






ARTICLE



<https://doi.org/10.1057/s41599-021-00953-0>

OPEN

# Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings

Jessica E. Steele <sup>1✉</sup>, Carla Pezzulo<sup>1</sup>, Maximilian Albert<sup>2</sup>, Christopher J. Brooks<sup>2</sup>, Elisabeth zu Erbach-Schoenberg<sup>1</sup>, Siobhán B. O'Connor<sup>1</sup>, Pål R. Sundsøy<sup>3</sup>, Kenth Engø-Monsen <sup>3</sup>, Kristine Nilsen<sup>1</sup>, Bonita Graupe<sup>4</sup>, Rajesh Lal Nyachhyon<sup>5</sup>, Pradeep Silpakar<sup>5</sup> & Andrew J. Tatem <sup>1,2</sup>

Call detail records (CDRs) from mobile phone metadata are a promising data source for mapping poverty indicators in low- and middle-income countries. These data provide information on social networks, call behavior, and mobility patterns in a population, which are correlated with measures of socioeconomic status. CDRs are passively collected and provide information with high spatial and temporal resolution. Identifying features from these data that are generalizable and able to predict poverty and wealth beyond a single context could promote broader usage of mobile data, contribute to a reduction in the cost of socioeconomic data collection and processing, as well as complement existing census and survey-based methods of poverty estimation with improved temporal resolution. This is especially important within the context of the sustainable development goals (SDGs), where poverty and related health indicators are to be reduced significantly across subnational geographies by 2030. Here we utilize measures of cell phone user behavior derived from three CDR datasets within a Bayesian modeling framework to map poverty and wealth patterns across Namibia, Nepal, and Bangladesh. We demonstrate five metrics of user mobility and call behavior that are able to explain between 50% and 65% of the variance in socioeconomic status nationally for these three countries. These key metrics prove useful in very different contexts and can be readily provided as part of an existing CDR platform or software package. This paper provides a key contribution in this regard by identifying such metrics relevant to estimating poverty. We highlight the inclusion of ancillary data and local context as an important factor in understanding model outputs when targeting poverty alleviation strategies.

<sup>1</sup>WorldPop, University of Southampton, Southampton, UK. <sup>2</sup>Flowminder Foundation, Stockholm, Sweden. <sup>3</sup>Telenor Research, Oslo, Norway. <sup>4</sup>Mobile Telecommunications Limited, Windhoek, Namibia. <sup>5</sup>Ncell Private Limited, Kathmandu, Nepal. ✉email: [steele.jessica@gmail.com](mailto:steele.jessica@gmail.com)

## Introduction

The first of the United Nations sustainable development goals (SDGs) is poverty eradication (United Nations General Assembly, 2015), and achievement of this goal depends on regular and reliable estimates of the number of people in poverty and where they live. This information can be difficult to attain yet is critically important to development agencies, foundations, NGOs, and governments working toward alleviating poverty within low- and middle-income countries (LMICs). Low socioeconomic status within a country is associated with significant health problems; for example, malaria, child mortality, and population growth have all been linked to poverty (Tusting et al., 2013; Målqvist, 2015; UNFPA, 2014). The geographic identification of poor populations at those at high-risk of poverty susceptibility is of paramount importance when developing measures to target the vulnerable. Detailed poverty maps that quantify the spatial distribution and magnitude of economic impoverishment are essential for progress toward poverty eradication, and the availability of high-quality, timely, and disaggregated data is necessary for evidence-based decision making for implementing the 2030 agenda (United Nations, 2017).

Subnational estimates of poverty are typically made using data from national censuses and household surveys within a small area estimation (SAE) modeling framework. This method utilizes the content detail of household surveys along with the coverage of the census to produce subnational estimates of the proportion of households living in poverty (Elbers et al., 2002a, 2002b; Hentschel et al., 1998). As censuses are a main input of SAE models, detailed and reliable estimates of poverty at high granularity would be possible with a timely and complete census for each country. However, the release of census data, and accompanying data on subnational unit boundaries, typically occurs every 10 years. Censuses can be delayed, missing, incomplete, unreliable, or unavailable for many low- and middle-income countries (LMICs), making the estimation of development indicators especially difficult in the highest burden areas. Furthermore, in some LMICs the administrative boundaries and thus spatial availability of census data can be too coarse to produce reliable subnational estimates, or accurate data on admin boundaries are altogether not available (Jerven, 2013). These factors have led researchers to explore new sources of data and methodologies for estimating socioeconomic status that are independent of data from censuses to meet the need for more frequent updates and finer spatial detail in estimating poverty.

One such source of data comes from features derived from call detail records (CDRs) collected by mobile network operators (MNOs). CDRs contain the metadata, but not the content, of communications between millions of people at a time. These data also often include information on user location and social ties, which combined can give insight into population-level movement and social networks (Blondel et al., 2015). Data on the amount and frequency of airtime recharges, i.e. top-ups, on mobile phones provides direct information on user consumption, which can be an important predictor of poverty measures (Steele et al., 2017). In addition, compared to the information collected via surveys, CDRs are largely considered free from the bias of self-reporting. Whereas information from observed behavior has been shown to differ from self-reported behavior due to the perception of subjects themselves (Eagle et al., 2009), CDRs are derived from only observed behavior, e.g. mobility records, calling patterns, top-up amounts and frequencies.

CDR data come with limitations and biases and it is important to consider these, as a dataset from one MNO in a country will

not necessarily reflect the general population. In many cases the ability to quantify self-selection bias in the mobile user population as compared to the general population is impracticable. However, there are some known factors that contribute to data bias and limitations when using CDRs as a proxy to study the general population. Studies have shown that mobile phone ownership, and thus data generation, is skewed toward the educated, males, urban populations, and wealthier people (Stork, 2011; Wesolowski et al., 2012). Furthermore, multiple telecom companies usually operate within a country and CDRs from a single operator represent only the portion of the population that comprises their market share. Individual access to a mobile phone is also dependent on mobile reception, the ability to afford a mobile handset and top-ups, and electricity to recharge the device (Stork, 2011). These factors vary geographically, and mobile phone ownership and coverage can be low in rural and remote areas. Moreover, not all CDR datasets comprise a year's worth of data, adding additional bias due to seasonal activities and mobility: studies have shown that time of year directly affects population densities and their consequent characteristics spatially, which influences and is reflected by CDR data (Wesolowski et al., 2017; zu Erbach-Schoenberg et al., 2016). All of these factors contribute to self-selection bias, where the poorest members of the population, and especially those with intersecting forms of social marginalization, may not have access to mobile phones, may not be included temporally due to seasonal activities, and are thus absent from the data.

Despite these drawbacks and biases, CDRs have been shown to provide useful information on the spatio-temporal variation in poverty and wealth. Studies using CDR data to infer socioeconomic status have formed a significant branch of work within the data for development domain, with digital technologies being used to design, implement, and monitor development projects (Data for Development, 2017; GMSA, 2016; GSMA, 2014; OPAL, 2017). Researchers are increasingly making use of the vast information present in CDRs to quantify socioeconomic status in individuals and populations at high spatial and temporal resolution (Blumenstock et al., 2015, 2010; Eagle et al., 2010; Frias-Martinez and Virseda, 2012; Njuguna and McSharry, 2017; Pokhriyal and Jacques, 2017; Smith-Clarke et al., 2014; Soto et al., 2011; Steele et al., 2017). However, to date, this has only been shown for individual countries. A more widespread adoption of mobile phone data as a component of poverty estimation will rely on several factors, including both the internal validity of context-specific studies, and the generalizability of these data across multiple settings. Furthermore, in order for MNOs to make their data available for social good, multiple criteria need to be met:

- Data privacy protection must be ensured, and legal guidelines followed,
- Methods used for calculating and aggregating metrics must be transparent and verifiable, and
- The additional burden on the MNO's resources on top of their core business must be minimal.

Organizations are working to coordinate these efforts and enable the inclusion of CDR-derived features in a shareable way to leverage and harness these data for development (e.g. OPAL (OPAL, 2017), United Nations Global Pulse (United Nations Global Pulse, 2018), Global Partnership for Sustainable Development Data (Global Partnership for SDGs, 2018), World Bank (World Bank, 2016), Data Pop Alliance (Data-Pop Alliance, 2018), etc.). Data access is one of the most challenging aspects of

utilizing CDRs for development goals and the availability of relevant user features—basic and advanced phone usage, handset type, revenue data, mobility and social network information, and top-ups—can vary greatly from country to country.

