PCA). PCA reprojects a set of possibly correlated covariates into a set of 'principal components' that are linearly uncorrelated and represent a linear combination of the covariates. With the method proposed here, the covariates are uncorrelated with each other, similarly to the vectors of PCA, but retain their original shape.

5.4.3. further details on the modelling architecture for high resolution mapping

As previously detailed, we applied machine learning techniques (ANN) and Bayesian geostatistical models to produce sex-disaggregated maps of literacy, agriculture-based occupation, and health facility births at high spatial resolution, across Nepal.

Both the ANN models and those implemented in INLA used cross-validation (repeated random sub-sampling), applied to 70% of the available DHS data, to tune the model parameters. For the neural network, parameters include: the number of neurons in each layer, the performance function, and the activation function of the hidden and output layers (Sigmoid, linear, etc.). As a training algorithm we used the Levenberg-Marquardt method. For the Bayesian method implemented in INLA, parameters include: likelihood models, prior distribution of the hyperparameters and building of the mesh.

In order to further increase model performance, a simplified version of the Selective Improvement by Evolutionary Variance Extinction (SIEVE) (de Rigo et al., 2005) was applied to the ANN. This simplified SIEVE was utilised for both the high resolution mapping and CDR-based gender prediction models.

The core of the SIEVE architecture is to iteratively select the best parameter vectors, so reducing exponentially the number of parameter vectors surviving each iteration. This reduction of parameter vectors is typically compensated for through the extension of the computational resources dedicated to training each parameter vector until the optimum vector passes the final SIEVE. The complete (non simplified) SIEVE architecture includes a "generative" phase (bypassed in the simplified SIEVE) within each step, where a cloud of new vectors are generated close to each surviving parameter vector from the previous sieves. We initially applied the simplified SIEVE to reduce the required computational time. The simplified version proved sufficiently robust to improve model performance.

After an initial testing phase, during which we compared many different model architectures, we decided to apply INLA as the default. INLA models are considerably less time demanding than ANNs, yet have similar predictive capacity (Bosco et al., 2017). For this reason, the majority of the maps we produced in this study are based on the INLA family of models. Artificial neural networks were, however, applied in instances where INLA resulted in poor results in prediction and in particular, when unusual data distributions required accommodation.

5.4.4. Model validation

In order to achieve the best possible accuracy for each of the modelled indicators, while retaining model's flexibility in predicting over untested areas, the predictive capacities of different models were compared, exploiting a two-step validation approach.

By exploiting a repeated random sub-sampling cross-validation on the set of data selected for model training, we selected the best model for each of the tested model architectures. We quantified the accuracy of the model (the relationship between predicted and observed values) by calculating the root mean square error (RMSE) and the mean absolute error (MAE). Although some authors suggest that inter-comparisons of average model-performance should be based exclusively on MAE (Willmott and Matsuura, 2005), we chose to additionally calculate the RMSE given its sensitivity to occasionally large predictive error. We used the remaining 30% of the data to validate the model. MAE, RMSE and the explained variance of the model (expressed in proportional terms) were calculated to measure model performance.

To produce the final map (at a resolution of 1x1 km^2) for each of the modelled indicators, the model with the highest explained variance and lowest RMSE and MAE was selected. Explained variance was calculated using the approach detailed in by Bosco et al. (2017).

In addition to the calculation of the RMSE and MAE for each of the models, we introduced another parameter to calculate the general bias of a model:

$$\text{general bias} = \frac{|\overline{\text{obs}} - \overline{\text{pred}}|}{\sigma_{\text{obs}}},$$

where pred is the mean of the predicted values, σobs and obs are the standard deviation and mean of observed data.

The measure of general bias was introduced to discriminate between systematic over or under-estimates of a model and models having an overall bias mitigated by the compensating local under/over-estimations. This is because MAE and RMSE cannot directly preserve information on the sign of the modelling errors.

5.4.5. Results

The results from this study show that spatially detailed sex-disaggregated estimates of a variety of development indicators can be produced from survey cluster data and mapped at high spatial resolution. In tables 5.1, 5.3 and 5.5 we present the performance of the applied model architectures for each of the investigated indicators with and without the inclusion of CDR covariates.

For each of the modelled indicators we show maps of the survey clusters and the indicator value in each of the clusters, maps of the predicted proportion of the modelled indicators, the level of uncertainty associated with these maps and a graph comparing DHS vs predicted values at province level. We also present the results of the covariate selection exercise, detailing which covariates were selected as the optimum performing set for the given indicator.

We organized the presentation of results by indicator, and disaggregated by sex, in the following order: proportion of females and males that are literate, proportion of females and males engaged in agriculture, and proportion of births in health facilities.

Literacy

Table 5.1 - Comparison of Bayesian model results for male and female literacy with and without the inclusion of CDR data in the modelling architecture. RMSE, MAE, explained variance, MSE and MSE of a trivial model were calculated.

Modelled Indicator	Modelling technique	MSE	RMSE	MAE	Exp.Var.	MSE (trivial)
Female literacy	INLA (GIS/RS)	0.016	0.13	0.1	0.65	0.047
Female literacy	INLA (with CDR data)	0.018	0.134	0.1	0.62	0.047
Male literacy	INLA (GIS/RS)	0.02	0.14	0.1	0.32	0.029
Male literacy	INLA (with CDR data)	0.02	0.14	0.1	0.32	0.029

Table 5.2 Summary output of the covariate selection procedure for male and female literacy with and without exploiting CDR data in the modelling architecture. Exp. Var. is the proportion of variance explained by each of the models.

Female literacy	GIS/RS data	GIS/RS and CDR data
N. of covariates	10	9

Exp. Var.	0.65	0.62
Selected Covariates	Enhanced Vegetation Index Net Primary Productivity Distance to rivers Distance to roads Potential evapotranspiration Distance to School Nightlights (threshold of 0.5) Ethnicity caste hill Ethnicity madhesi Distance from areas with nightlight values over 0.5	Incoming call duration (median) Total call counts (median) Mid Infrared Index (max) Distance to rivers Distance to roads Distance to schools Protected areas Ethnicity madhesi Ethnicity muslims

Male literacy	GIS/RS data	GIS/RS and CDR data
N. of covariates	7	8
Exp. Var.	0.32	0.32
Selected Covariates	Distance to urban areas Landcover Distance to roads Precipitations Ethnicity madhesi Ethnicity muslims Distance from areas with nightlight values over 0.5	Handset weight (median) Distance to rivers Distance to roads Precipitations Ethnicity madhesi Ethnicity muslims Potential evapotranspiration Nightlights

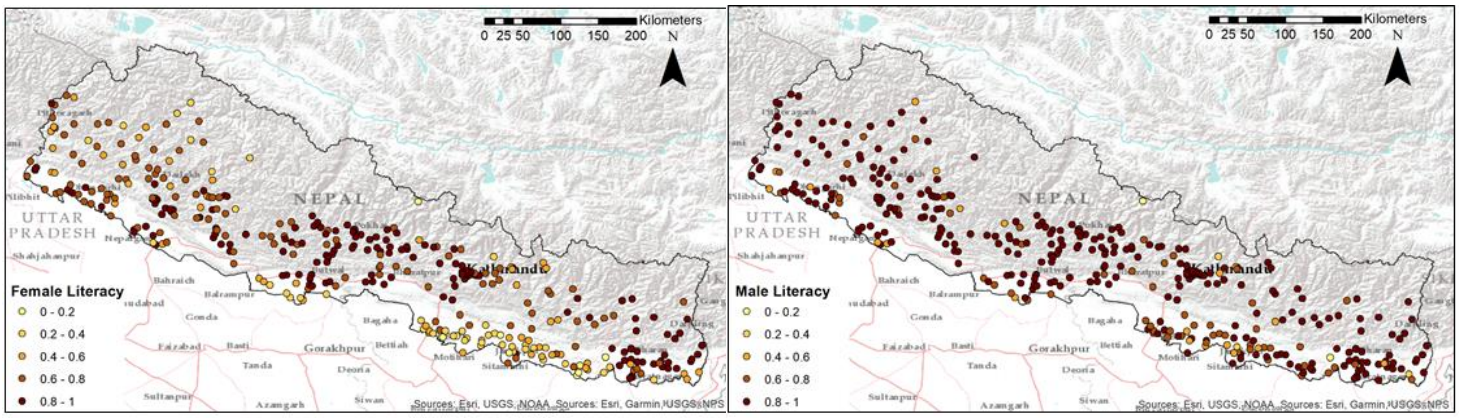

Figure 5.1. Map of the cluster-point DHS survey data for female and male literacy in Nepal.

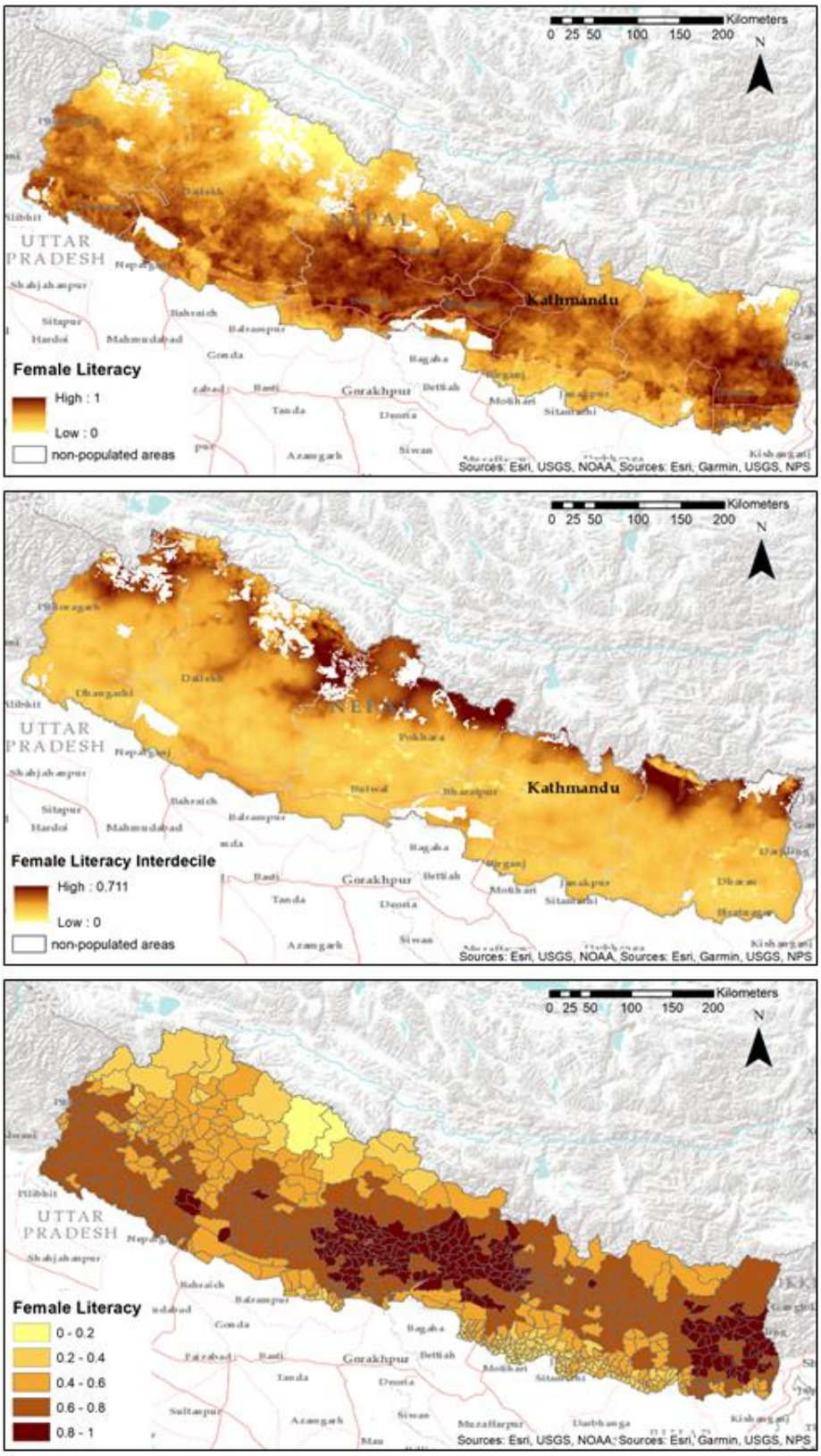

Figure 5.2. Maps of the proportion of female literacy from GIS, RS and CDR data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of female literacy weighted by population aggregated at municipality and rural municipality level (bottom).

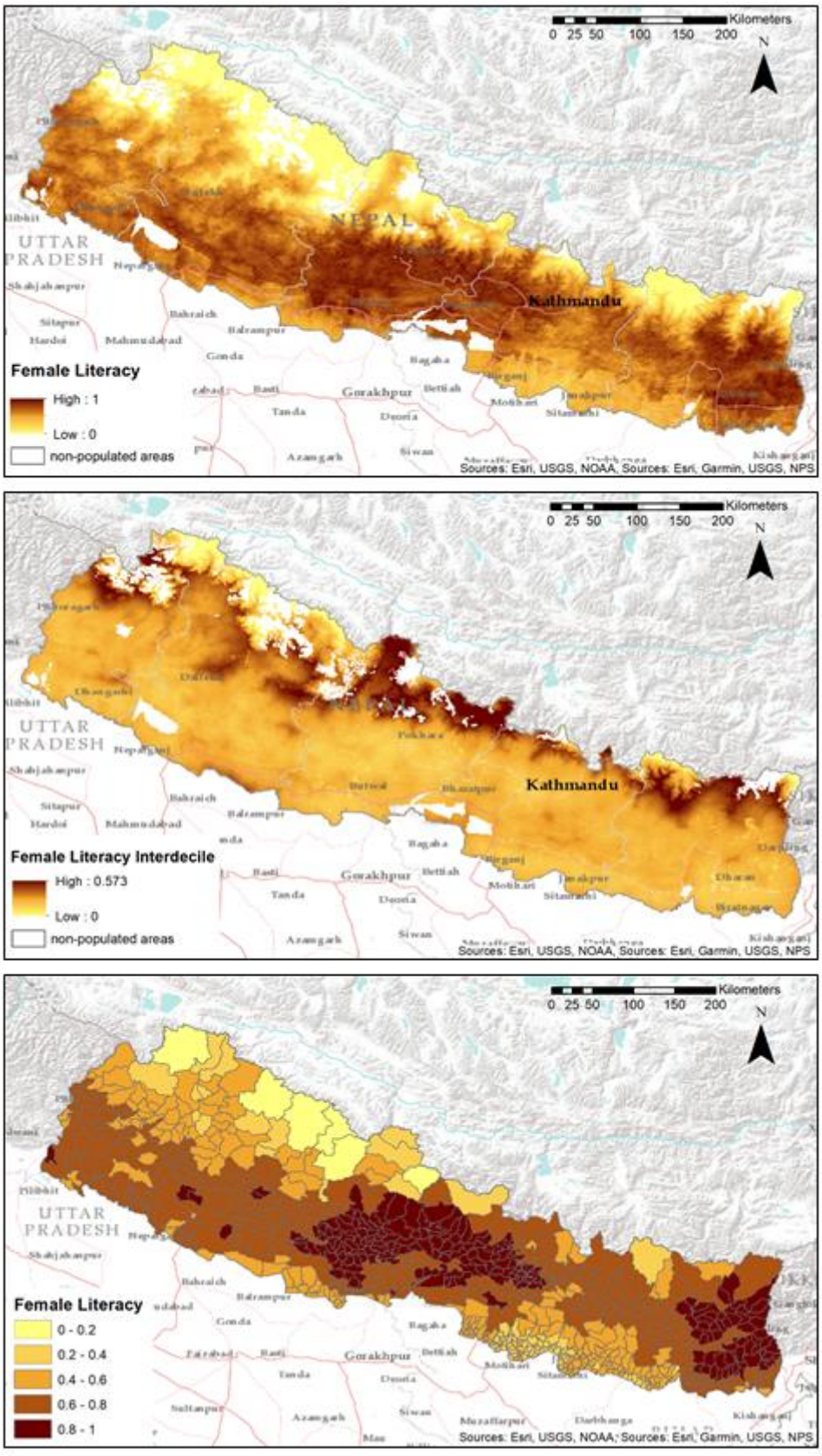

Figure 5.3. Maps of the proportion of female literacy created by exploiting only GIS and RS data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of female literacy weighted by population aggregated at municipality and rural municipality level (bottom).