Here, we use a common set of CDR-derived features from Namibia, Nepal, and Bangladesh to comparatively estimate poverty within a robust modeling framework across three very different geographies in Africa and Asia. We first produced national-scale poverty maps using aggregate user features derived from all available mobile phone metadata within each individual country. We then produced national-scale poverty maps for each country using only aggregate user features that were available in all three countries, as we are interested in how a generalized set of easily replicable CDRs performs across settings in estimating asset-based poverty. Model performance was evaluated using out-of-sample cross-validation statistics (coefficient of determination ( $r^2$ ) and the root-mean-square-error (RMSE)) calculated on randomly selected test subset of data. We report on the CDR covariates, both generalizable and country-specific, that are significant predictors of poverty within and across all three countries. We further calculated the numbers of people living in poverty as predicted by each model in order to compare the spatial distributions of socioeconomic status. This allowed us to highlight ancillary data within a local context as being a key consideration when making use of big data for strategic development efforts and monitoring of SDGs.

**Methods**

All data used in this study were processed to ensure that projections, resolutions, and extents matched. We estimated the approximate reception area of each mobile phone tower via Voronoi tessellation (Okabe et al. 2009) and based the spatial scale of analysis on these coverage areas. Each Voronoi polygon was then assigned aggregate features based on the mean, sum, or mode of the corresponding CDR data. The household survey data were matched to the Voronoi based on the lat/long coordinate representing the centroid of each DHS cluster; where multiple clusters fell within the same polygon, we used the mean aggregate value.

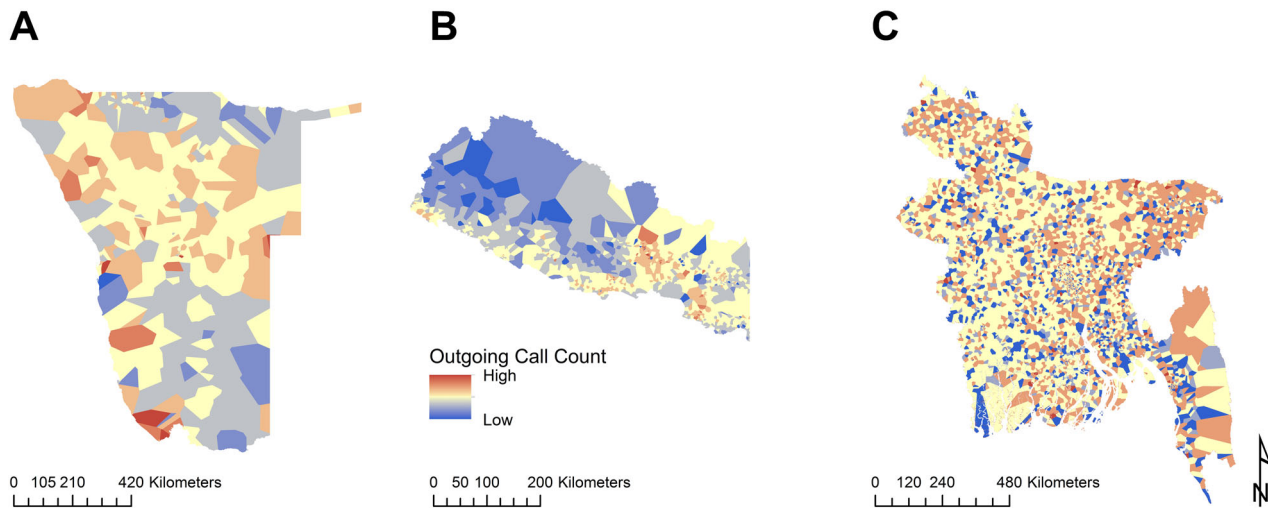
**Mobile phone data.** The CDR metrics used in this study were derived from network data provided by MTC Namibia, NCell in Nepal, and Grameenphone in Bangladesh. These countries were chosen based on data availability where agreements with MNOs overlapped HH survey data from the DHS. Table 1 provides the network details of MNO data including market share, penetration rate, subscribers, and number of Voronoi polygons used for this study. The primary purpose for the collection of CDRs by an MNO is to enable subscriber billing. The re-purposing of this data source for alternative use, as presented here, does not come without additional technical overheads, legal and regulatory challenges. Data availability at any given MNO is driven by numerous factors including, but not limited to, data warehouse resource availability, data retention policies, and the prioritization of operational concerns (see Supplementary Information, *Notes on processing, data biases, and limitations of CDR data* for more information).

All CDR indicators were calculated on an individual level and then aggregated up to the tower level based on the Voronoi cell representing the individual’s home location (as defined below and in Steele et al., 2017, SI Section A2). For example, Fig. 1 shows the outgoing call count for each Voronoi cell in the study countries. First, we calculated the outgoing call count for each user in the data; then the data were aggregated within each Voronoi polygon by calculating the mean outgoing call count for all users with

**Table 1 Summary table with the mobile phone penetration rate in each country, market share of each MNO, number of subscribers in each dataset, time period covered by the data, and number of towers.**

Country	Penetration rate, unique subscribers (Country)	Mobile network operator	Market share (MNO)	Number of SIMs (MNO)	Time period	Number of tower voronoi
Bangladesh	42% [Ref 1, Table 5]	Grameenphone	42.2% [Ref 2, p.2]	~48 million [Ref 3] [also: ref 2, p. 2]	November 2013–March 2014 <sup>a</sup>	6858
Namibia	51% [Ref 4, Fig. 3]	MTC	76%	~3 million	1 January 2013–31 December 2013	642
Nepal	39% [Ref 5, Table 5]	Ncell	49.3% [Ref 5, Table 4]	~12 million [Ref 6, Table 2]	1 January 2015–7 April 2015	2560

<sup>a</sup>Excluding January 2014. [Ref 1] [https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/02/Country\\_Overview\\_Bangladesh.pdf](https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/02/Country_Overview_Bangladesh.pdf). Accessed 11 Mar 2020. Barbara Aresé Lucini and Tim Hatt. Country overview: Bangladesh. GSMA Intelligence, published August 2014. [Ref 2] [https://www.grameenphone.com/sites/default/files/investor\\_relations/files/MAD%20202014.pdf](https://www.grameenphone.com/sites/default/files/investor_relations/files/MAD%20202014.pdf). Accessed 11 Mar 2020. Grameenphone Ltd.: 2nd Quarter 2014 Business Highlights. [Ref 3] <http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-january-2014>. [Ref 4] [https://web.archive.org/web/20180427174946/https://www.gsma.com/mobileeconomy/archive/GSMA\\_MIE\\_Sub-SaharanAfrica\\_2013.pdf](https://web.archive.org/web/20180427174946/https://www.gsma.com/mobileeconomy/archive/GSMA_MIE_Sub-SaharanAfrica_2013.pdf). Accessed 11 Mar 2020. Sub-Saharan Africa Mobile Economy 2013. GSMA Intelligence, 2013. [Ref 5] <https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2015/10/NCell-report-October-2015.pdf>. Accessed 11 Mar 2020. Rahul Shah, “Mobile for Development Utilities—Assessing the opportunity to improve energy access: Ncell in Nepal”, September 2015. [Ref 6] <http://nta.gov.np/wp-content/uploads/2017/11/NTA-MIS-95.pdf>. Accessed 5 Dec 2018. NTA MIS Report, Issue 75 vol 123, published April 2015.



**Fig. 1** Mean outgoing call count per cell tower for Namibia (A), Nepal (B), and Bangladesh (C). The coverage areas for the cell towers are approximated by Voronoi tessellation.

coincident home cells. This process is repeated for each covariate and only the aggregate features are used in covariate selection, model fitting, and prediction. To preserve user anonymity, the operators remove all personally identifying information from the data before analysis:

- (i) All customers are de-identified and only telecom employees have had access to any detailed data
- (ii) The processing of detailed CDR/top-up data resulted in aggregations of the data on a tower level granularity; the tower-level aggregation makes re-identification impossible.

Hence, the resulting aggregated dataset is anonymized and involves no personal data.