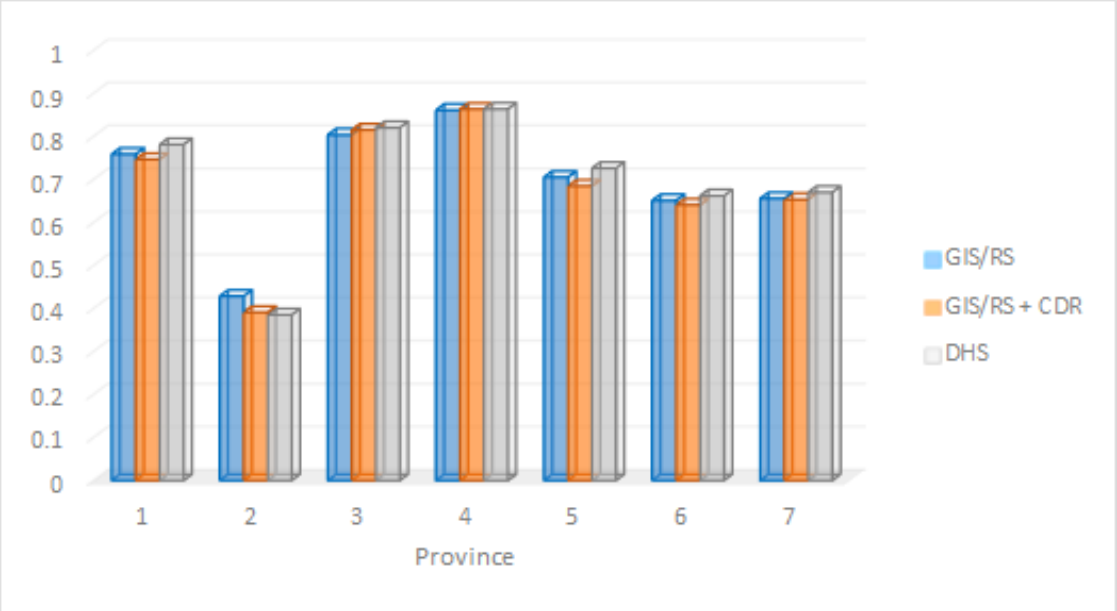

Figure 5.4. Predicted values of female literacy (aggregated at province level), derived from GIS/RS covariates (blue), GIS/RS covariates and CDR information (orange), and DHS survey data (grey).

Agriculture based occupation

Table 5.3 - Comparison of modelling results for male and female engagement in agriculture with and without the inclusion of CDR data in the modelling architecture. RMSE, MSE, MAE, explained variance and MSE of a trivial model.

Modelled Indicator	Modelling technique	MSE	RMSE	MAE	Exp.Var.	MSE (trivial)
Female Agri	INLA (GIS/RS)	0.052	0.22	0.16	0.55	0.11
Female Agri	INLA (with CDR data)	0.054	0.23	0.17	0.53	0.11
Male Agri	INLA (GIS/RS)	0.055	0.23	0.19	0.36	0.087
Male Agri	ANN (GIS/RS)	0.056	0.23	0.19	0.36	0.087
Male Agri	INLA (with CDR data)	0.054	0.23	0.19	0.37	0.087

Table 5.4 Summary output of the covariate selection procedure for male and female engaged in agriculture with or without exploiting CDR data in the modelling architecture. Exp. Var. is the proportion of variance explained by each of the models.

Female empAgri	GIS/RS data	GIS/RS and CDR data
N. of covariates	4	6
Exp. Var.	0.55	0.53

Selected Covariates	Global Urban Footprint Distance to schools Nightlights (threshold of 0.5) Distance from areas with nightlight values over 0.5	Handset weight (median) Outgoing call duration (median) Distance to schools Nightlights (threshold of 1) Distance from areas with nightlight values over 1 Distance from areas with GHS values over 0.2

Male empAgri	GIS/RS data	GIS/RS and CDR data
N. of covariates	7	10
Exp. Var.	0.36	0.37
Selected Covariates	Friction_surface Food security and related distance Mid Infrared Index Precipitation Distance to schools Nightlights (threshold of 0.5)	Handset weight (median) Incoming call duration (median) Precipitation Distance to roads Distance to schools Nightlights (threshold of 1) Distance from areas with nightlight values over 1 Distance from areas with GHS values over 0.2 and 0 Ethnicity madhesi

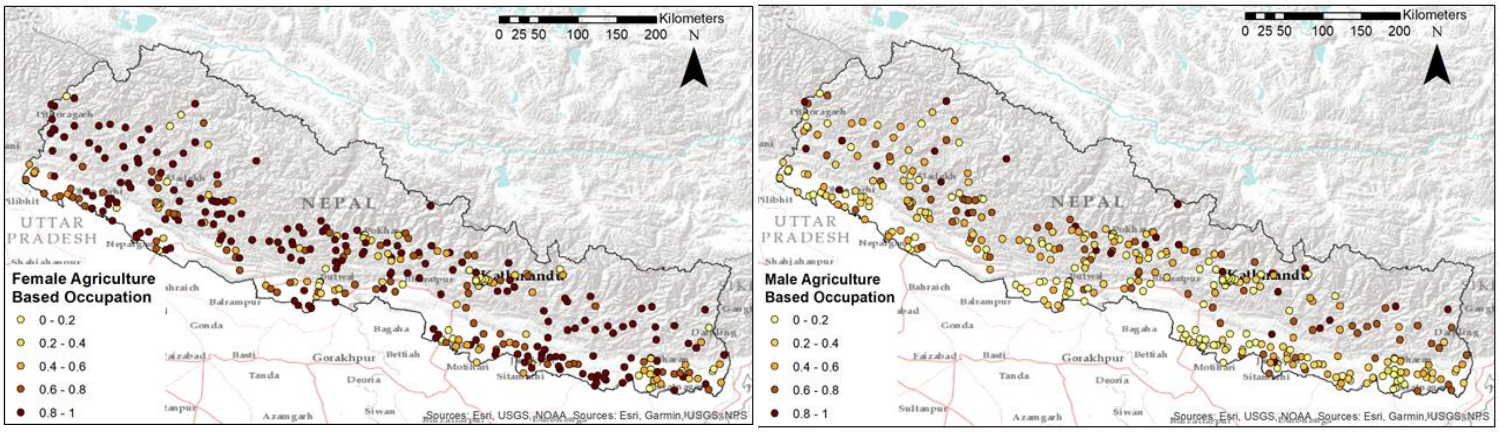

Figure 5.5. Map of the cluster-point DHS survey data for female and male engagement in agriculture in Nepal.

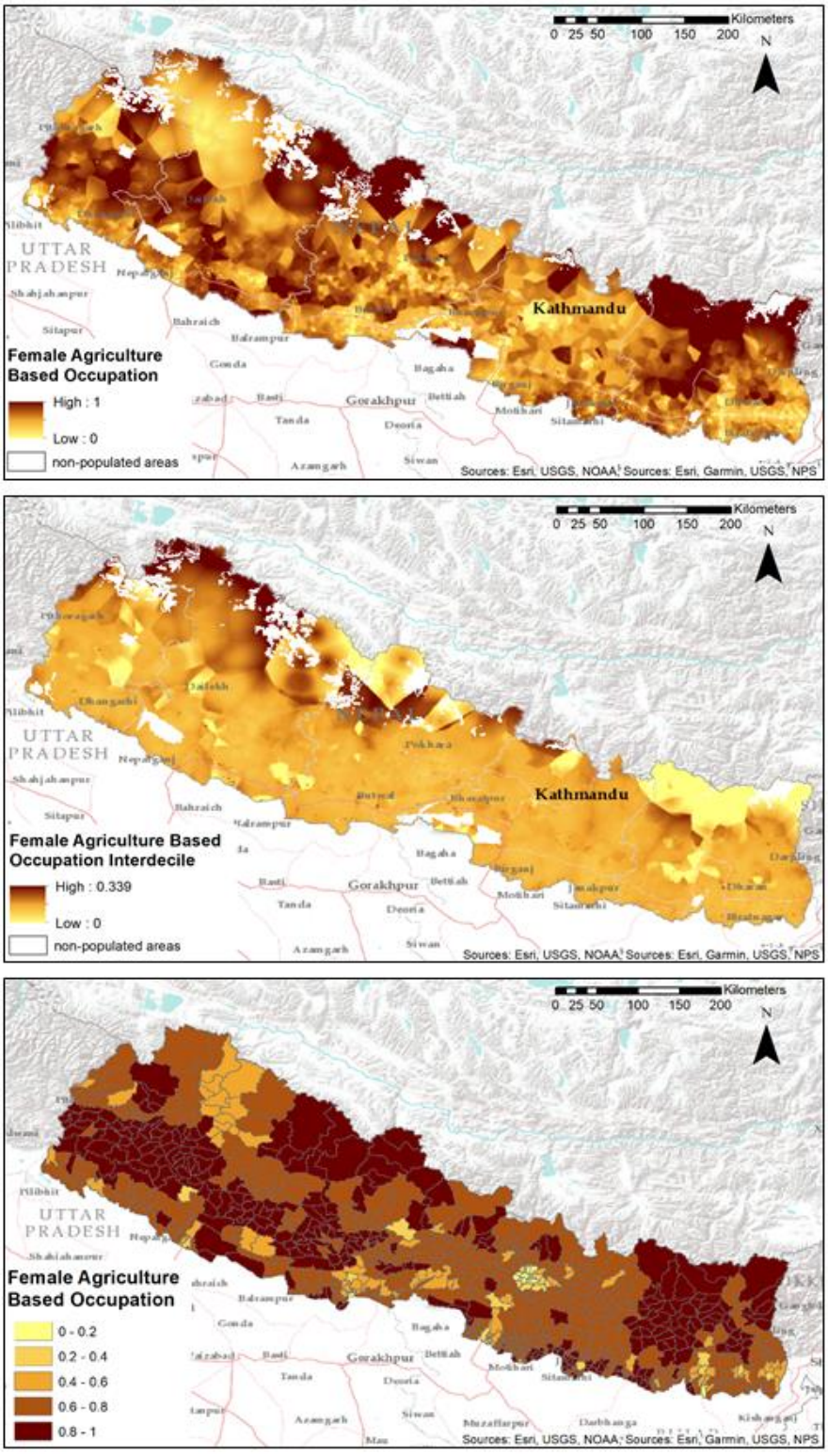

Figure 5.6. Maps of the proportion of female engagement in agriculture from GIS, RS and CDR data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of female employment in agriculture weighted by population aggregated at municipality and rural municipality level (bottom).

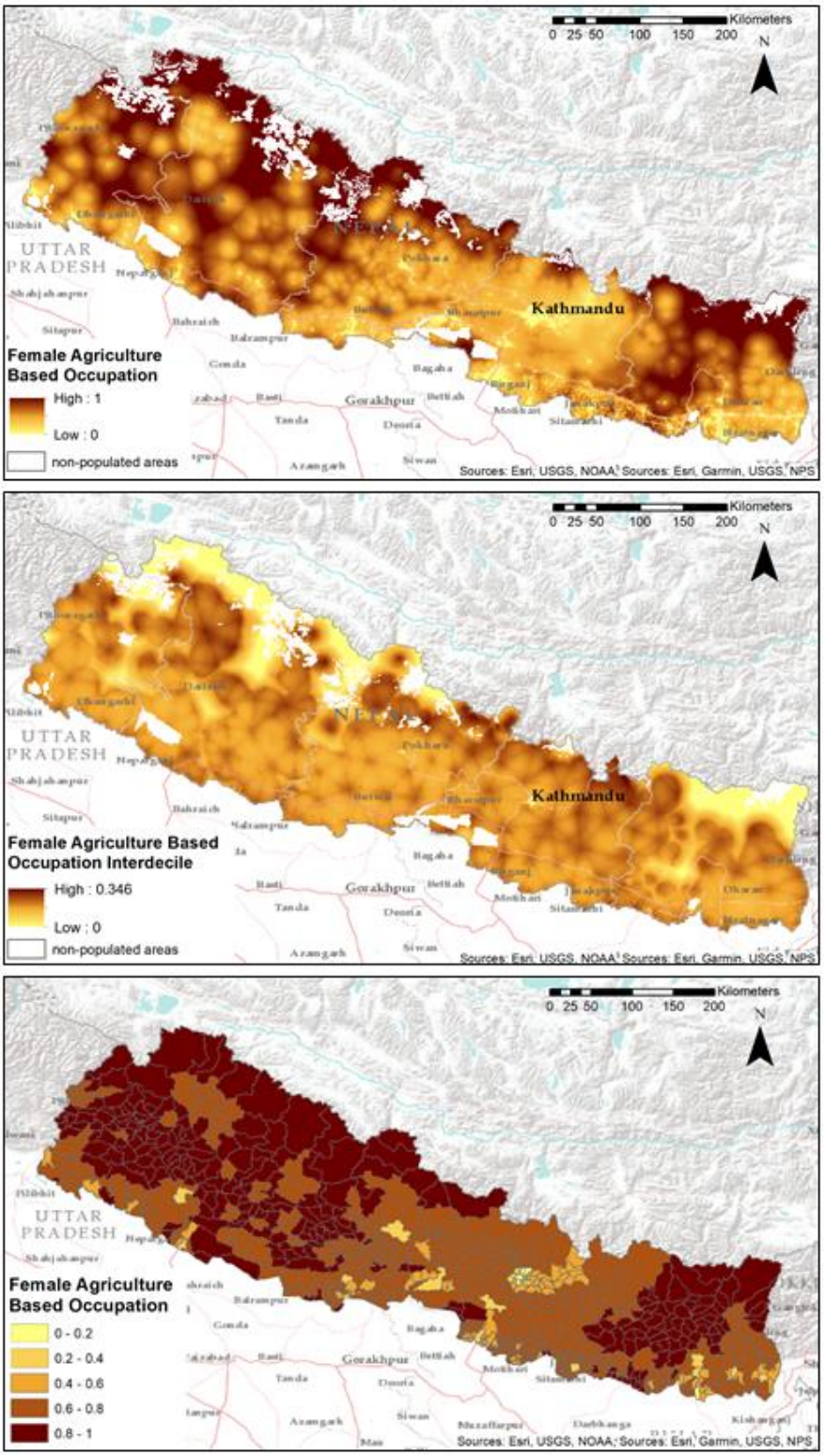

Figure 5.7. Maps of the proportion of female engagement in agriculture created by exploiting only GIS and RS data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of female engagement in agriculture weighted by population aggregated at municipality and rural municipality level (bottom).

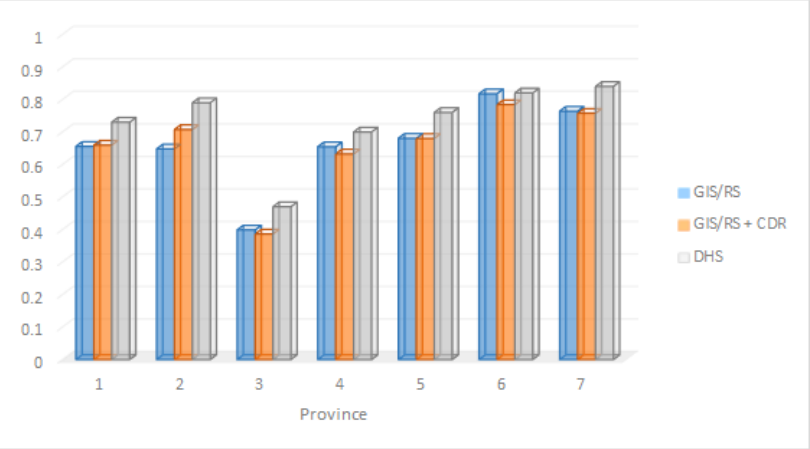

Figure 5.8. Predicted values of female engagement in agriculture (aggregated at province level), derived from GIS/RS covariates (blue), GIS/RS covariates and CDR information (orange), and DHS survey data (grey).

Births at Health Facilities

Table 5.5 - Comparison of Bayesian model and ANN results for the proportion of births at health facilities with and without the inclusion of CDR data in the modelling architecture. RMSE, MAE, explained variance and MSE of a trivial model were calculated.

Modelled Indicator	Modelling technique	MSE	RMSE	MAE	Exp.Var.	MSE (trivial)
Health fac. births	INLA (GIS/RS)	0.044	0.2	0.17	0.53	0.094
Health fac. births	ANN (GIS/RS)	0.043	0.2	-	0.54	0.094
Health fac. births	INLA (with CDR data)	0.048	0.22	0.17	0.49	0.094

Table 5.6 Summary output of the covariate selection procedure for modelling births at health facility with and without exploiting CDR data in the modelling architecture. Exp. Var. is the proportion of variance explained by each of the models.

Health fac. Births	GIS/RS data	GIS/RS and CDR data
N. of covariates	6	8
Exp. Var.	0.53	0.49
Selected Covariates	Global Urban Footprint Distance to schools Distance to roads Protected areas Crop Suitability Gross Primary Productivity	Incoming_call_duration (median) Distance to roads Protected areas Distance to schools Distance to rivers Ethnicity madhesi Ethnicity muslims Distance from areas with GHS values over 0.2

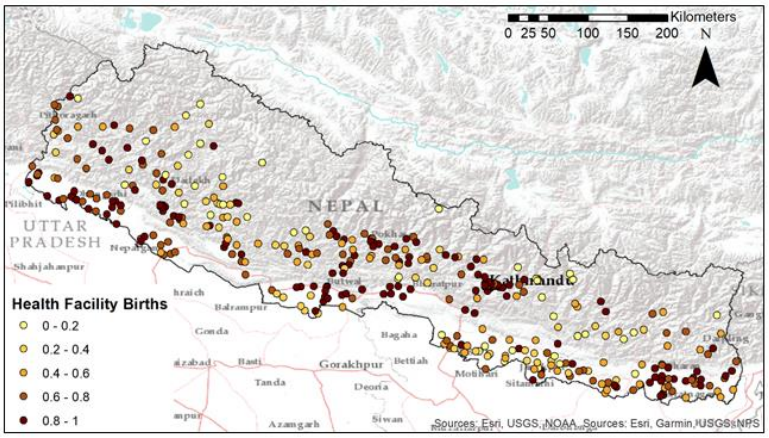

Figure 5.9. Map of the cluster-point DHS survey data for health facility births in Nepal.

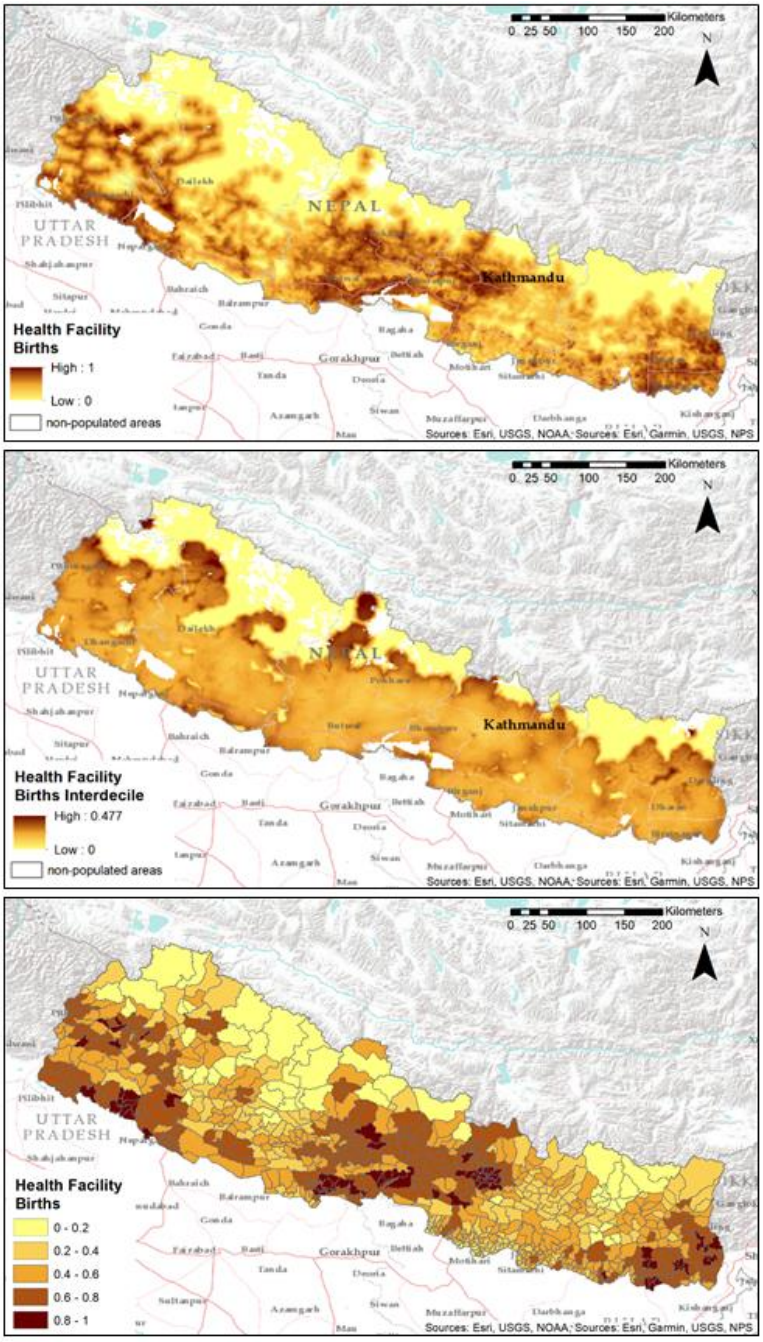

Figure 5.10. Maps of the proportion of health facility births from GIS, RS and CDR data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of health facility births weighted by population aggregated at municipality and rural municipality level (bottom).

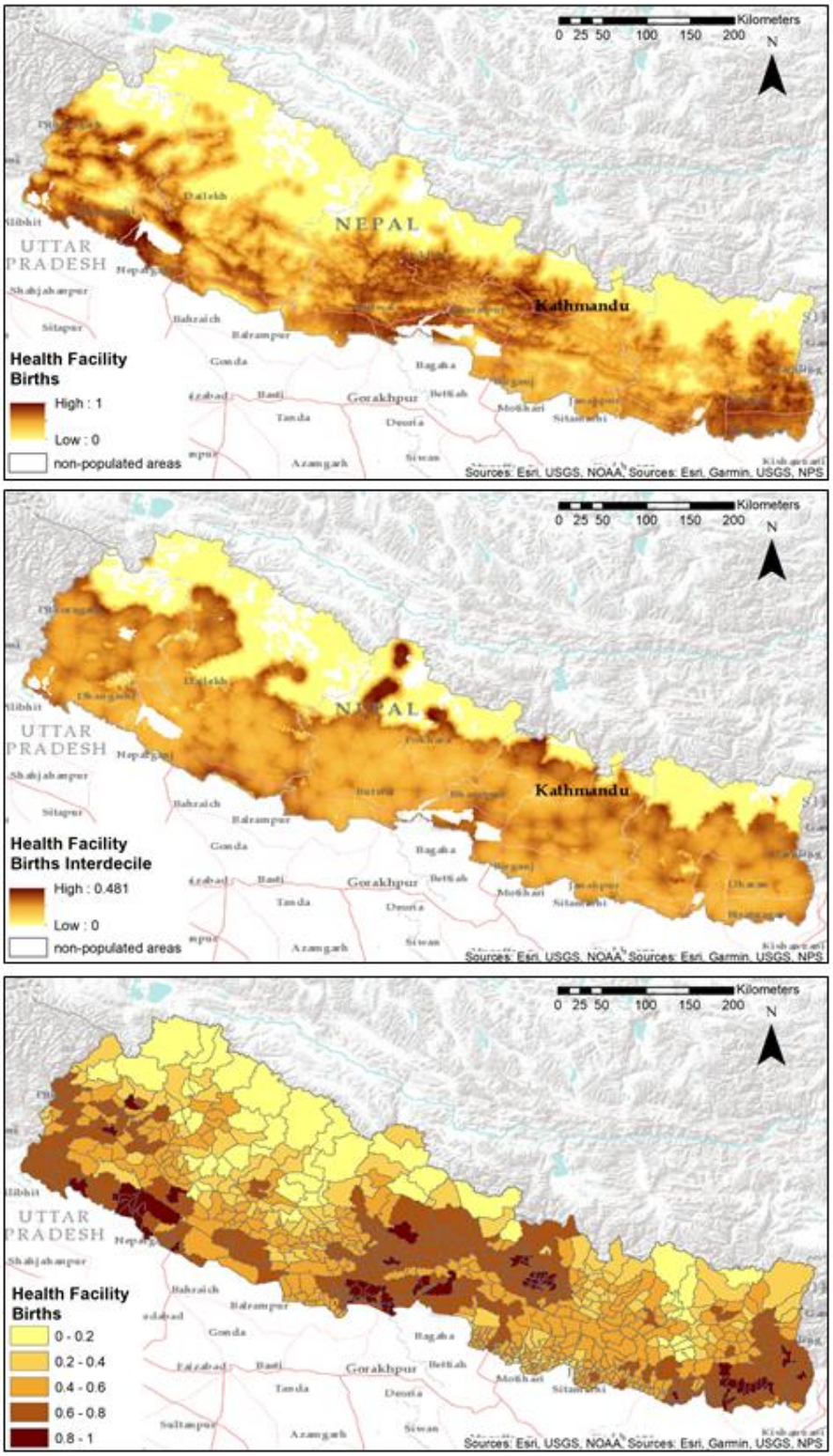

Figure 5.11. Maps of the proportion of health facility births created by exploiting only GIS and RS data. The maps show the median (top) and interdecile (middle row) values at 1km2 resolution and the values of health facility births weighted by population aggregated at local-level (bottom). The maps were created applying the INLA model, this technique was here preferred to ANNs because with a similar prediction ability but a lower modelling uncertainty.

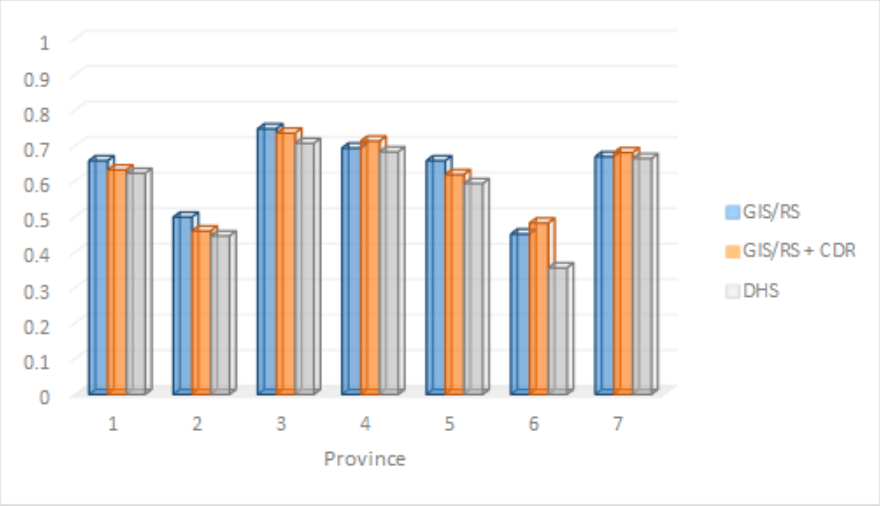

Figure 5.12. Predicted values of health facility births, aggregated at province level, derived from GIS/RS covariates (blue), GIS/RS covariates and CDR information (orange), and DHS survey data (grey).