**Namibia.** CDR features were derived from the network data of MTC, the leading mobile phone provider in Namibia. The data set spans 12 months between 1 January and 31 December 2013 and contains 2,936,046 users. We calculated covariates for each individual and then grouped all individuals with the same home location based on their last call of the day. We conducted a sensitivity analysis on home location definition, as we are interested in capturing users at their home and not a workplace or other regularly visited location. We analyzed the following alternatives for defining user home locations: nighttime location (most used tower between 8 p.m. and 6 a.m. inclusive); most used tower irrespective of time; and location of the last call of the day. We eliminated the nighttime definition as ~130,000 users had no nighttime location and would have been omitted from our analyses. In comparing the other two definitions, we mapped users and ultimately chose last call of the day as it placed people in residential areas better than most frequently used tower. That is, the most used tower definition placed more users in central urban commercial areas than are reasonably expected to reside there. The last call of the day definition more consistently placed users where they are likely to reside in areas of non-commercial land use. We used raw data (individual call/text entries) for most of the covariates and unfilled daily locations to calculate the home location for each individual. Table 2 details the data processed for Namibia.

**Nepal.** CDR features were derived from the network data of NCell mobile phone metadata collected between 1 January and 7 April 2015. These data were processed into features of user mobility, social networks, basic phone usage, and selected phone

features (Table 2). Again, each user was assigned a home location based on their last call of the day, and the mean value of each indicator was calculated for all users sharing the same home tower and used in model fitting and prediction.

**Bangladesh.** CDR features were derived from the network data of Grameenphone (GP) mobile phone metadata collected over 4 months between November 2013 and March 2014. GP, the largest mobile network operator in Bangladesh, had 48 million customers at the time of the analysis. Table 2 details the data processed for Bangladesh used in this study. Each user was assigned a home location based on their most used tower, and the mean value of each indicator was calculated for all users sharing the same home tower. Further details are reported in the Supplemental Information, Section A.2 and Tables S1 and S2A of Steele et al. (2017).

**Geolocated survey data.** We utilized demographic and health surveys data from USAID (Rutstein and Rojas, 2006). These surveys are designed to collect household data on marriage, fertility, family planning, and other health indicators in nearly all lower income countries (Rutstein and Rojas, 2006). By assembling characteristics on living standards correlated with a household's economic status (i.e. the ownership of a television, telephone, radio; descriptions of floor type, ceiling materials, other facilities), the DHS program calculates a wealth index for each country (Rutstein and Johnson, 2004). In the DHS, it is inferred that a household's assets and access to amenities are related to its relative economic position in the country (Rutstein, 2008).

We used the per-cluster mean wealth index calculated from the 2013 Namibia DHS, the 2011 Nepal DHS, and the 2011 Bangladesh DHS. These nationally representative surveys are based on two-stage stratified sampling of households, where enumeration areas (EAs or clusters) are first selected with probability proportional to the EA size (see Supplementary Information *Computing the DHS Wealth Index* for more information). The first stage provides a listing of households for the second stage, where a sample is selected per cluster to create statistically reliable estimates of key demographic and health variables (ICF International, 2012; National Institute of Population Research and Training et al., 2013). In Namibia, 550 clusters were first selected with probability proportional to the EA size (267 clusters in urban areas and 283 in rural areas); in Nepal, 289 clusters (95 clusters in urban areas and 194 in rural areas); and in



**Table 2 CDR variables and their description in this study for each country.**

Variable	Description	Namibia	Nepal	Bangladesh
Incoming calls	Total call count in	X	X	X
Outgoing calls	Total call count out	X	X	X
Percent nocturnal calls	Proportion of call count out during “off peak” tariff time <sup>a</sup>	X	X	X
Incoming call duration	Length of time for calls in		X	X
Outgoing call duration	Length of time for calls out		X	X
Number of places	Count of unique places visited (cell towers)	X	X	X
Entropy of places	The entropy of user’s visited places (cell towers) <sup>b</sup>	X	X	X
Radius of gyration	Measure of the typical distance a user traveled from their average location <sup>c</sup>	X	X	X
Percent interactions from home	Proportion of interactions that are routed through the tower determined to be the home tower for an individual	X	X	
Number of active days	Number of days with at least one communication	X		
Incoming texts	Total SMS count in	X		X
Outgoing texts	Total SMS count out	X		X
Frequent places	Number of locations that account for 80% of the locations where the user was observed		X	
Number of interactions	Sum of incoming and outgoing interactions		X	X
Interactions per contact	The number of interactions a user had with each of its contacts		X	X
Call fraction	Fraction of calls out of all communications	X		
Entropy of contacts	The entropy of user’s contacts <sup>b</sup>		X	X
Percent pareto	The percentage of user’s contacts that account for 80% of its interactions		X	
[...]	d			X

<sup>a</sup>Nighttime rates make it cheaper to call during “off peak” times.  
<sup>b</sup>See SI Section B for detailed information on how entropy is calculated.  
<sup>c</sup>See SI Section B for detailed information on how radius of gyration is calculated.  
<sup>d</sup>There are -130 more covariates that were only available in Bangladesh. See Steele et al. (2017), Table S1 for details.

Bangladesh, 600 clusters (207 in urban areas and 393 in rural areas). Geolocations representing the center of each sampling unit were collected in the field, enabling the use of robust statistical methods (accounting for smaller sample sizes and uncertainties in the data) to move from national estimates of poverty to subnational estimates necessitated by the SDGs.

**Calculating people in poverty.** We calculated numbers of people in poverty using WorldPop population data (Worldpop Research Group, 2017) from the most recent census year for each country (2011 for all three countries). Here the underlying assumption is each individual person takes the poverty status from his or her household as measured by the DHS wealth index. Population data were then overlaid with model outputs in ArcGIS and the sum of people in each Voronoi polygon (tower area) was calculated. The DHS divides its asset index into five quintiles and the lowest two quintiles, “poorer and poorest,” are considered to be poor (Rutstein and Johnson, 2004; Rutstein and Rojas, 2006). Following this categorization, we computed the total number of people predicted as poorer and poorest for each model (six in total) and each tower area. This allowed us to compare differences in the distribution of poverty incidence as estimated by the full and generalized models both spatially and in total numbers of people calculated to be poor.

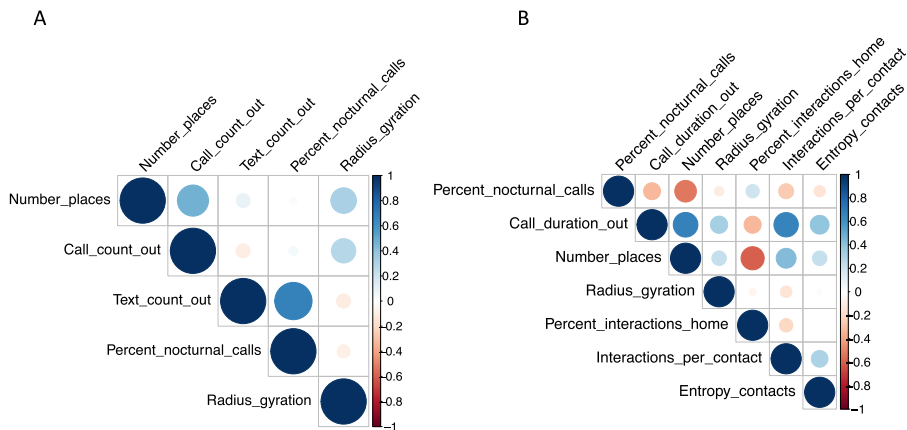
The quality of the WorldPop population data is a function of the spatial resolution of the administrative units rather than the population densities themselves. Smaller disparities between the scale of the source (administrative units) and target (100-m pixels) results in better population estimates. For the three countries modeled here, the average out-of-bag prediction error (mean squared residuals over 500 trees) for the population data was 0.47; 0.56; and 0.39 for Bangladesh, Namibia and Nepal, respectively. The average pseudo- $r^2$  for the models was 0.66; 0.95; and 0.80 for Bangladesh, Namibia and Nepal, respectively.