5.4.6. Discussion

The focus of the SDGs on "reaching the furthest behind first" creates a need for new approaches capable of identifying who and where these people are in order to reach them. Moreover, to reach the SDGs and to be able to track progress towards meeting these goals, it is necessary to regularly update such information.

Census data can provide information that is sufficiently spatially detailed to generate a detailed geographical snapshot for some development indicators. The possibility to observe changes and dynamics over time is limited, however, by the decennial (at best) timing of the population census.

Here we explored the potential for spatial interpolation methods to generate accurate maps depicting variations and inequalities between and among men and women, proxied by spatial and sex-disaggregated variations in key development indicators measured by a quinquennial household survey. We also explored the possibility to provide frequent updates to the modelled layers by exploiting (the daily availability of new) CDR data.
The obtained results highlight the practical challenges involved in using spatial interpolation methods to map survey-derived development indicators at high-resolution. The relatively small number of geolocated data points (PSUs / clusters) typically available in nationally representative household data, inhibits the full exploitation of the available model architectures.

We modelled three indicators, disaggregated by sex and mapped at high resolution across Nepal. We obtained explained variance (in validation) of around 55% to 65% for all selected indicators predicted for women. Equivalent results could not, however, be obtained for men. BGS and ANNs achieved similar performance.

Some of the covariates evidenced strong correlation with the selected indicators. For example, the Pearson correlation of the global urban footprint with female agriculture is around 0.65 and

around 0.55 for the correlation with the distance from areas with nightlight values over 0.5. The same covariates present a lower correlation with male agriculture (0.42 and 0.44 respectively). The same pattern is observed for the sex-disaggregated literacy data.

A number of factors underlie the differences in model performance for the data on male and female development indicators. The amount - and spatial scale - of variation displayed for each indicator is a major factor, as is the extent to which the indicator is associated with the available geospatial covariates. If limited information is present in the covariates, the model fails to predict the phenomena well, but good performance is always obtained where good correlation exists.

In general, good correlation between literacy and urbanisation has been shown previously (Arouri et al., 2014). Correlations have also been shown between urban areas and delivery in health facility births (Tey and Lai 2013, Zere et al., 2011), employment in non-farm activities and proximity to urban centres (Deichmann et al., 2009). Remotely-sensed satellite data for extracting environmental metrics such as land cover types have been linked to female literacy (Watmough et al., 2013).

The tables listing the covariates running within the applied models (paragraph 5.4.5), show that introducing CDR data to the modelling architecture produced no consistent improvement in the model performance for either the male or female data.

Notwithstanding this finding, there is evidence of correlation between the set of covariates derived from CDR data and the modelled indicators. For example, the proportion of females in agriculture shows a direct correlation with the handset weight (0.62), the proportion of males in agriculture and of births in health facilities have a correlation of 0.46 and 0.42 respectively with incoming call duration.

Despite this new source of information, the model's predictive performance did not improve as hypothesised. The difficulty encountered in exploiting this additional information may have arisen from the (unavoidable) application of an unsupervised method for selecting the covariates when CDR data are used. The high number of covariates and the small set of input data (the geo-located survey cluster data), forced the selection of a suboptimal set of data on the basis of distance correlation values - in order to avoid model overfitting and reduce the possibility of chance correlation (section 5.4.2). The risk in applying this methodology is that covariates containing a large part of the signal are removed, a priori, from model contention.

Even in the event of good model performance, known sources of error increase uncertainty. The DHS survey's introduction of a random displacement on cluster location in order to protect respondents' anonymity is one such source of uncertainty for the model (Bosco et al., 2017). To mitigate this potential source of error we extracted mean values through a defined buffer around the survey clusters (Perez-Haydrich et al., 2013). While the extent of the impact of this displacement can vary between indicators and different surveys, in general, its effects should be modest (Gething et al., 2015).

The maps produced for female literacy, agriculture-based occupation, and health facility births have sufficient accuracy to enable re-aggregation to a geographical level (e.g. municipality) relevant for decision makers and for planning and resource-allocation purposes. Given that many of the modelled covariates are frequently updated, there is potential for the regular (or even continuous) revision and monitoring of these indicators to inform development policy and resource allocation throughout the country.

The results also highlight the need for caution - and for further investigation. Future analysis, should consider a larger set of geo-located clusters with indicator data, ideally for multiple time intervals.

In order to assess the ability of the available techniques to produce accurate maps of development indicators at administrative level, we re-aggregated the maps we created at 1km^2 resolution to Province level. We then compared the results with the equivalent province-level estimates obtained from the NDHS 2016 (MOH et al, 2017).

The graphs in Figure 5.4, 5.8 and 5.12 show the high accuracy of our models (with and without exploiting CDR data) in predicting female literacy, agriculture-based occupations and the percentage of births delivered in health facilities. For example, Figure 5.4 reports a maximum distance between the proportion of literate women based on DHS survey data - and our results - of 0.04. These strong results are similarly observed for health facility births, where the map created (also exploiting CDR data) differs from the NDHS 2016 province level estimates by between just 0.01 and 0.03 (with the exception of province number 6 where the difference between the two maps is over 10%). Similarly, good performance is also observed for the map related to feminisation of agriculture, where the map exploiting CDR data has a maximum distance of 0.08 with DHS values in each of the seven provinces.

Although the addition of CDR data does not seem to increase predictive performance overall (subject to verification, applying fully supervised techniques on a larger set of data), the evidence is that this new source of information is capable of improving sub-national model performance. For example, in province number two, both for health facility births and feminisation of agriculture, the addition of information derived from CDR data substantially increased the models predictive power.

While the study results suggest spatial interpolation methods offer a way to provide regular updates on development indicators, further investigation of the minimum and maximum time windows for the remote sensing and CDR data, and for other time-variant covariates, remains necessary for future modelling. Further effort is required to obtain additional datasets covering a range of time intervals to confirm this hypothesis.

5.5. Sex-disaggregate CDR data by carrying out a large-scale phone based survey

Mobile phone survey data, described in section 1 of Material and methods, were analysed in concert with CDR data, in order to generate a gender prediction model for application to the

entire CDR user base. The final objective of this model was to split the CDR database by gender (based on differences in men's and women's observable SIM use episodes), in order to support the development of tools to map women's and men's mobility and migration patterns, trajectories, and dynamics.

Multiple approaches exist to assign demographics - including gender - to de-identified CDR data. We tested Bayesian geostatistical (BGS) models and machine learning (ML) techniques on an initial subset of data collected during the original (2017 - 2018) phone survey. While the applied ML techniques (ANNs, RF and GBT) exhibited similar performance, the Bayesian geostatistical methods performed poorly in terms of predictive power. The difference in model performance most likely originated in the difficulty of exploiting spatial autocorrelation with this type of data.

Based on our past experience of implementing artificial neural networks within varied model architectures, we took the decision to apply this technique to develop the gender prediction models.

5.5.1. Selection of Covariate Layers

As was the case for the high resolution mapping workstream (section 5.4.2.), the predictive capacity of the gender-prediction models largely depends on robust selection of a set of covariates for model inclusion.

The available, truncated, survey data (n = 1,280) inhibited scope to fully exploit the information available in the CDR data. Of the more than 50 explanatory variables derived from the CDR data, a subset of 20 variables was ultimately retained. Covariate selection proceeded in two phases. The objective was to reduce the number of candidate covariates, so limiting the twin risks of model overfitting and chance correlation, while maximising the information contained in the surviving covariate set. Preliminary screening was performed using the unsupervised methodology (based on distance correlation) detailed above (section 5.4.2). This resulted in a subset of covariates, selected to minimise pairwise correlation while retaining their original shape. In the second selection phase a jackknife approach (described in section 5.4) was applied to identify the most informative subset of surviving covariates.

The final set of 20 covariates were normalised (standardised to have a mean of zero and unit standard deviation as dimensionless quantities) to eliminate differences in measurement unit, and mitigate differences in value range.

5.5.2. Modelling Architecture

In this study arm, we applied the same feed-forward neural network architecture used to create the high resolution maps in the preceding section. The specific architecture is a feed-forward multilayer perceptron implemented in Matlab language within the GNU Octave computing environment (see section 5.2). As with the ANNs developed for the high resolution mapping, the database was split into training, validation, and test sets. The training and validation sets (respectively, 60% and 20% of the data) were used to tune modelling parameters, such as the

activation function or the number of neurons in each layer, using the Levenberg-Marquardt training algorithm.

The high rate of SIM sharing (47%) reported in the primary survey presented difficulties when it came to building a robust set of training data. Of the 47% of survey respondents reporting SIM sharing, the majority (35% of all the users) reported use by at least one man and at least one women (the remainder reported shared usage among the same sex). In order to build a model capable of predicting gender from SIM use behaviour, it is first necessary to identify gender with patterns of phone use. The high rate of reported mixed-gender SIM usage problematises the binary classification of individual CDRs.

As discussed in section 1.3, previous published work to predict gender from behaviour traces in mobile phone data (e.g. Jahani et al., 2017, Frias-Martinez, et al., 2010), has acknowledged the theoretical implications of SIM sharing, while, in practice, holding to the 'single-SIM / single user' assumption. In the absence of prior work to accommodate SIM sharing in demographic prediction models, we proceeded systematically. We first tested model performance on an artificial set of data. This artificial set was created by stripping all of the SIM records associated with mixed-gender SIM use from the dataset. The aim of this exercise was to assess the performance of our model against that achieved by previous studies. A classifier based on a feed-forward multilayer perceptron was tuned and validated on this artificial set of data. We obtained interesting results, which we present in the next sections.

Following the initial testing, we assessed the scope for gender-prediction models to be extended to the whole of the CDR dataset. We considered a range of possible solutions to the classification problem introduced by SIM sharing. A large proportion of SIM shared by both females and males implies that a classic binary-classification modelling is becoming useless to acceptably represent the reality described by the surveyed data. This entailed a departure from the initial objective of binary classification to a proportion based scheme (continuous-values regression modelling), whereby we assigned a value to each CDR, representing the relative 'shares' of reported SIM use by gender. Values ranged from 0 to 1, with 1 identifying use by (only) women, and 0 identifying use by (only) men, with intermediate values corresponding to the relative proportion of SIM use by women and men. Section 1 of the report details the response codes related to 'own' and 'others' SIM usage. The classification scheme measures each identified 'SIM users' use frequency on a five item scale (as reported by the survey respondent).

In order to address the classification problem, we proceeded as follows: Based on the relative numerical distribution of the categorical classes (the five-item response scale reported in section 1.3), we inferred the frequency of usage associated with each class, treating the data-transformation as an optimisation problem of discrete density estimation. The information provided by the distribution of the classes enables the relative frequency of SIM usage associated with each item response level to be inferred, such that the estimated frequency minimises the amount of additional prior information, or equivalently, maximises the entropy (as defined in information theory and statistics) (Watson and Elliot, 2016). The entropy score of a given set of survey responses (categorical classes), in relation to their corresponding

quantitative frequency (unknown: to be estimated) is a measure of the unpredictability of the data-driven frequency, or equivalently, its information content.

The resulting set of frequencies constitute the unknown density distribution associated with the classes. In accordance with the principle of maximum entropy, the distribution with the highest entropy is chosen as the least-informative default. Maximum entropy provides an efficient and mathematically robust tool for inferring constraint-dependent probability distributions (de Martino and de Martino, 2018).

We applied the categorical2freq function of the Mastrave modelling library (de Rigo, 2019) within the GNU Octave computing environment. This tool allows all of the survey data corresponding to men's and women's SIM use (e.g. a situation in which one woman uses the SIM 'often' -the 'main user'- and two men (perhaps a husband and son) use the SIM 'rarely') to be exploited to identify the distribution maximising the entropy. Due to the constrained information on sharing practices obtained for the survey sample, bootstrap resampling was applied to estimate the uncertainty originating in data scarcity (as well as standard sources of uncertainty characteristic of survey data) (detailed in Section 1). With a statistical resampling (100 bootstrap runs) of the survey data, we created a set of 100 possible distributions of the proportion of time that women and men spent in using a single SIM.

Seeking to maximise the predictive power of our classifier, and of the overall regression model, we employed within our ANN the simplified SIEVE method detailed in section 5.4.3. (as applied to the high resolution mapping workstream). The model results are detailed in section 5.5.4.

5.5.3. Model Validation

In order to maximise model performance, within a robust model architecture, we split the data into training, validation and test sets (employing a 60-20-20 ratio).
As standard, the performance of a classifier is expressed in terms of the accuracy and AUC (average area under the curve). Here we measured model performance (of our classifier) in terms of classification accuracy. We also calculated values of sensitivity and specificity to detect differences in model performance for gender (proportion) prediction.

The generalisability of the methods we employed to assign (predominant) SIM user gender to anonymised individual-SIM level CDR data depends, largely, on the scope for correlations detected between the linked CDR and survey data to hold beyond the 'training phase'.

To assess model performance for both the tuning - and the final validation of the regression modelling architecture, we calculated the RMSE, MAE and explained variance, as described above (section 5.4.4).

5.5.4. Results

The study results highlight the opportunities, challenges, and limits of models to discern women's and men's SIM use practices from trace data in the CDR. We first present the results

obtained by applying a machine-learning classifier to the subset of data for single-gender SIM use (i.e. sole SIM use and shared SIM use by (only) women or (only men). We then present results for models applied to the full population of SIM users.

Data related to non-mixed-gender SIM-sharing customers

We developed an ANN based classifier, and produced a model with good predictive capacity, for non-mixed-gender SIM-users. The accuracy of this model was over 70% in validation. These results are presented below, and discussed in the next section.

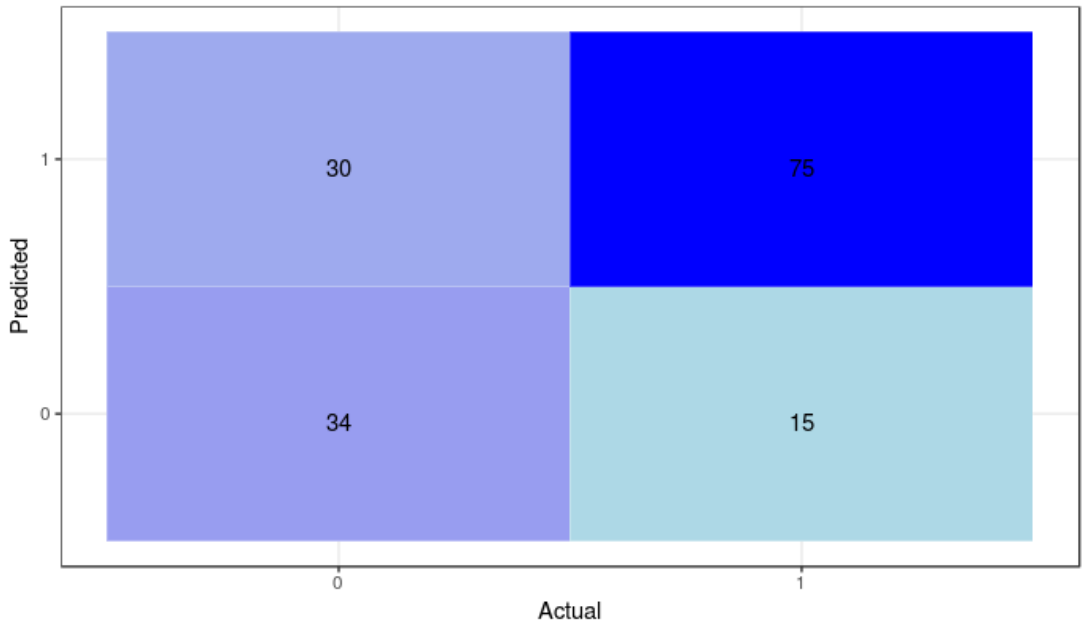

Figure 5.13. Confusion matrix for values predicted in validation, applied to the dataset of non-mixed-gender SIM-users.

Table 5.7. Summary of the classifier applied to predict gender for non-mixed-gender SIM-users. Exp. Var. is the proportion of variance explained by the model.

	Model	Accuracy	Sensitivity	Specificity	Exp.Var.
Training	ANN classifier	0.72	0.72	0.72	0.64
Validation	ANN classifier	0.70	0.77	0.58	0.58

Table 5.8. Summary output of the modelling prediction capacity and covariate selection procedure for modelling gender. Exp. Var. is the proportion of variance explained by the model.

	ANN classifier (GNU Octave)
N. of covariates	8
Exp. Var.	0.58
Selected Covariates	Distance travelled Incoming interevent time Outgoing call duration Incoming call duration

	Subscriber out-degree Unique location counts Displacement Total recharge counts

Data related to all customers

Encouraged by the promising results obtained for the subset of non-mixed-gender SIM users, we developed a model based on the same ANN architecture to split the entire CDR dataset according to gendered usage, based on all of the available survey data on SIM use and sharing. The aim of this model was to predict the proportion of usage by women and men for a single SIM. The results are presented here and discussed in the next section.

Table 5.9. Summary output for model predictive capacity and covariate selection procedure for predicting gendered SIM use using a neural network applied to the whole dataset (containing information on both non-mixed and mixed-gender sim-sharing). Based on 100 bootstrap runs. Exp.Var. is the proportion of variance explained by the model.

	ANN (GNU Octave)
N. of covariates	8
Exp. Var.	0.07 - 0.13
Selected Covariates	Distance travelled Incoming interevent time Outgoing call duration Incoming call duration Subscriber out-degree Unique location counts Displacement Total recharge counts

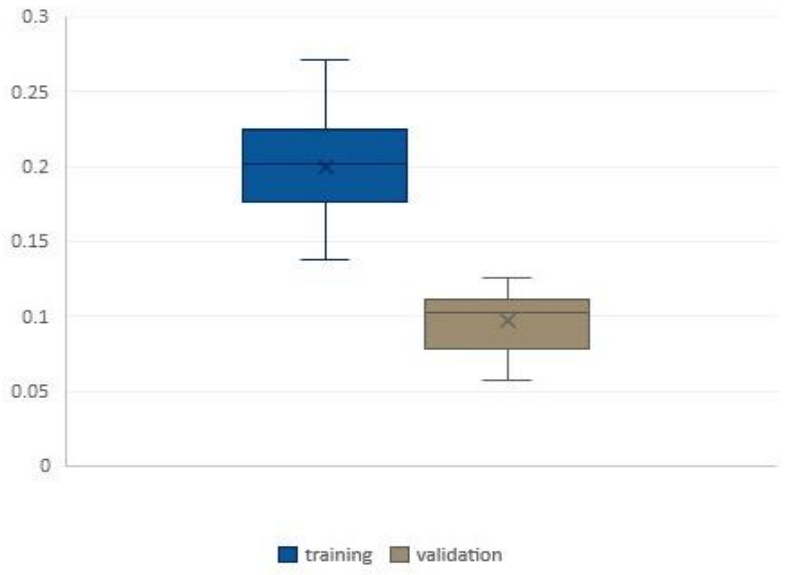

Figure 5.14 Variability of the model predictive capacity (explained variance values from 100 x bootstrap) in training and validation.

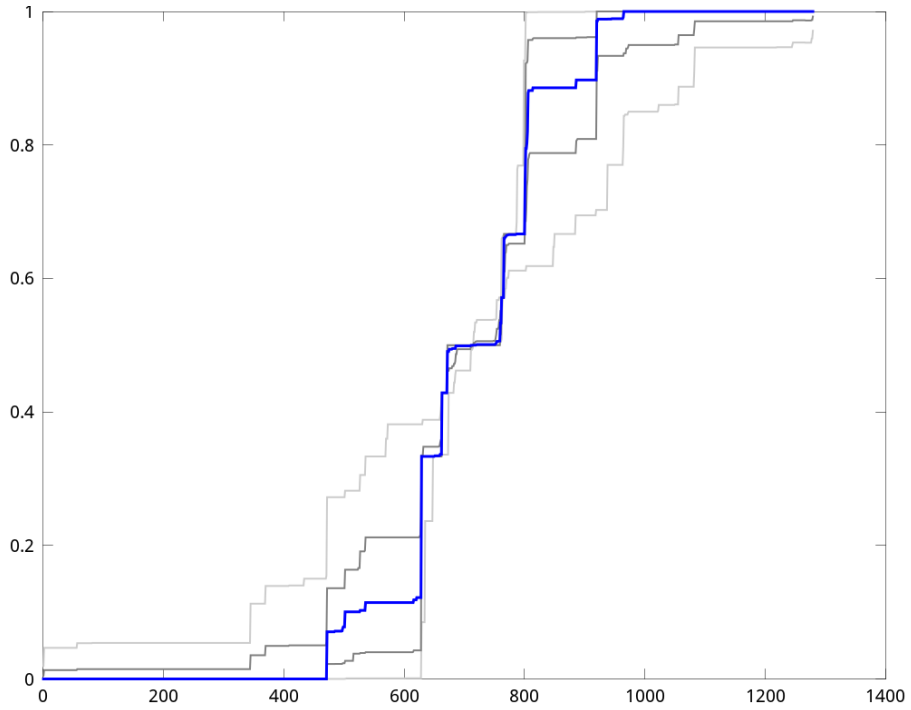

Figure 5.15. Distribution of gendered SIM use(in the range [0 1], where 1 is 100% usage of the SIM by women). Statistical resampling performed with the categorical2freq function in the Mastrave modelling library. The graphs show the statistical uncertainty linked with the applied technique. Blue: median; gray: quantiles 25% and 75%; light grey: quantiles 5% and 95%. The large quantile uncertainty is an indication of data-scarcity (relative small amount of available survey records, n = 1280), and possibly of other sources of data uncertainty linked with the complex survey.

5.5.5. Discussion

The results presented in section 5.5.4 demonstrate the difficulty of predicting gendered SIM usage from behavioural traces in deidentified mobile phone data (number of calls, duration of calls, charging habits, etc.). While slightly more than half (53%) of Ncell subscribers do not permit / invite others to use their SIM, over a third of subscribed SIMs (35%) are subject to mixed-gender use. Compounding the effects of the relatively small sample of survey data, and the consequent higher uncertainty in reconstructing the distribution of gendered SIM use (see Fig. 5.15), the 'noise' introduced by widespread SIM sharing may have further inhibited the models predictive capacity.

To better understand the impact of SIM sharing for predicting gendered SIM use, we tested the predictive capacity of a classifier developed on a subset of the linked survey-mobile phone data, from which we omitted mixed-gender SIM use (results presented in section 5.5.4). While we consider this test data to be 'artificial', the results obtained are comparable with those of previous studies. For instance, Jahani et al. (2017), report that their best performing model achieved an accuracy of 73.6% (for a subscriber base in an unidentified EU country), and 72.9% (for a subscriber base in an unidentified South Asian country), with a training set of 5000 linked-survey observations (when doubling and tripling the set of training data results improved only marginally) (*Ibid*).

The accuracy obtained for our ANN based classifier is slightly over 70% in validation. The difference in accuracy, of a few percentage points, is likely attributable to the (relatively) small set of data with which we had to work. The small training set dictated the adoption of a suboptimal analytical approach, inhibiting, for example, robust supervised training of the model. Instead, we applied an unsupervised technique for covariate selection (distance correlation, as described in sections 5.4.2 and 5.5.1), to mitigate against chance correlation and model overfitting.

Based on the promising results obtained for the subset of data exempting mixed-gender SIM use, we considered strategies to extend the modelling to incorporate mixed-gender SIM use. The original analytical strategy relied on a binary classification scheme to assign a single gendered user to each SIM in the set of training data. The binary classification scheme could not be justified in the presence of evidence on widespread mixed-gender SIM usage. Instead, we assigned each SIM in the set of data a value indicating the proportion [0 1] of use by men and women. The survey items on gendered SIM use contained limited information to support a proportion-based approach (detailed in section 1.3). Reported use frequency was recorded on a standard five-item response scale (never, exceptionally, rarely, sometimes, often) for both the respondent and for each additional person the respondent reported to use the SIM. Gender was recorded for each reported user, as well as for the respondent (gender was classified in terms of kin relationships for related SIM users and as “woman”, “man”, “third gender” for respondents and non-kin relations). Response items were developed to prioritise survey feasibility, validity, and reliability.

The survey items on SIM use-frequency record the relative use of a single SIM by multiple individuals. The items support the classification of gendered SIM use in terms of relative shares of use - for each SIM - by men and women (i.e. greater / equal / lesser shares). The modelling strategy sought to compensate for the absence of an objective measure of use-frequency, definitively linking reported-use-episodes to observable-use-episodes for each (gendered) user. The results demonstrate that the relative scarcity of available survey records, and the absence of precise information conveyed by them (e.g. ideally, recorded as a percentage or unit of time) on individual SIM use, inhibits the models predictive power.

Building on the survey data on gendered sim use, we applied the principle of maximum entropy to classify each SIM according to proportion of use by women and men. The results, reported in table 5.9 and in figure 5.14, indicate poor model performance. The model’s lack of predictive capacity is likely a function of the low cardinality of the final set of data (imposing sub-optimal unsupervised techniques to select the covariate set, and imposing a higher uncertainty in reconstructing the distribution of gendered SIM use, see Fig. 5.15), and the imprecision of the information on mixed gender phone usage. It is important to note that this situation of imprecise information may be expected to be common for similar future surveys. Hence, a methodological modelling answer (as the one here discussed) is likely to remain - even in similar future works - an essential component to exploit this type of imperfect available information. The results of this exploratory research are promising, but subject to the recommendation to complement available systematic databases with carefully designed survey protocols (section 1) to estimate

the otherwise unknown data uncertainty, and partially compensate its severe effects on modelling mobility and migration with the available data.

Besides this key recommendation, a future valuable extension of this work to predict user gender based on behaviour traces in deidentified mobile phone data, would be to explore prediction models with aggregated data. The datasets employed by the present study may contain sufficient information to identify the proportion of men and women, when aggregated.

For example, exploiting Monte Carlo method based techniques, it is likely that the proportion of phone use attributable to men and women can be detected for the municipality and rural municipality (i.e. local government) level. Future work will ideally explore the potential and limits of the proposed approach before collecting and collating new sets of data - containing additional information.

Conclusions and Future Work

Artificial Neural Networks and Bayesian Geostatistical models were applied to create high resolution, sex-disaggregated maps for literacy, engagement in agriculture and for births at health facilities (work package one). The same modelling techniques were also exploited to predict user gender for the CDR dataset (work package two).

In the first of two work packages, results highlight that some of the maps produced in this study are sufficiently accurate to be summarised at an administrative-unit level relevant for policy and decision makers to plan and allocate resources. The maps of female literacy, agriculture based occupation and the map related to health facility births have levels of accuracy that make them suitable for planning purposes. Equivalent results were not obtained for men.

Based on the wider literature, and informed by a workshop organized with Kathmandu-based social scientists to discuss our findings, we hypothesise that cultural aspects are at the basis of these differences in the performance of the models. In order to be able to capture these sociological aspects, new covariates encompassing factors relating to socio-economic differences should be added into our model architecture.

Within the study, we also explored the possibility to provide frequent updates to the modelled layers by exploiting (the daily availability of new) CDR data. The results we obtained show the challenge in fully exploit this new source of information. The relatively small number of geolocated data points, typical of nationally representative household data, inhibits the full exploitation of the model architectures. Although the use of CDR data within the models does not seem to extend their predictive ability (subject to verification, on a larger set of data), the evidence is that this new source of information is capable of improving model performance at local level (section 5.4.6).

With geolocated household surveys being undertaken regularly, the potential exists for continuous monitoring of these and other indicators as well as their dynamics. These maps can support the development of policy to promote women's equal opportunities.

The results also highlight that caution is needed, and further investigation is necessary, and possibly in multiple time intervals. Creating a superior set of covariates with a higher correlation with modelled indicators and especially for males (also including sociological aspects), is recommended to lead to improved estimates.

With surveys and covariates available for multiple time intervals, there is the potential to undertake multi-temporal mapping to measure progress towards meeting sustainable development goals at fine spatial disaggregation.

The study's second work package sought to predict gender for a 'population' of SIM subscribers, based on behaviours observable in their CDR and 'top-up' records. When omitting mixed-gender SIM sharing records, we obtained model performance equivalent to that reported for previous

published studies. Model performance declined with the introduction of mixed-gender SIM sharing records (detailed in section 5.4.4).

The findings highlight the need to interrogate the Single-SIM/Sole-user assumption maintained in the demographics prediction branch of data-intensive modelling. The potential for 'SIM sharing' to destabilise results is widely acknowledged, but little investigated, in this literature (Jahani et al 2017).