**Statistical analyses.** Statistical analyses were implemented using the R statistical software package (R Core Team, 2015). All

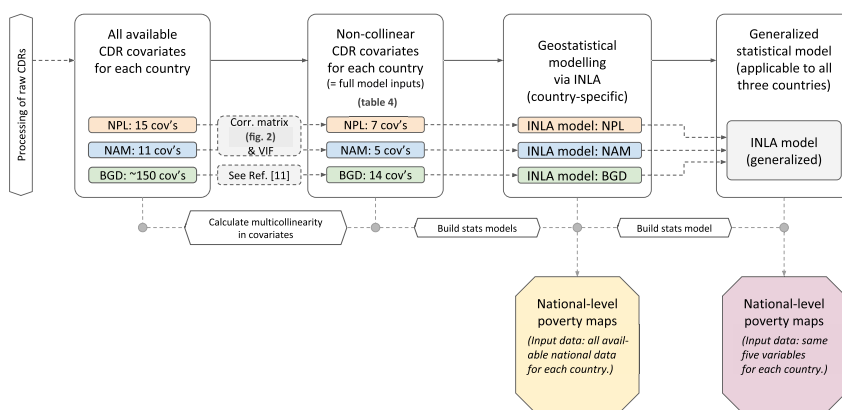
analytical steps were undertaken on a per-country basis, initially using all mobile phone data available within each country. Prior to model building and prediction, all CDR data were log transformed for normality. To assess multicollinearity in the data, we computed a bivariate Pearson correlation (see Fig. 2) to identify correlations of  $r > 0.70$ . In choosing data for the modeling process, we compared results from the Pearson correlation to computations of the variance inflation factor (VIF), whereby we removed the covariate with the highest VIF from iterations until all remaining covariates had  $VIF < 4$ . In Bangladesh only, the number of covariates available in the dataset necessitated additional processing to reduce the number of covariates from approximately 150 to 14 non-collinear variables (see Steele et al., 2017, Section 2.4 for details).

Figure 3 diagrams the data and modeling steps undertaken for this study. Models were built using a randomly selected 70% of the data to guard against overfitting. We employed hierarchical Bayesian areal models to build relationships between poverty and CDR data at sampled locations, and predict poverty estimates at unsampled locations across each country. These models were chosen due to the advantages in modeling geolocated household survey data—this modeling framework allows for straightforwardly imputing missing data, specifying prior distributions in model parameters and spatial covariance, and estimating uncertainty in predictions with a full posterior distribution for each estimate (Blangiardo et al., 2013; Blangiardo and Cameletti, 2015). All models were implemented using integrated nested Laplace approximations (INLA) (Rue et al., 2009), which uses an approximation for inference to avoid the computational demands and convergence issues, which can be problematic for MCMC algorithms (Rue and Martino, 2007).

Following previous work, the areal models are fit using R-INLA, with the Besag model for spatial effects specified inside the function (Blangiardo and Cameletti, 2015; Rue et al., 2009; Rue and Martino, 2007; Steele et al., 2017; The R-INLA project, 2016, 2015). Within the Besag model, gamma hyperpriors on the precision parameters  $\tau_\phi$  and  $\tau_\theta$  are meant to make a prior which



**Fig. 2** Bivariate Pearson correlation plots for Namibia (A) and Nepal (B). These plots show all non-collinear CDR data from the input data detailed in Table 2. The variables shown here specify the model inputs for the full models.



**Fig. 3** Data inputs and methodological steps used in this study. This diagram illustrates the processes undertaken to produce national-level poverty maps for each country from raw CDR data.

places equal emphasis on both spatial and non-spatial variance, where the precision of  $\phi$ ,  $\tau_\phi$  is given the hyperprior gamma (1, 1) and the precision of  $\theta$ ,  $\tau_\theta$  is given the hyperprior gamma (3.27, 1.81) (Elbers et al., 2002a). The model accounts for spatial covariance in the data through incorporating a spatially varying random effect, which is formed by the Voronoi polygons themselves as all of the data are aligned to mobile tower locations. The Voronois are clustered across each country at varying spatial scales and neighbors are defined within a scaled precision matrix (Sørbye and Rue, 2014) built using the geographical adjacency of the mobile phone towers to explicitly incorporate the neighborhood structure of the data This allows observations to have decreasing effects on predictions that are further away (Besag and Kooperberg, 1995). In the Besag model, Gaussian Markov random fields (GMRFs) are used to model spatial dependency structures and unobserved effects. GMRFs penalize local deviation from a constant level based on the precision parameter  $t$ , where the hyperpriors are loggamma distributed (Sørbye and Rue, 2014). The hyperprior distribution governs the smoothness of the field used to estimate spatial autocorrelation (Sørbye and Rue, 2014). The spatial random vector  $x = (x_1, \dots, x_n)$  is thus defined as

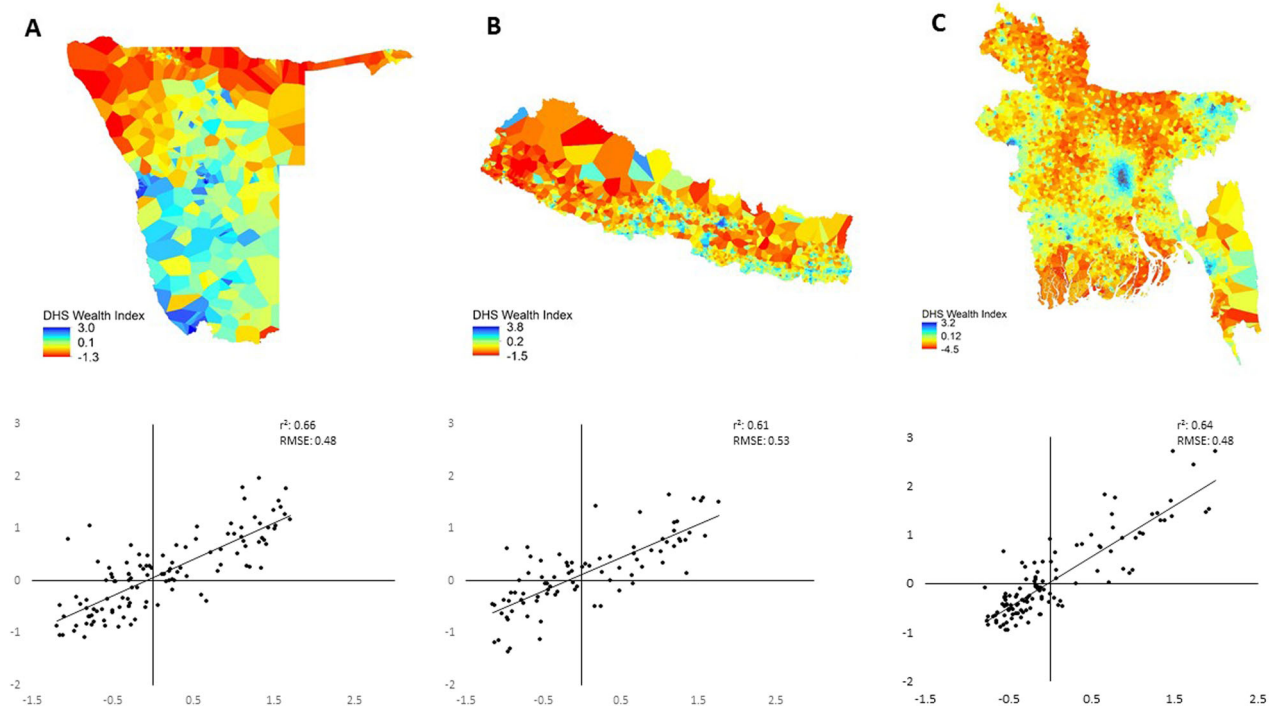
$$x_i | x_j, i \neq j, \tau \sim \mathcal{N}\left(\frac{1}{n_i} \sum_{i \sim j} x_j, \frac{1}{n_i \tau}\right),$$

where  $n_i$  is the number of neighbors of node  $i$ ,  $i \sim j$  indicates that the two nodes  $i$  and  $j$  are neighbors.

Using the fitted models, we produced estimates of the wealth index per Voronoi polygon as a posterior distribution with complete modeled uncertainty around estimates. The posterior mean and standard deviation for each polygon were then used to generate prediction maps (Fig. 4) with associated uncertainty (Fig. S4). Predictive performance of models was assessed using out-of-sample validation statistics calculated on a random 30% test subset of data; root-mean-square-error (RMSE) and the coefficient of determination ( $r^2$ ) was calculated for all models (Table 3). We also generated scatter plots of observed versus predicted values for visualization purposes (Fig. 4). The same modeling framework, with the same likelihoods, priors, and random spatial effect for each country was used for generalized models including only the common set of five CDR-derived features. We produced national estimates of poverty with associated uncertainty for each country as described above and again assessed model performance using out-of-sample validation statistics on a random 30% test set of data for comparison (Table 3).

**Results**

**Poverty mapping.** We produced national-scale poverty estimates using hierarchical Bayesian spatial models, with socioeconomic data from the Demographic and Health Surveys (DHS) and independent variables derived from CDR metadata. The spatial scale of analysis was based on approximating the mobile tower coverage areas using Voronoi tessellation (Okabe et al., 2009) and all data were aligned to these Voronoi polygons. CDR data are



**Fig. 4 National-level poverty estimates for each country.** These maps illustrate the wealth index predictions for each Voronoi polygon, with associated out-of-sample validation statistics (scatterplot below corresponds to above map, showing predicted (y-axis) vs. observed (x-axis) values) for Namibia [ $n = 141$ ] (A), Nepal [ $n = 85$ ] (B), and Bangladesh [ $n = 117$ ] (C).

**Table 3 Cross-validation statistics based on a random 30% test set of data for models using all CDR features (Full model) and a common set of CDR features (Generalized model).**

Country	Model	$r^2$	RMSE
Namibia	Full model	0.66	0.48
	Generalized model	0.65	0.48
Nepal	Full model	0.61	0.53
	Generalized model	0.60	0.54
Bangladesh	Full model	0.64	0.48
	Generalized model	0.50	0.57

aggregated at the tower level and the resultant values apply to the entire spatial extent of each Voronoi. We aligned the socio-economic data from the DHS to the Voronois by matching the lat/long of each household cluster to the polygon in which its centroid fell. We used the DHS wealth index (Rutstein, 2008; Rutstein and Johnson, 2004), an asset-based indicator of poverty calculated from nationally representative household survey data. We modeled the mean wealth index score of sampled populations within each Voronoi polygon, and where multiple household clusters fell within the same Voronoi, we modeled the mean aggregate value.

CDR-derived covariates varied for each country based on availability (see Table 2). Broadly, we utilized measures of user mobility, including the number of unique towers visited, entropy of places, and users’ radius of gyration—an indicator of movement trajectories (González et al., 2008); basic phone usage, such as the percentage of nocturnal calls made and outgoing/incoming counts of texts and calls; and social network features, including the number of interactions per contact and the entropy of users’ contacts. These social network features have been shown to correlate with economic well-being (Eagle et al., 2010). In

Bangladesh only, we were able to access and use revenue and consumption data based on users’ recharge amounts and frequencies. All CDR-derived covariate data were aligned with wealth index data in each Voronoi and fit as areal models using integrated nested Laplace approximations (INLA) (Rue et al., 2009) to estimate poverty per tower area with associated uncertainty (Fig. 4A–C and Supplementary Information, Fig. S4A–C).

All CDR data comprising the full models for each country are presented in Table 4; we used all non-collinear CDR data for each country from Table 2 in these models. We then comprised CDR data for the generalized models by examining CDR features that were statistically significant in at least one country’s full model, and were also available in all three countries. Table 5 shows these results—statistically significant covariates from the full models are listed here with the data from the generalized models highlighted in bold italics. The variables for the generalized models are the same for each country and include: number of unique towers visited, outgoing call count, percent nocturnal communications, radius of gyration, and entropy of places.

We find models utilizing only the common set of CDR features perform nearly identically to the full suite of predictors in Namibia and Nepal, and comparatively less well in Bangladesh (Table 3). The differences in predictive performance were modest: Namibia full model  $r^2 = 0.66$ , generalized model  $r^2 = 0.65$ ; Nepal full model  $r^2 = 0.61$ , generalized model  $r^2 = 0.60$ ; Bangladesh full model  $r^2 = 0.64$ , generalized model  $r^2 = 0.50$ . Only in Bangladesh was there a notable increase in model error associated with reduced data inputs: Namibia full model RMSE = 0.48, generalized model RMSE = 0.48; Nepal full model RMSE = 0.53, generalized model RMSE = 0.54; Bangladesh full model RMSE = 0.48, generalized model RMSE = 0.57.

The number of unique towers visited and percent nocturnal calls had the strongest effect on poverty predictions in the models built using the common CDR dataset (see Supplementary

**Table 4 Full model specifications for each country.**

Namibia
1 + number of places + outgoing call count + outgoing text count + percent nocturnal calls + radius of gyration
Nepal
1 + percent nocturnal calls + outgoing call duration + number of places + radius of gyration + percent interactions from home + interactions per contact + entropy of contacts
Bangladesh
1 + recharge average per tower + percent nocturnal calls + number of places + entropy of contacts + outgoing internet sessions + sum outgoing internet sessions + incoming voice duration + count incoming content management system + count sum incoming content management system + volume of incoming multimedia messages + recharge amount per transaction + count incoming multimedia messages + count incoming texts + weekly recharge amount

**Table 5 Mobile phone data used in country-specific and generalized poverty models.**

	Namibia	Nepal	Bangladesh
<b>Number of places</b>	<b>X</b>	<b>X</b>	<b>X</b>
<b>Outgoing call count</b>	<b>X</b>		
Outgoing text count	X		
<b>Percent nocturnal communications</b>	<b>X</b>	<b>X</b>	<b>X</b>
<b>Radius of gyration</b>	<b>X</b>	<b>X</b>	
Incoming text count			X
<b>Entropy places</b>			<b>X</b>
Entropy contacts		X	X
Outgoing call duration		X	
Top-up measure			X
Multimedia messaging			X
Internet usage			X

Statistically significant CDR-derived features for country-level models are denoted with an X. From these, covariates that were statistically significant in at least one country-level model and available for all three countries make up the CDR data for generalized models, and are italicized in bold.

Information, Tables S1–S3). In addition, outgoing call counts were important in Namibia, whereas radius of gyration and entropy of places were prominent in Nepal. Given the full suite of available predictors, the number of unique towers visited and percent nocturnal calls remained significant. Results also show a few, key covariates unique to each country as these models included data that were not available for all three countries at the time of this study. In Namibia, this includes outgoing text counts and the number of users whose home location is at each tower as important covariates. In Nepal, entropy of contacts and the percentage of interactions from users’ home tower were significant. In Bangladesh, incoming text counts were important covariates, as well as measures of top-ups, multimedia messaging, and Internet usage.

**Spatial distribution of poverty.** The DHS divides its asset index into five quintiles and the lowest two quintiles are considered poor (Rutstein and Johnson, 2004; Rutstein and Rojas, 2006). To explore the spatial distributions of poverty, we calculated the total number of people in the lowest two quintiles for each model. The mean wealth index score modeled in Namibia and Nepal is bimodal in distribution, with a higher proportion of households falling into lower quintiles (see Supplementary Information, Fig. S1A, B). In Bangladesh, the mean wealth index score modeled is positively skewed, with far greater numbers of households in lower quintiles (Fig. S1C). We would expect a more accurate model to reflect these input data and predict greater numbers of people in poverty nationally in Bangladesh, and better differentiation of poverty and wealth in Namibia and Nepal (more