To our knowledge, this study represents the first time that SIM sharing has been rigorously assessed and incorporated into model architectures for demographic prediction. The results indicate that, notwithstanding high levels of mobile phone ownership in Nepal, shared SIM use is common. This finding is consistent with prior survey-based research conducted both in the Global North and the Global South, with 'SIM sharing' found to occur in settings with close-to-universal mobile phone ownership, and in the absence of economic compulsions (discussed in section 1.3).

The complexity of the "SIM sharing" construct presents challenges in terms of survey operationalisation and measurement. From a modelling perspective, the survey's measurement of SIM sharing is suboptimal, artificially limiting the information available to assess SIM sharing behaviour. In contrast, an optimal modelling approach , i.e. accessing detailed use frequency for each 'sharer', recorded as a percentage or unit of time, is impractical from a survey perspective, and likely to (wildly) inflate response bias originating in respondent-related error, for all but the most numerate of target populations[13].

Despite our efforts to resolve the problem through the application of innovative techniques (e.g. the principle of maximum entropy detailed in section 5.4.4), the results indicate poor model performance. The model's lack of predictive capacity is likely a function of the limited cardinality and available imprecise information on mixed gender phone usage. It is important to note that this situation of imprecise information may be expected to be present in similar future surveys. Hence, a methodological modelling answer (as the one discussed in this work) is likely to remain an essential component to exploit this type of imperfect available information.

While the 'mismatch' between optimal and feasible measurements of SIM sharing may be expected to be lessened by the development of better proxies, it is unlikely to be wholly eliminated. For this reason, to progress the field, it will be necessary to develop or adapt methodological approaches capable of exploiting imperfect proxy data on SIM sharing practices. While theoretically appealing, 'survey linkage' does not offer a panacea. Alternative research designs, incorporating experimental approaches, might also be valuably explored.

[13]Panel survey methods may offer a promising avenue for future research. Subsequent to panel recruitment, a sequence of follow-up contacts, timed to assess the presence of systematic variation in the characteristics of the call recipient, would provide an objective indication of the presence and timing of SIM usage.

References

Almaatouq, A., Prieto-Castrillo, F., Pentland, A., (2016) Mobile Communication Signatures of Unemployment arXiv:1609.01778v1 [cs.SI] 6 Sep 2016

Arouri, M.E.H., Youssef, A.B., Nguyen-Viet, C., Soucat, A., 2014. Effects of urbanization on economic growth and human capital formation in Africa. PGDA Working https://halshs.archives-ouvertes.fr/halshs-01068271

Banerjee, S., Gelfand, A.E., Polasek, W., 2000. Geostatistical modelling for spatial interaction data with application to postal service performance. Journal of statistical planning and inference 90(1), 87-105. https://doi.org/10.1016/S0378-3758(00)00111-7

Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264): 1073-1076

Bosco, C., Tejedor Garavito, N., de Rigo, D., Tatem, A., Pezzulo, C., Wood, R., Chamberlain, H. and Bird, T., 2018. Geostatistical Tools to Map the Interaction between Development Aid and Indices of Need. AidData working paper, 49.

Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., et al., 2017. Exploring the high-resolution mapping of sex-disaggregated development indicators. Journal of The Royal Society Interface, 14(129). https://doi.org/10.1098/rsif.2016.0825

Bosco, C., Sander, G., 2015. Estimating the effects of water-induced shallow landslides on soil erosion. IEEE Earthzine 7(2), 910137+. http://earthzine.org/?p=910137 , https://doi.org/10.1101/011965 , INRMM-MiD:13455081

Bosco, C., de Rigo, D., Dijkstra, T. A., Sander, G., Wasowski, J., 2013. Multi-scale robust modelling of landslide susceptibility: regional rapid assessment and catchment robust fuzzy ensemble. IFIP Advances in Information and Communication Technology 413, 321-335, ISSN:1868-4238. doi:10.1007/978-3-642-41151-9_31

Burgert, C.R., Colston, J., Roy, T., Zachary, B., 2013. Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys. DHS Spatial Analysis Reports No. 7 Calverton, Maryland, USA: ICF International.

Busse, B. and Fuchs, M., 2014. Recruiting respondents for a mobile phone panel: The impact of recruitment question wording on cooperation, panel attrition, and nonresponse bias. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10(1), 21-30. http://dx.doi.org/10.1027/1614-2241/a000064

Busse, B. and Fuchs, M., 2013. Prevalence of Cell Phone Sharing. Survey Methods: Insights from the Field. Retrieved from http://surveyinsights.org/?p=1019

Busse, B. and Fuchs, M., 2012a. The components of landline telephone survey coverage bias:The relative importance of no-phone and mobile-only populations. Quality & Quantity, 46(4), 1209-1225.

Busse, B. and Fuchs, M., 2012b. Recruiting Respondents for a Mobile Phone Panel. The Impact of Recruitment Question Wording on Cooperation, Panel Attrition and Nonresponse Bias Methodology.

Deichmann, U., Shilpi, F. and Vakis, R., 2009. Urban Proximity, Agricultural Potential and Rural Non-farm Employment: Evidence from Bangladesh. World Development, 37(3). 645-660. doi:10.1016/j.worlddev.2008.08.008

De Martino, A. and De Martino, D., 2018. An introduction to the maximum entropy approach and its application to inference problems in biology." Heliyon 4(.4), e00596+. https://doi.org/10.1016/j.heliyon.2018.e00596

de Rigo, D. 2019. Estimating maximum-entropy implicit frequency from categorical information: the module "categorical2freq" of the Mastrave modelling library. In: Semantic Array Programming with Mastrave - Introduction to Semantic Computational Modelling.

de Rigo, D., 2015. Study of a collaborative repository of semantic metadata and models for regional environmental datasets' multivariate transformations. Ph.D. thesis, Politecnico di Milano, Milano, Italy. http://hdl.handle.net/10589/101044 , INRMM-MiD: 13769492

de Rigo, D., Corti, P., Caudullo, G., McInerney, D., Di Leo, M., San-Miguel-Ayanz, J., 2013. Toward open science at the European scale: Geospatial Semantic Array Programming for integrated environmental modelling. Geophysical Research Abstracts 15, 13245+. https://doi.org/10.6084/m9.figshare.155703

de Rigo, D.: Semantic Array Programming with Mastrave – Introduction to Semantic Computational Modelling, The Mastrave project, available at: http://mastrave.org/doc/MTV-1.012-1 (last access: November 2014), 2012a.

de Rigo, D.: Semantic array programming for environmental modelling: application of the Mastrave library, in: International Congress on Environmental Modelling and Software: Managing Resources of a Limited Planet, Leipzig, Germany, 1–5 July 2012, 1167–1176, 2012b.

de Rigo, D. 2012c. Detecting general multi-dimensional nonlinear correlations: the module "dist_corr" of the Mastrave modelling library. In: Semantic Array Programming with Mastrave - Introduction to Semantic Computational Modelling. http://mastrave.org/doc/mtv_m/dist_corr

de Rigo, D., Castelletti, A., Rizzoli, A.E., Soncini-Sessa, R., Weber, E., 2005. A selective improvement technique for fastening neuro-dynamic programming in water resources network management. In: Proceedings of the 16th IFAC World Congress (IFAC Praha 2005). http://folk.ntnu.no/skoge/prost/proceedings/ifac2005/Papers/Paper4269.html , http://hdl.handle.net/11311/255236,INRMM-MiD:10793225.

de Rigo, D., Rizzoli, A. E., Soncini-Sessa, R., Weber, E., Zenesi, P., 2001. Neuro-dynamic programming for the efficient management of reservoir networks. In: Proceedings of MODSIM 2001, International Congress on Modelling and Simulation. Vol. 4. Modelling and Simulation Society of Australia and New Zealand, pp. 1949-1954. https://doi.org://:10.5281/ZENODO.7481

Eaton, J. W., Bateman, D., and Hauberg, S., 2008. GNU Octave Manual Version 3. A high-level interactive language for numerical computations, Network Theory Limited, ISBN: 0-9546120-6-X.

Environmental Systems Research Institute (ESRI), 2015. ArcGIS Release 10.4. Redlands, CA.

Frias-Martinez, V., Frias-Martinez, E. and Oliver, N., 2010. A gender-centric analysis of calling behavior in a developing economy using call detail records. In: 2010 AAAI Spring Symposium Series.

Gething, P., Tatem, A., Bird, T., Burgert-Brucker, C.R., 2015. Creating spatial interpolation surfaces with DHS data. Spatial Analysis Reports, no. 11. Rockville, MD: ICF International.

Ghandour L.A., El Hayek G.Y., Mehio Sibai A., 2019. Cell Phone Survey. In: Liamputtong P. (eds) Handbook of Research Methods in Health Social Sciences. Springer, Singapore;

Goodman, J., 2005. Linking mobile phone ownership and use to social capital in rural South Africa and Tanzania, The Vodafone Policy Paper Series, vol. 3

Government of Nepal (2015). The Constitution of Nepal. Kathmandu: Govt of Nepal. Available from: http://www.wipo.int/edocs/lexdocs/laws/en/np/np029en.pdf, (Accessed: 2nd September 2019)

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

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

IOM (2016). Barriers to women's land and property access and ownership in Nepal. Kathmandu: International Organization for Migration.
Available online at: https://www.iom.int/sites/default/files/our_work/DOE/LPR/Barriers-to-Womens-Land-Property-Access-Ownership-in-Nepal.pdf (Accessed 2nd September 2019)

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L. and de Montjoye, Y.A., 2017. Improving official statistics in emerging markets using machine learning and mobile phone data. EPJ Data Science, 6(1), p.3. https://doi.org/10.1140/epjds/s13688-017-0099-3

Kreinovich, V.Y., 1991. Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem. Neural Networks, 4 (3), 381-383. https://doi.org/10.1016/0893-6080(91)90074-f

Lee, S., Ryu, J.H., Kim, I.S., 2007. Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. Landslide, 4, 327-338. https://doi.org/10.1007/s10346-007-0088-x

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lopez, A., 2000. The South Goes Mobile. UNESCO COURIER

Matinga, M.N., Gill, B., Winther, T., 2019. Rice Cookers, Social Media, and Unruly Women: Disentangling Electricity's Gendered Implications in Rural Nepal. Front. Energy Res. 6:140. https://doi.org/10.3389/fenrg.2018.00140

Ministry of Health, Nepal, New ERA and ICF, 2017. Nepal Demographic and Health Survey 2016. Kathmandu, Nepal: MOH/Nepal, New ERA/Nepal, and ICF.

Murtaugh, P.A. 2009. Performance of several variable selection methods applied to real ecological data. Ecol. Lett. 12, 1061–1068. https://doi.org/10.1111/j.1461-0248.2009.01361.x

Perez-Heydrich C, Warren JL, Burgert CR, Emch ME., 2013 Guidelines on the use of DHS GPS data. Spatial Analysis Reports, no. 8. Calverton, MD: ICF International.

Pradhan, B., Lee, S., 2009. Landslide risk analysis using artificial neural network model focusing on different training sites. International Journal of Physical Sciences 4(1), 1-15

Press, S. J., 2002. Hierarchical Bayesian Modeling. in Subjective and Objective Bayesian Statistics: Principles, Models, and Applications, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9780470317105.ch14

QGIS Development Team, 2018. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

R Development Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rangaswamy, N., Singh, S., 2009. Personalizing the shared mobile phone, Human–Computer Interaction Institute

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. Journal of the Royal Statistical

Society: Series B (Statistical Methodology) 71(2), 319-392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Samuel, J. Shah, N. Hadingham, W., 2005. Mobile communications in South Africa, Tanzania & Egypt: results from community & business surveys, The Vodafone Policy Paper Series, vol. 3

Šćepanović S, Mishkovski I, Hui P, Nurminen JK, Ylä-Jääski A (2015) Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator. PLoS ONE 10(4): e0124160. doi:10.1371/journal.pone.0124160

Schmid M.D., 2009. A neural network package for Octave. User's Guide Version: 0.1.9.1. http://www.plexso.com/61_octave/neuralNetworkPackageForOctaveUsersGu.pdf

Sebusang, S., Masupe, S., Chumai, J., 2005. Botswana, in: A. Gillwald (Ed.), Towards an African E-index: ICT Access & Usage, The LINK Centre, Wits University School of Public & Development Management, Johannesburg, 2005

Secomandi, N., 2000. Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands. Computers & Operations Research, 27 (11–12), 1201–1225. https://doi.org/10.1016/S0305-0548(99)00146-X

Stallman, R. M., 2009. Viewpoint: Why "open source" misses the point of free software. Communications of the ACM 52 (6), 31-33. https://doi.org/10.1145/1516046.1516058

Stork, C., 2005. Namibia, in: A. Gillwald (Ed.), Towards an African E-index: ICT Access & Usage, The LINK Centre, Wits University School of Public & Development Management, Johannesburg.

Sundsøy, P., 2016a. Can Mobile Usage predict illiteracy in a developing country? arXiv Preprint: 1607.01337

Sundsøy, P., Bjelland, J., Reme, B., Iqbal, A., Jahani, E., 2016b. Deep learning applied to mobile phone data for Individual income classification. In: ICAITA 2016 International Conference on Artificial Intelligence and applications

Sundsøy, P., Bjelland, J., Reme, B-A,. Jahani, E., Wetter, E., Bengtsson, L., 2016c. Estimating individual employment status using mobile phone network data. https://arxiv.org/abs/1612.03870

Sundsøy P., Bjelland J., Reme BA., Jahani E., Wetter E., Bengtsson L., 2017. Towards Real-Time Prediction of Unemployment and Profession. In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol 10540. Springer

Székely, G.J., Rizzo, M.L. and Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. The annals of statistics, 35(6), pp.2769-2794. https://doi.org/10.1214/009053607000000505

Tey, N. and Lai, S., 2013. Correlates of and Barriers to the Utilization of Health Services for Delivery in South Asia and Sub-Saharan Africa. The Scientific World Journal.1-11. https://doi.org/10.1155/2013/423403

Tukey, J.W., 1958. Bias and confidence in not-quite large samples. Annals of Mathematical Statistics, 29, 614.

UNDP, 2018. Human development Indices and Indicators. 2018 statistical update. 123 pp. New York, UNDP.