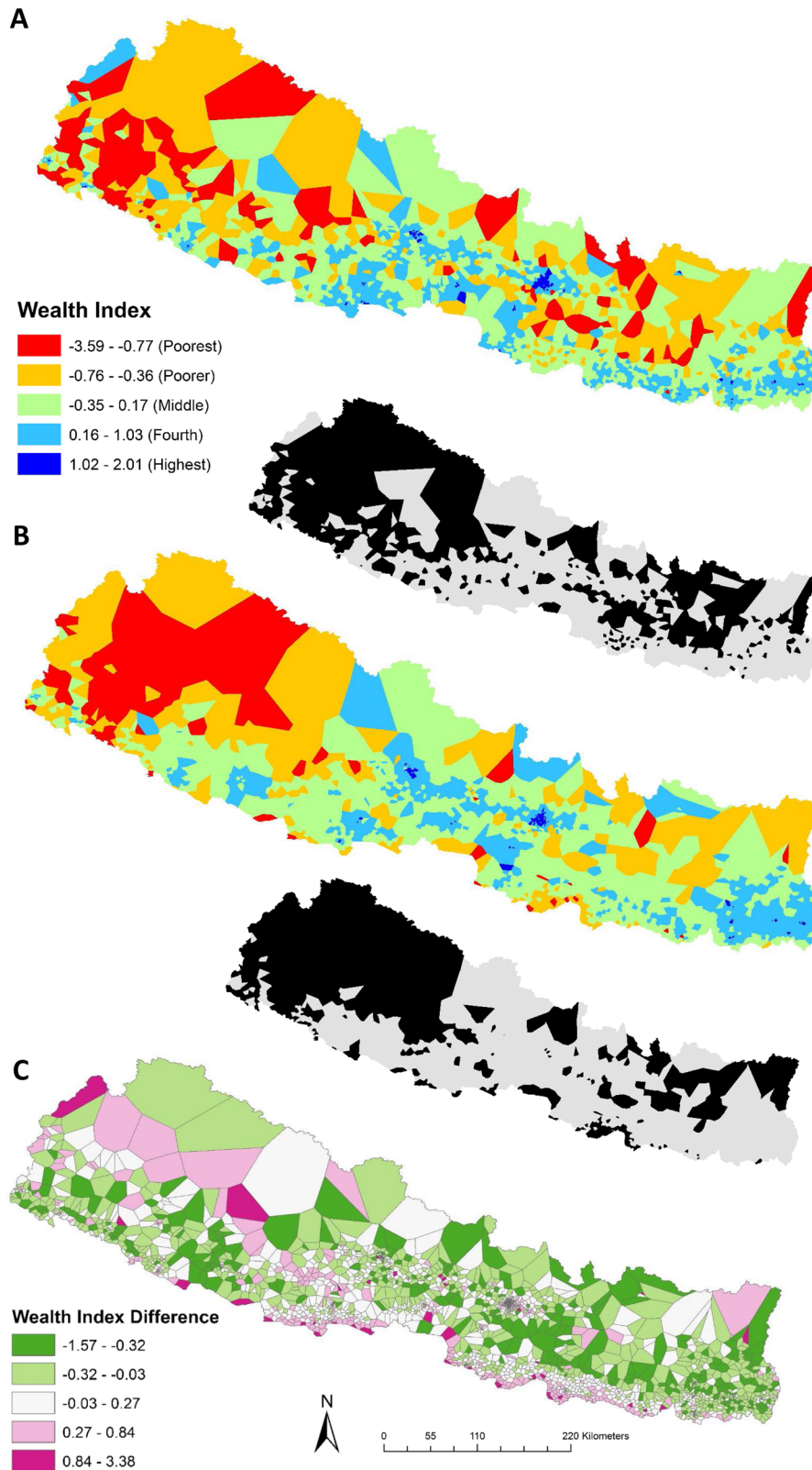
poverty, more wealth, and fewer middle-class geographies). Models using all CDR-derived features produced marginally better outputs in terms of prediction and error, while also including additional data specific to each country; thus, we expect greater numbers of people in poverty to be predicted by the full models. Models using only the common subset of CDR data could leave out poor people and fail to capture differences across geographies due to incompleteness of data, higher model error, or lower predictive power.

In Namibia, the full model predicts 909,432 people in poverty versus 857,761 predicted by the generalized model. There are small shifts in the spatial distribution of poverty where areas are predicted to be poorer or richer, but in general the patterns in the urban centers and north/south regional trends hold (see Supplementary Information, Fig. S2). In Bangladesh, the differences are striking, both in respect to the total numbers of people in poverty (full and generalized models predict 17,107,057 and 9,832,711 poor people, respectively) and in their spatial distribution. The additional CDR data used in the full model (text counts, top-up data, multimedia messaging data, and Internet usage) produce a map with greater precision and distinction between poverty and wealth (Fig. S3) as compared to the generalized model, which it predicted most areas in the middle class.

Nepal demonstrated a different outcome, where the generalized model predicted greater numbers of people in poverty (generalized: 6,707,748 and full: 6,436,490). Likewise, the spatial distribution of the predictions shifted appreciably between the models (Fig. 5). To explore this, we looked at the covariates that were having the greatest effect on model outputs along with existing benchmarks of population density and socioeconomic status. Entropy of places, radius of gyration, and number of unique towers visited have the greatest effect on outputs from the generalized model (see Table S2), which are all measures of user mobility. Also, higher wealth is predicted in the national parks, where an increase in mobility from tourism could be a contributing factor (Nepal Ministry of Culture, Tourism & Civil Aviation, 2014). In terms of absolute change, more and greater levels of poverty are predicted across the southern regions of Nepal in the generalized model.

By incorporating recent population (Worldpop Research Group, 2017) and poverty data for Nepal (Bank, 2013; Haslett et al., 2014; Nepal and Bohara, 2010), it became evident that although poverty incidence in rural Nepal—predominately in the north and northwest—is higher than in urban Nepal, the numbers of absolute poor are higher across the southern regions—especially in the south-southeast—due to higher population densities. In the case of our two models, the generalized model more accurately reflects this. The generalized model, as driven by mobility data, predicts greater levels of poverty in regions of high population density that concurrently have lower mobility. This more accurately reflects the higher total numbers





**Fig. 5 National wealth index maps produced for Nepal.** This figure shows maps produced using all available noncollinear CDR data (A), a CDR subset comprised of 5 generalizable features (B), and the difference between these two models (C). For (A) and (B), the subset maps show, in black, poor areas predicted by each model (the Demographic and Health Surveys class the two lowest quintiles, poorer and poorest, as poor).

of people in poverty and demonstrates poorer people to have lower mobility than wealthier people, matching findings elsewhere (Wesolowski et al., 2013).

## Discussion

The results here demonstrate that five easily replicable, population-level CDR-derived features are able to account for 50–65% of the variance in socioeconomic status nationally across Namibia, Nepal, and Bangladesh, highlighting how a smaller set of data are able to contribute to monitoring and mapping poverty metrics across countries. This work represents the first attempt to generalize CDR-derived features across countries to predict poverty. We are able to identify aggregate information reflecting user's mobility and call behavior as having a key role in explaining the distribution of poverty in very different contexts. The results provide evidence-based support for including aggregated, anonymized CDRs wherever possible as a non-trivial data component for strategic poverty measurement and monitoring, and demonstrate that CDRs do give reasonable estimates of the distributions of socioeconomic status across LMICs.

Although our aim is not to determine causation, or the determinants of poverty, we thought that data related to user mobility and call patterns would correlate well with socioeconomic status and considered the following explanations:

1. We expect higher levels of mobility (as measured by the radius of gyration and entropy of places) lead to a higher level of socioeconomic status (Wesolowski et al., 2013), with the idea being that wealthier people are more mobile and visit more places than poorer people.
2. We expect a higher percentage of nocturnal calls correlates to a lower level of socioeconomic status, with the idea being that nighttime rates are cheaper so poorer people will do more of their communications during these “off peak” times.
3. We expect a higher count of outgoing calls leads to a higher socioeconomic status, with the idea being that the initiating party pays for outgoing calls. Whereas receiving a call does not result in a charge.

Mobility features were most important in explaining the variation in poverty across Nepal, whereas in Namibia call pattern data were more significant. Both types of data were needed in Bangladesh to achieve 50% explained variance in the generalized model. Text counts (incoming and outgoing) were unobtainable for Nepal at the time of this study, but were important features in mapping poverty in Namibia and Bangladesh. This is not surprising as texting is customary in these countries and people may text each other more than call. In Nepal, event durations were more important than event counts—length of incoming/outgoing calls was a better predictor of socioeconomic status in this context, suggesting that measures of event duration capture important information on consumption and expenditures. Additional data—especially top-ups, Internet usage, and SMS communications are expected to improve poverty maps wherever these data are available. Further exploration is needed in terms of the relationship between distance-based CDR features and tower locations to quantify the extent to which some of the mobility and entropy covariates are providing unique information versus being a function of tower distribution. The final sets of covariates here were determined largely by availability, MNO agreements, and relevance in previous work.

Poverty maps produced with CDR-derived features need to be interpreted within a particular locality. Poverty is highly context-specific and factors associated with poverty can vary considerably from country to country. Understanding key country-specific

data together with information on how people use their phones is essential. In Bangladesh, direct measures of consumption—top-ups, data on Internet usage and multimedia messaging features—were necessary to capture the variability in poverty and detect the poorest households present in the household survey data. As demonstrated in the Nepal models, the spatial distribution and estimates of numbers of absolute poor can shift significantly based on different types of input data. Incorporating additional information on geographical conditions and phone usage yields reciprocal benefits. For example, population densities and demographic data provide insight into the mobility patterns of low-income people, and this information highlights how well model outputs are estimating fine-scale variation in poverty. This could reduce the inadvertent exclusion of people who are poor from estimates, and ultimately programs designed to reduce poverty.

In the absence of a census, the method applied here is able to estimate poverty reasonably well as measured by the DHS wealth index. We chose the wealth index for this study as it is widely available. The DHS produces estimates approximately every 5 years for many LMICs (The DHS Program Country List, 2018), depending on survey type, instrument, and sample size (The Demographic and Health Survey (DHS), 2018). As such, it could be feasible using these data to construct high-resolution estimates of asset-based poverty 2–3 times before 2030. With additional household survey data—using similar methodologies and variables to construct a wealth index—more points in time could be produced. Therefore, it must be noted that the method applied here captures long-term poverty trends (i.e. 5-year changes in assets and living standards) rather than short-term developments (i.e. 6–12 months changes in consumption or expenditure). Ideally we would also test income- and consumption-based metrics of poverty to better understand how well the applied method and data could capture these short-term changes in socioeconomic status but those data were not available at the time of this study. To that end, it would be incredibly useful to incorporate other types of survey data to test how well CDRs or other types of ‘big data’ for that matter can estimate short-term changes in poverty and wealth during intercensal periods and integrate these estimates temporally. Evaluating the extent to which features derived from CDRs can capture these short-term fluctuations would be requisite for a proper evaluation of their usefulness as compared to traditional surveys.

We inevitably had a mismatch in years of CDR and survey data for Nepal and Bangladesh. Where both datasets were concurrent in Namibia, we achieved the best results—highest predictive power and lowest error, with no appreciable difference between the full and generalized models, highlighting the importance of matching data sources temporally. As demonstrated in previous work (Chen and Nordhaus, 2011; Elbers et al., 2002a; Head et al., 2017; Jean et al., 2016; Njuguna and McSharry, 2017; Noor et al., 2008; Steele et al., 2017; Watmough et al., 2016), data from satellites and user-generated GIS platforms are important data sources expected to improve predictions, especially in rural areas where mobile towers can be sparse. Fewer model features provide computational tractability of analysis and interpretability for policy makers or non-specialists. Nevertheless, as computing power and algorithm development progress, we will be able to extract these types of measures faster and with increased accuracy. This will make understanding how people use technology, and their geographical conditions, even more important when making inferences from big data to derive solutions to improve lives.

Progressing the global development agendas requires identification of the poor, and CDRs can contribute to these efforts by providing timely, accurate updates on socioeconomic status in

populations for monitoring and evaluation. These data also offer the potential of dynamic measurement and the ability to evaluate change over time. Although significant challenges in accessing these data and distributing outputs remain, we are optimistic that studies such as this demonstrating the usefulness of aggregate, anonymous CDR data will encourage mobile operators to continue to collaborate with researchers, development agencies, and governments working toward development goals. Part of this process necessarily includes getting data on the political agenda to connect the supply and demand of data, and create enabling environments for data to flow across systems and users. This could yield increased efficiency and foster the incorporation of real-time data into how the SDGs are being addressed.

### Data availability

Data are available for the replication of results only by contacting the corresponding author.

Received: 11 September 2018; Accepted: 11 October 2021;

Published online: 22 November 2021

### References

- Bank AD (2013) Nepal: country partnership strategy (2013–2017). Asian Development Bank.
- Besag J, Kooperberg C (1995) On conditional and intrinsic autoregressions. *Biometrika* 82:733–746. <https://doi.org/10.1093/biomet/82.4.733>
- Blangiardo M, Cameletti M (2015) Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.
- Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spat Spatio-Temporal Epidemiol* 4:33–49. <https://doi.org/10.1016/j.sste.2012.12.001>
- Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. *EPJ Data Sci* 4. <https://doi.org/10.1140/epjds/s13688-015-0046-0>
- Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–1076. <https://doi.org/10.1126/science.aac4420>
- Blumenstock J, Gillick D, Eagle N (2010) Who's calling? Demographics of mobile phone use in Rwanda. *Transportation* 32:2–5
- Chen X, Nordhaus WD (2011) Using luminosity data as a proxy for economic statistics. *Proc Natl Acad Sci USA* 108:8589–8594. <https://doi.org/10.1073/pnas.1017031108>
- Data for Development (2017) <http://www.d4d.orange.com/en/Accueil>. Accessed 23 May 2017.
- Data–Pop Alliance (2018) Data–pop alliance. <http://datapopalliance.org/>. Accessed 17 May 2018.
- Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328:1029–1031. <https://doi.org/10.1126/science.1186605>
- Eagle N, Pentland A. (Sandy), Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106:15274–15278. <https://doi.org/10.1073/pnas.0900282106>
- Elbers C, Lanjouw JO, Lanjouw P (2002a) Micro-level estimation of poverty and inequality. *Econometrica* 71:355–364
- Elbers C, Lanjouw JO, Lanjouw P (2002b) Micro-level estimation of welfare (Research working paper no. 2911). World Bank Development Research Group, Washington, USA
- Frias-Martinez, V. and Virseda, J. (2012) On the Relationship between Socio-Economic Factors and Cell Phone Usage. In Best, M. L. et al. (eds), *Fifth International Conference on Information and Communication Technologies and Development, ICTD '12, Atlanta, GA, USA, March 12-15, 2012*, ACM, New York, pp. 76–84. <https://doi.org/10.1145/2160673.2160684>
- Global Partnership for SDGs (2018) Global Partnership for Sustainable Development Data. <http://www.data4sdgs.org/>. Accessed 23 May 2017.
- GMSA (2016) Discussing ‘Data for Development’.’ *Mob Dev*. <http://www.gsma.com/mobilefordevelopment/programme/digital-identity/discussing-data-for-development>. Accessed 23 May 2017.
- González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. *Nature* 453:779–782. <https://doi.org/10.1038/nature06958>
- GSMA (2014) *Mob Dev*. <http://www.gsma.com/mobilefordevelopment/country/global/gmsa-guidelines-on-the-protection-of-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-ebola-outbreak>. Accessed 23 May 2017.
- Haslett S, Jones G, Isidro M, Sefton A (2014) Small area estimation of food insecurity and undernutrition in Nepal. Central Bureau of Statistics, National Planning Commissions Secretariat, World Food Programme, UNICEF and World Bank, Kathmandu, Nepal
- Head, A. et al. (2017) Can Human Development be Measured with Satellite Imagery? In Saif, U. (ed). *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, ICTD 2017, Lahore, Pakistan, November 16 - 19, 2017*, ACM, New York, pp. 1–11. <https://doi.org/10.1145/3136560.3136576>
- Hentschel J, Lanjouw JO, Lanjouw P, Poggi J (1998) Combining census and survey data to study spatial dimensions of poverty (Policy Research Working Paper No. 1928). The World Bank Development Research Group and Poverty Reduction and Economic Management Network Poverty Division.
- ICF International (2012) Demographic and health survey sampling household listing manual, MEASURE DHS. ICF International, Calverton, USA
- Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. *Science* 353:790–794. <https://doi.org/10.1126/science.aaf7894>
- Jerven M (2013) Poor numbers: how we are misled by african development statistics and what to do about it, Cornell studies in political economy. Cornell University Press, Ithaca
- Målvqvist M (2015) Abolishing inequity, a necessity for poverty reduction and the realisation of child mortality targets. *Arch Dis Child* 100:S5–S9. <https://doi.org/10.1136/archdischild-2013-305722>
- National Institute of Population Research and Training, Mitra and Associates, ICF International (2013) Bangladesh Demographic and Health Survey 2011. Dhaka, Bangladesh and Calverton, USA.
- Nepal M, Bohara A (2010) Micro-level estimation and decomposition of poverty and inequality in Nepal. *Nepal Himal Res Pap Arch*. [https://digitalrepository.unm.edu/nsc\\_research/37](https://digitalrepository.unm.edu/nsc_research/37)
- Nepal Ministry of Culture, Tourism & Civil Aviation (2014) Nepal Tourism Statistics 2013. Ministry of Culture, Tourism & Civil Aviation, Singha Durbar, Kathmandu
- Njuguna C, McSharry P (2017) Constructing spatiotemporal poverty indices from big data. *J Bus Res* 70:318–327. <https://doi.org/10.1016/j.jbusres.2016.08.005>
- Noor AM, Alegana VA, Gething PW, Tatem AJ, Snow RW (2008) Using remotely sensed night-time light as a proxy for poverty in Africa. *Popul Health Metr* 6:5. <https://doi.org/10.1186/1478-7954-6-5>
- Okabe A, Boots B, Sugihara K, Chiu SN (2009) Spatial tessellations: concepts and applications of Voronoi diagrams. John Wiley & Sons.
- OPAL (2017) Glob. Partnersh. Sustain. Dev. Data. <http://gpsdd.squarespace.com/dc-opal>. Accessed 23 May 2017.
- Pokhriyal N, Jacques DC (2017) Combining disparate data sources for improved poverty prediction and mapping. *Proc Natl Acad Sci USA* 114:E9783–E9792. <https://doi.org/10.1073/pnas.1700319114>
- R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian Inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Ser B* 71:319–392
- Rue Hå, Martino S (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J Stat Plan Inference Special Issue: Bayesian Inference Stoch Processes* 137:3177–3192. <https://doi.org/10.1016/j.jspi.2006.07.016>
- Rutstein S (2008) The DHS Wealth Index: approaches for rural and urban areas (DHS Working Papers No. 60). United States Agency for International Development, Demographic and Health Research Division, Macro International Inc., Calverton, USA
- Rutstein S, Johnson K (2004) The DHS Wealth Index (DHS Comparative Reports No. 6). ORC Macro, Demographic and Health Surveys, Calverton, Maryland
- Rutstein S, Rojas G (2006) Guide to DHS statistics. ORC Macro, Demographic and Health Surveys, Calverton
- Smith-Clarke, C., Mashhadi, A. and Capra, L. (2014) Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks. In Jones, M. et al. (eds). *CHI '14: CHI Conference on Human Factors in Computing Systems, Toronto, Ontario, Canada, April 26 - May 1, 2014*, ACM, New York, pp. 511–520. <https://doi.org/10.1145/2556288.2557358>
- Sorbye SH, Rue H (2014) Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spat Stat Spatial Statistics Miami* 8:39–51. <https://doi.org/10.1016/j.spasta.2013.06.004>
- Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E (2011) Prediction of socioeconomic levels using cell phone records. In: Konstan JA, Conejo R, Marzo JL, Oliver N (Eds.) *User modeling, adaptation and personalization*. Lecture notes in computer science. Springer, Berlin, Heidelberg, pp. 377–388
- Steele JE, Sundsøy PR, Pezzulo C, Alegana VA, Bird TJ, Blumenstock J, Bjelland J, Engø-Monsen K, Montjoye Y-A, de Iqbal AM, Hadiuzzaman KN, Lu X, Wetter E, Tatem AJ, Bengtsson L (2017) Mapping poverty using mobile