Valliant, R., Dever, J. A., Kreuter, F. (2012) Practical tools for designing and weighting survey samples. London: Springer

Watmough, G., Atkinson, P., Hutton, C., 2013. Exploring the links between census and environment using remotely sensed satellite sensor imagery. Journal of Land Use Science, 8(3) , 284-303. https://doi.org/10.1080/1747423X.2012.667447

Watson, S.K., Elliot, M., 2016. Entropy Balancing: A maximum-entropy reweighting scheme to adjust for coverage error. Quality & Quantity. doi: 10.1007/s11135-015-0235-8

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79. https://doi.org/10.3354/cr030079

Wright-Steenson, M., Donner, J., 2009. Beyond the Personal and Private: Modes of Mobile Phone Sharing in Urban India. in The Reconstruction of Space and Time: Mobile Communication Practices, Eds. Rich Ling and Scott W. Campbell, New Brunswick: Transaction Publishers, 231 - 250

Zere, J.E., Oluwole, D., Kirigia, J.M., Mwikisa, C.N., Mbeeli, T., 2011. Inequities in skilled attendance at birth in Namibia: a decomposition analysis. BMC Pregnancy and Childbirth. 11: 34. 1-10. https://doi.org/10.1186/1471-2393-11-34

Appendix A

Table A1.1. Set of subscriber CDR features extracted for gender classification.

Data type	Feature name	Description
Voice call	Displacement	Displacement from subscribers monthly Home Location, measured in km
Voice call	Nocturnal Calls	Percentage of calls placed at night (between 8pm and 4am)
Voice call	Incoming call count	Count of subscriber's incoming calls
Voice call	Outgoing call count	Count of subscriber's outgoing calls
Voice call	Incoming call duration	Average duration of a subscriber's incoming calls
Voice call	Outgoing call duration	Average duration of a subscriber's outgoing calls
Voice call	Subscriber out-degree	Number of unique subscribers which a subscriber has made calls to
Voice call	Subscriber in-degree	Number of unique subscribers which a subscriber has received calls from
Voice call	Total international calls	Number of international calls made or received by a subscriber
Voice call	Outgoing international calls	Number of international calls made by a subscriber
Voice call	Incoming international calls	Number of international calls received by a subscriber
Voice call	Radius of gyration	Mobility measure
Voice call	Unique location counts	Number of unique cell towers from which a subscriber has made or received a call
Voice call	Diameter of influence	Maximum distance between all the towers used by a subscriber
Voice call	Distance travelled	Average distance travelled
Voice call	Interevent time	Average time between call events
Voice call	Incoming interevent time	Average time between incoming call events
Voice call	Outgoing interevent time	Average time between outgoing call events
Topup	Total recharge amount	Total topup expenditure per month
Topup	Total recharge counts	Number of topup transactions per month
Topup	Average recharge amount	Average topup transaction value per month
Topup	Median recharge amount	Median topup transaction value per month
Topup	Minimum recharge amount	Minimum topup transaction value per month
Topup	Maximum recharge amount	Maximum topup transaction value per month
Topup	Std deviation recharge amount	Std deviation topup transaction value per month

Table A1.2 Set of aggregate subscriber CDR features extracted for development indicator estimation.

Data type	Feature name	Aggregation over subscribers					Description
		sum	median	mean	std	proportion	
Voice call	Home location	✗					Modal daily location of a subscriber
Voice call	Percentage nocturnal calls		✗	✗	✗		Percentage of calls placed at night (between 8pm and 4am)
Voice call	Incoming call count	✗	✗	✗	✗		Count of a subscriber's incoming calls
Voice call	Outgoing call count	✗	✗	✗	✗		Count of a subscriber's outgoing calls
Voice call	Total call count	✗	✗	✗	✗		Count of all a subscriber's calls
Voice call	Proportion outgoing calls		✗	✗	✗		Ratio of a subscriber's outgoing call count to total call count
Voice call	Total incoming call duration	✗	✗	✗	✗		Summed durations of a subscriber's incoming calls
Voice call	Mean incoming call duration	✗	✗	✗	✗		Mean duration of a subscriber's incoming calls
Voice call	Median incoming call duration	✗	✗	✗	✗		Median duration of a subscriber's incoming calls
Voice call	Total outgoing call duration	✗	✗	✗	✗		Summed durations of a subscriber's outgoing calls
Voice call	Mean outgoing call duration	✗	✗	✗	✗		Mean duration of a subscriber's outgoing calls
Voice call	Median outgoing call duration	✗	✗	✗	✗		Median duration of a subscriber's outgoing calls
Voice call	Number of contacts	✗	✗	✗	✗		Number of contact of a subscriber
Voice call	Subscriber out-degree	✗	✗	✗	✗		Number of unique subscribers which a subscriber has made calls to
Voice call	Subscriber in-degree	✗	✗	✗	✗		Number of unique subscribers which a subscriber has received calls from
Voice call	Entropy of contacts	✗	✗	✗	✗		Entropy of a subscriber's contacts
Voice call	Number of interactions	✗	✗	✗	✗		Number of interactions by a subscriber with their contacts
Voice call	Median number of interactions per contact	✗	✗	✗	✗		Median number of interactions by a subscriber with each contact
Voice call	Mean number of interactions per contact	✗	✗	✗	✗		Mean number of interactions by a subscriber with each contact
Voice call	Percentage pareto interactions		✗	✗	✗		The Pareto proportion for a subscriber's interactions i.e. the fraction of a subscriber's contacts who account for 80% of their interactions.
Voice call	Radius of gyration		✗	✗	✗		Mobility measure
Voice call	Number of places	✗	✗	✗	✗		Number of unique cell sites from which a subscriber has made or received a call
Voice call	Entropy of places		✗	✗	✗		Entropy of cell sites from which a subscriber has made or received a call
Voice call	Percentage interactions at home		✗	✗	✗		Percentage of calls made or received by a subscriber at their home location

Voice call	Frequent places		✗	✗	✗		Number of locations that account for 80% of the locations where the subscriber is seen
Voice call	Location introversion for all interactions	✗					Number of all interactions made by a subscriber at a cell site at which the interaction counterparty is also located
Voice call	Location introversion for all incoming interactions	✗					Number of all incoming interactions received by a subscriber at a cell site at which the interaction counterparty is also located
Voice call	Location introversion for all outgoing interactions	✗					Number of all outgoing interactions made by a subscriber at a cell site at which the interaction counterparty is also located
Device	Device brand					✗	Brand of the device most used by a subscriber. Device brand distribution is expressed as a proportion of the total number of subscribers for each cell site. We use the most common device of a subscriber during January 2015, and assign to the Home Location of the subscriber. We only consider brands with greater than 10000 subscribers.
Device	Device Operating System					✗	Operating System (OS) of the device most used by a subscriber. Device OS distribution is expressed as a proportion of the total number of subscribers for each cell site. We use the most common device of a subscriber during January 2015, and assign to the Home Location of the subscriber. We only consider OSs with greater than 10000 subscribers.
Device	Device size		✗	✗	✗		Diagonal size in millimetres of the device most used by a subscriber. We use the most common device of a subscriber during January 2015, and assign to the Home Location of the subscriber.
Device	Device weight		✗	✗	✗		Weight in grammes of the device most used by a subscriber. We use the most common device of a subscriber during January 2015, and assign to the Home Location of the subscriber.
Device	Device display size		✗	✗	✗		Diagonal size of the display resolution in pixels of the device most used by a subscriber. We use the most common device of a subscriber during January 2015, and assign to the Home Location of the subscriber.

Table A3.3: Definitions of the shortlisted indicators

Indicator		Definition	Denominator	Numerator
1.	Literacy	Literacy is defined as the ability to read a short sentence – wholly or in part – from a card. Cards are prepared in several languages but are non-exhaustive. Respondents for whom no appropriate language-card is available do not receive a literacy result. Blind and visually impaired respondents are similarly excluded. Respondents with higher than secondary schooling are automatically recorded as literate, with no demonstration requirement.	Number of women aged 15-49 (n=12,848) Number of men aged 15-49 (n = 4,060)	Number of women with higher than secondary level schooling or who can read a short sentence – wholly or in part – from a card. (n = 9,020) Number of men with higher than secondary level schooling or who can read a short sentence – wholly or in part – from a card. (n = 3,631)
2.	Educational attainment	Educational attainment is defined as the highest level of schooling attended or completed:	Number of women aged 15-49 (n=12,862)	Number of women who attended secondary school (n = 6,435)

		1. No education 2. Incomplete primary 3. Complete primary 4. Incomplete secondary 5. Complete secondary 6. Higher	Number of men aged 15-49 (n = 4,063)	Number of men who attended secondary school (n = 2,872)
3.	Market labour participation	Market labour participation is defined as participation in the labour force within a 12 month reference period.The DHS defines 'work' broadly to include paid (whether cash and/or in-kind) and unpaid work. 'Own housework' (reproductive labour) is omitted.	Number of women aged 15-49 (n = 12,862) Number of men aged 15-49 (n = 4,063)	Number of women aged 15-49 who have undertaken market labour within a 12-month reference period. (n = 8809) Number of men aged 15-49 who have undertaken market labour within a 12-month reference period. (n = 3,476)
4.	Agriculture-based occupation	Agriculture-based occupation is defined by the respondent's principal occupation (agriculture) and employment status within a 12 month reference period. It includes paid (cash and in-kind) and unpaid work in agriculture, whether on an employed, family enterprise, or self-employed basis.	Number of women aged 15-49 who are currently undertaking market labour or who have undertaken market labour within a 12-month reference period. (n = 8809) Number of men aged 15-49 who are currently undertaking market labour or who have undertaken market labour within a 12-month reference period. (n = 3,476)	Number of women aged 15-49 whose primary occupation (excluding own housework) was agriculture- based, during the 12-month reference period. (n = 8809) Number of men aged 15-49 whose primary occupation (excluding own housework) was agriculture-based, during the 12-month reference period. (n = 8809)
5.	Stunting in childhood	Stunting – impaired growth due to chronic malnutrition - is defined as a length/height- for-age z score more than two standard deviations below the World Health Organization (WHO) Child Growth Standards median.	Number of living female children born 0-59 months prior to survey (n = 1,172) Number of living male children born 0-59 months prior to survey (n=1,274)	Number of female children with a length/height-for- age z-score more than -2.0 SD below the WHO Child Growth Standards median (n = 415) Number of male children with a length/height-for- age z-score more than -2.0 SD below the WHO Child Growth Standards median (n =468)
6.	Complacency about GBV against women	Complacency about GBV against women is defined as a stated belief that a husband is justified in hitting or beating his wife for one or more of the following: 1) Burning food 2) Arguing with him 3) Going out without telling him 4) Neglecting the children 5) Refusing sexual intercourse with him	Number of women aged 15-49 (n = 12,862) Number of men aged 15-49 (n = 4,063)	Number of women aged 15-49 who agree that a husband is justified in hitting or beating his wife in at least one of the listed circumstances. (n = 3,693) Number of men aged 15-49 who agree that a husband is justified in hitting or beating his wife in at least one of the

				listed circumstances. (n = 1,051)
7.	Births in health facilities	Number of live deliveries in a government or private health facility.	Number of live births to women aged 15-49, within five- year reference period. (n = 5,038)	Number of live births to women aged 15-49 delivered in a health facility, within five-year reference period. (n = 2,809)

Table A1.4. GIS and remote sensing dataset used to create the geospatial covariates

Dataset	Description	Data type	Year
Accessibility	Accessibility to cities and friction surface from Malaria Atlas Project	Continuous	2015
Protected Areas	WDPA protected terrestrial/maritime areas	Vector	2016
Land cover	European Space Agency (ESA) land cover data	Categorical	2015
Ethnicity	Geo-referencing Ethnic Power Relations dataset	Vector	1946-2013
Global Human Settlements	European Commission Global Human Settlements (GHS) layer	Continuous	2014
Global Urban Footprint	DLR Global Urban Footprint (GUF) layer	Continuous	2011
Urban Extents	CIESIN Global Rural Urban Mapping Project (GRUMP)	Categorical	2000
Population	Worldpop population count	Continuous	2015
Elevation/Slope	Calculated from Shuttle-Radar Topography Mission (SRTM) DEM data	Continuous	2000
Precipitation	Precipitation from Centre for Environmental Data Analysis (CEDA)	Continuous	2001 - 2014
Potential Evapotranspiration	PET from Centre for Environmental Data Analysis (CEDA)	Continuous	2001 - 2014
Temperature	Temperature from Centre for Environmental Data Analysis (CEDA)	Continuous	2001 - 2014
Roads	Open Streetmap roads vector dataset	Vector	2018
Rivers	Open Streetmap rivers vector dataset	Vector	2018
Schools	Open Streetmap school locations dataset	Vector	2018
Crop Dominance	NASA GFSAD crop dominance, irrigated cropland extracted	Continuous	2010
Crop Suitability	GLUES overall crop suitability	Continuous	2011-2040
Health Facilities	Health facilities locations from Humanitarian Data Exchange (HDX)	Vector	2018
Nighttime Lights	Visible Infrared-Imaging Radiometer Suite (VIIRS) night-time lights data	Continuous	2016
NDVI	MODIS Normalised Difference Vegetation Index	Continuous	2009-2017

EVI	MODIS Enhanced Vegetation Index	Continuous	2009-2017
MIR	MODIS Middle Infra-red reflectance	Continuous	2009-2017
NPP	MODIS Net Primary Productivity	Continuous	2009-2014
GPP	MODIS Gross Primary Productivity	Continuous	2009-2017
Total Evapotranspiration	MODIS Total Evapotranspiration	Continuous	2009-2017
Potential Evapotranspiration	MODIS Potential Evapotranspiration	Continuous	2009-2017

Geospatial Covariates

Accessibility and Friction Surface

The Malaria Atlas project produced a 1km grid of accessibility to the nearest city for the year of 2015, where they quantify global accessibility to high density urban centres as measured by travel time to the nearest densely-populated area. The friction surface layer enumerates land-based travel speeds for all pixels, incorporating topographic conditions and features such as rivers, railways and national borders. Both raster datasets were selected for use in the analysis. Further details can be found at https://map.ox.ac.uk/research-project/accessibility_to_cities/

Protected Areas

Protected terrestrial and maritime areas from the world protected areas database (WDPA) vector dataset of protected areas (2016) were obtained (https://www.protectedplanet.net/c/world-database-on-protected-areas).
The dataset provides a polygon and point layer, with buffer zones around the protected areas. These buffer zones and points were removed and the protected areas were transformed to produce a binary raster dataset. Distance to the areas was calculated and the focal statistics tool was applied to produce a final set of continuous raster layers.

Land cover

We downloaded classified land cover for the year 2015 from European Space Agency (ESA) Climate Change Initiative (CCI) (https://www.esa-landcover-cci.org/?q=node/164). Data was resampled and reclassified to produce two sets of layers, the first containing cropland classes (both rain-fed and irrigated) and the second containing these cropland classes with the addition of mosaic cropland. The data was aggregated to 0.008333 dec. degrees using the maximum method for assigning cell values. A distance raster and smoothed raster using the focal statistics tool were created for both sets of land cover classes.

Ethnicity

Information on ethnicity in Nepal was extracted from the Ethnic Power Relations (EPR) geo-referenced dataset (https://icr.ethz.ch/data/epr/core/). This identifies politically relevant

groups and their access to state power in every country of the world from 1946-2013. It is provided in vector format, so individual ethnic groups were extracted for Nepal and converted to raster for analysis, where distance and focal statistics were calculated.

Global Human Settlements

The Global Human Settlements Area Layer (GHSL) from the Joint Research Centre was used which identifies built-up area presence from Landsat satellite data. The 2014 continuous dataset was selected, where three re-classifications were applied to separate built-up areas. Data was reprojected, resampled and 0.1 was used as a standard threshold to extract built-up areas of interest, with 0.2 and 0 also selected for comparison. Distance and focal statistics were calculated to produce continuous rasters for each reclassification. Further details can be found at https://data.jrc.ec.europa.eu

Global Urban Footprint

The global urban footprint (GUF) dataset created by DLR Earth Observation Centre (EOC) was obtained (https://www.dlr.de/dlr/). This defines built-up areas as regions featuring man-made building structures with a vertical component using satellite data where values of 255 represent built-up areas. Built-up areas were extracted, reclassified and the dataset was aggregated using the mean. Distance from built-up areas was also calculated for use in the analysis.

GRUMP Urban Extent

The GRUMP Urban Extents grid was obtained from Columbia University Centre for International Earth Science Information Network (CIESIN). This distinguishes urban and rural areas in binary raster format based on population count, settlement points and night-time lights data for the year 2000. The data was reclassified to produce a binary raster of urban extent, where distance and focal statistics were applied. Further details on the dataset can be found at http://sedac.ciesin.columbia.edu/data/set/grump-v1-urban-extents

Population

Population count data was utilised from existing WorldPop datasets (http://www.worldpop.org.uk/). WorldPop population per pixel (2015 UN-adjusted) was obtained, which provides gridded population counts adjusted to match UN population estimates, created using a random forest estimation technique. Population was aggregated using the sum of cell values.

Elevation and Slope

Elevation data for Nepal was obtained using the Shuttle-Radar Topography Mission (SRTM) digital elevation model (DEM), downloaded from CGIAR (http://srtm.csi.cgiar.org/). This continuous dataset is provided in tiles at 90m spatial resolution. In ArcGIS, tiles were mosaicked to produce an elevation layer and slope was calculated from this.

CEDA Climate Variables

Data obtained from the Centre for Environmental Data Analysis (CEDA) (http://catalogue.ceda.ac.uk/) was used to extract three variables, precipitation, potential evapo-transpiration and temperature for Nepal. Annual mean monthly values were calculated for each of these variables for the years 2001-2014, where R was used to download, pre-process and calculate these averages.

Distance to Roads

Distance to roads was calculated using Open Street Map (OSM) data extracted for Nepal from https://extract.bbbike.org/. Main road classes were selected for use in the distance covariate which included those classified as primary, secondary or tertiary. An additional layer including smaller residential roads was also created. Distances were calculated in ArcGIS.

Distance to Rivers

Distance to rivers was calculated using OSM data on waterways extracted for Nepal from https://extract.bbbike.org. Distances were calculated in ArcGIS.

Distance to Schools

Distance to schools was calculated using OSM amenity data which was extracted for Nepal as point and polygon locations from https://overpass-turbo.eu/ . All data was converted into a single point dataset before calculating distances in ArcGIS.

Crop Dominance

The global crop dominance dataset from NASA Global Food Security Support Analysis Data (GFSAD) was selected which is created using satellite data for the year 2010 at a 1km resolution. Irrigated cropland areas were extracted for use where distance was calculated and focal statistics applied to produce continuous raster layers. Further details can be found at

Crop Suitability

The overall crop suitability from GLUES was obtained (http://geoportal-glues.ufz.de/stories/globalsuitability.html) for the years 2011-2040, shows the potential number of suitable crop cycles for 16 crops, considering rain-fed and irrigation on currently irrigated areas. The raster was clipped and snapped to the Nepal mask for use in analysis.

Distance to Health Facilities

Distance to healthcare facilities was calculated using point location data produced by the Survey Department of Nepal and World Health Organisation (WHO), obtained from the

humanitarian data exchange (HDX) (https://data.humdata.org/dataset/nepal-health-facilities-cod). Distances were calculated in ArcGIS.

VIIRS Nighttime Lights

Visible Infrared-Imaging Radiometer Suite (VIIRS) night-time lights data from on-board the NPP-Suomi satellite was selected for the area of Nepal. NOAA provides both annual and monthly composites of average radiance, the most recent annual composite available for 2015 was selected and monthly composites for the year 2016 were obtained to create an annual average using a script in R. To create the layer for 2016, files that exclude any data impacted by stray light and the cloud free days composites were utilised. Further details can be found at https://www.ngdc.noaa.gov/eog

MODIS Vegetation Indices and Mid Infrared reflectance

The MOD13Q1 Version 6 product provides a Vegetation Index (VI) value at a per pixel basis. There are two primary vegetation layers that include the NDVI and the Enhanced Vegetation Index (EVI), which are provided globally as 16 day composites at a 250m resolution. These were extracted for Nepal between December 2009 and January 2017, along with the MIR. R scripts were used to pre-process the data and calculate a range of statistics including the mean, median, minimum, maximum, sum and sum per year. Final mosaicking, reprojecting and resampling was carried out in Arcmap. Further product details can be found at https://lpdaac.usgs.gov/products/mod13q1v006/

MODIS Gross Primary Productivity and Net Primary Productivity

The MOD17A2H MODIS/Terra version 6 product was utilised which provides the gross primary productivity as a cumulative 8-day composite at a 500m resolution. It is based on the radiation-use efficiency concept and data was extracted between December 2009 and January 2017 for Nepal. The MOD17A3H MODIS/Terra annual Net Primary Productivity product was obtained between the years 2009-2014 as annual composites at 500m. The NPP is derived from the sum of the PSN products which are produced by differencing the GPP and Maintenance Respiration. R scripts were used to pre-process the data and calculate a range of statistics including the mean, median, minimum, maximum, sum and sum per year. Final mosaicking, reprojecting and resampling was carried out in Arcmap. Further product details can be found at https://lpdaac.usgs.gov/products/mod17a2hv006/

MODIS total evapotranspiration and total potential evapotranspiration

ET kg/m^2 and PET kg/m^2 provided in the MOD16A2 MODIS/Terra version 6 product are provided as 8-day composites at 500m resolution, where pixel values are the sum of all eight days within this period. Data was extracted between December 2009 and January 2017 for use. R scripts were used to pre-process the data and calculate a range of statistics including the mean, median, minimum, maximum, sum and sum per year. Final mosaicking, reprojecting and

resampling was carried out in Arcmap. Further product details can be found at https://lpdaac.usgs.gov/products/mod16a2v006/

Maps of the uncertainty values for female literacy, engagement in agriculture and health facility births weighted by population and aggregated at local level

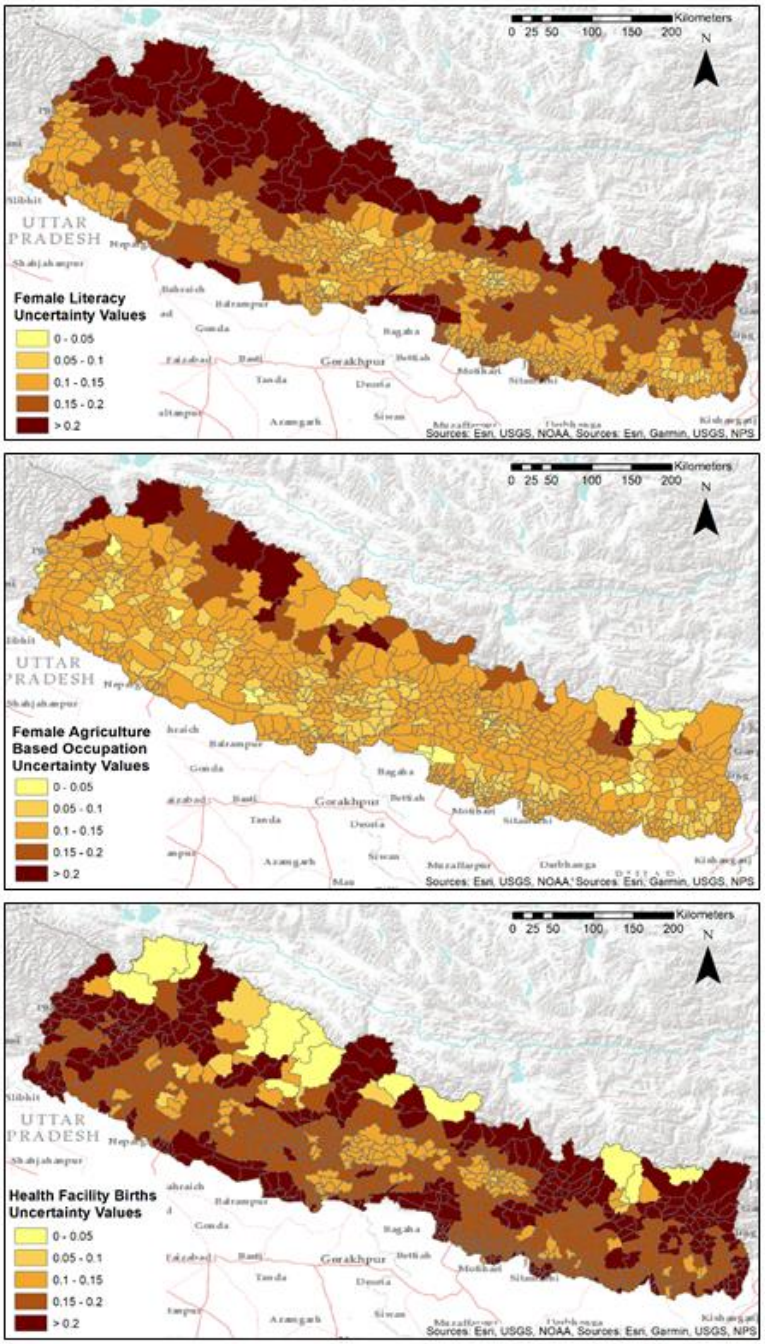

Figure A5.1. Maps of the uncertainty values for female literacy (top), female engagement in agriculture (middle row) and health facility births (bottom) created by exploiting GIS, RS and CDR data. The maps show the values of uncertainty weighted by population aggregated at local-level.

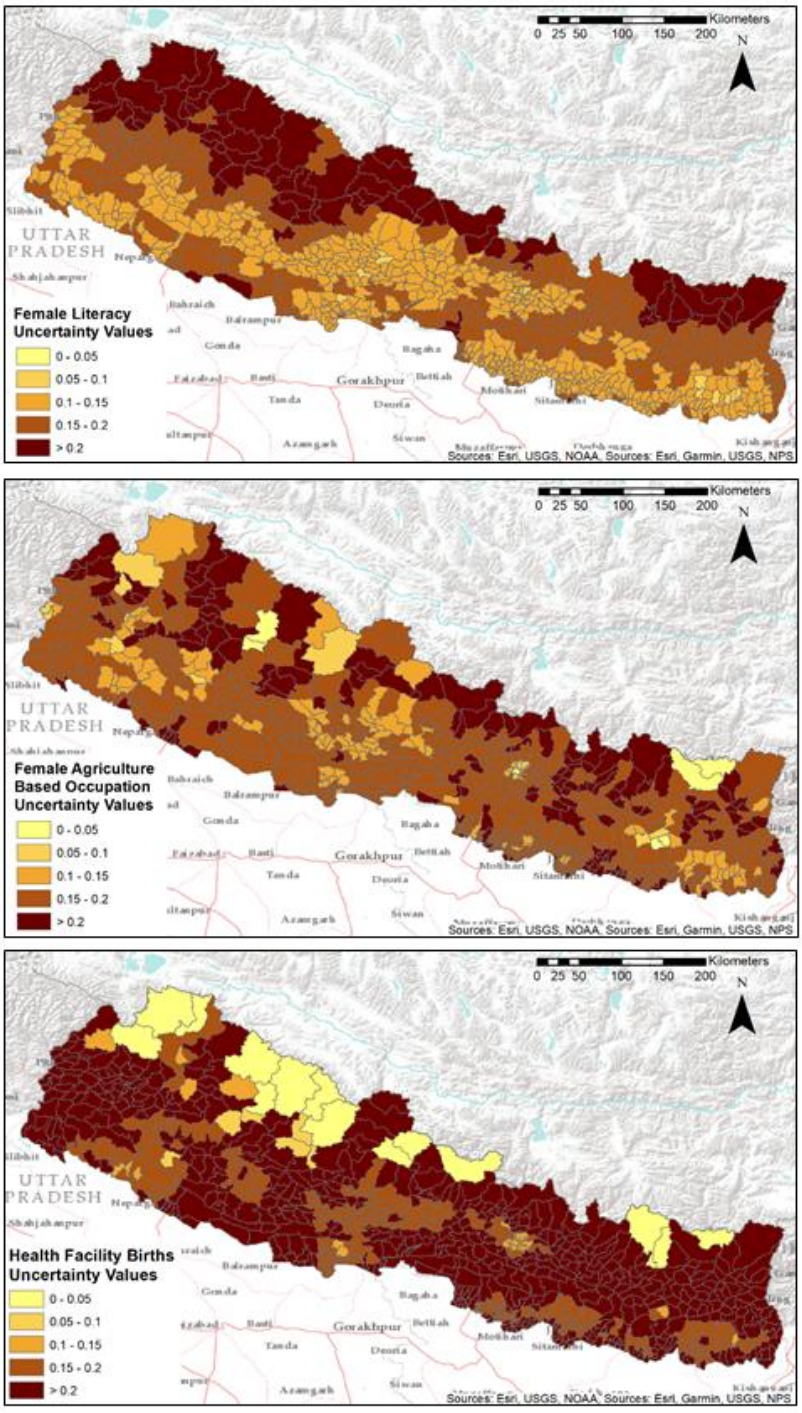

Figure A5.2. Maps of the uncertainty values for female literacy (top), female engagement in agriculture (middle row) and health facility births (bottom) created by exploiting only GIS and RS data. The maps show the values of uncertainty weighted by population aggregated at local-level.