- phone and satellite data. *J. R. Soc. Interface* 14:20160690. <https://doi.org/10.1098/rsif.2016.0690>
- Stork C (2011) Are mobile phones replacing the use of public phones in Africa? *info* 13:75–90. <https://doi.org/10.1108/14636691111131466>
- The Demographic and Health Survey (DHS) (2018) <https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm>. Accessed 17 May 2018.
- The DHS Program Country List (2018) <https://dhsprogram.com/Where-We-Work/Country-List.cfm>. Accessed 17 May 2018.
- The R-INLA project (2016) <http://www.r-inla.org/models/latent-models>. Accessed 21 Jan 2016.
- The R-INLA project (2015) Besag model for spatial effects [WWW Document].
- Tusting LS, Willey B, Lucas H, Thompson J, Kafy HT, Smith R, Lindsay SW (2013) Socioeconomic development as an intervention against malaria: a systematic review and meta-analysis. *The Lancet* 382:963–972. [https://doi.org/10.1016/S0140-6736\(13\)60851-X](https://doi.org/10.1016/S0140-6736(13)60851-X)
- United Nations E. and S.C. (2017) Progress towards the Sustainable Development Goals. United Nations E. and S.C.
- United Nations General Assembly (2015) Transforming our world: the 2030 Agenda for Sustainable Development (Resolution adopted by the General Assembly No. A/RES/70/1).
- United Nations Global Pulse (2018) <https://www.unglobalpulse.org/>. Accessed 17 May 2018.
- Watmough GR, Atkinson PM, Saikia A, Hutton CW (2016) Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: an example from Assam, India. *World Dev* 78:188–203. <https://doi.org/10.1016/j.worlddev.2015.10.031>
- Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. *J R Soc Interface* 10:20120986. <https://doi.org/10.1098/rsif.2012.0986>
- Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2012) Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS ONE* 7:e35319. <https://doi.org/10.1371/journal.pone.0035319>
- Wesolowski A, zu Erbach-Schoenberg E, Tatem AJ, Lourenço C, Viboud C, Charu V, Eagle N, Engø-Monsen K, Qureshi T, Buckee CO, Metcalf CJE (2017) Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nat Commun*. 8:2069. <https://doi.org/10.1038/s41467-017-02064-4>
- World Bank (2016) World development report 2016: digital dividends. World Bank. <https://doi.org/10.1596/978-1-4648-0671-1>
- Worldpop Research Group (2017) <http://www.worldpop.org.uk/>. Accessed 13 Sept 2017.
- zu Erbach-Schoenberg E, Alegana VA, Sorichetta A, Linard C, Lourenço C, Ruktanonchai NW, Graupe B, Bird TJ, Pezzulo C, Wesolowski A, Tatem AJ (2016) Dynamic denominators: the impact of seasonally varying population numbers on disease incidence estimates. *Popul Health Metr* 14:35. <https://doi.org/10.1186/s12963-016-0106-0>

## Acknowledgements

We thank NCell and Axiata Group, in particular Suren J. Amarasekera, Vishal Upadhyay, Prabin Pandey, Surendra Maharjan, Bhaba Bajracharya, and Eswari Prasad Sharma. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. JES is supported by funding from the Bill & Melinda Gates Foundation (OPP1106936 and OPP1134076). CP is supported by the Bill & Melinda Gates Foundation (OPP1106427). EE is supported by a Wellcome Trust Sustaining Health Grant (106866/Z/15/Z). AJT is supported by funding from the Bill & Melinda Gates Foundation (OPP1182408, OPP1106427, 1032350, OPP1134076), the Clinton Health Access Initiative, a Wellcome Trust Sustaining Health Grant (106866/Z/

15/Z), and funds from DFID and the Wellcome Trust (204613/Z/16/Z). This work forms part of the WorldPop Project ([www.worldpop.org](http://www.worldpop.org)) and Flowminder Foundation ([www.flowminder.org](http://www.flowminder.org)).

## Author contributions

JES was responsible for designing the research, data processing and management, statistical analyses, writing, interpretation and production of the final manuscript. CP was responsible for survey data processing and management, and interpretation and drafting of the final manuscript. MA contributed to the research design, CDR processing and interpretation, and writing and production of the final manuscript. EE, CJB, MA, PRS, KE, BG, RLN, and PS were responsible for CDR data access, management and/or processing of CDR data, and interpretation of the final manuscript. SO contributed data processing and management, GIS analyses, and map production. KN contributed to drafting and editing of the final manuscript. AJT was responsible for overall scientific management and interpretation/editing of the final manuscript. All authors gave final approval for publication.

## Competing interests

The authors declare no competing interests.

## Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

## Informed consent

This article does not contain any studies with human participants performed by any of the authors.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-021-00953-0>.

**Correspondence** and requests for materials should be addressed to Jessica E. Steele.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021