Using mobile call detail records (CDRs) and remote sensing data for spatial mapping of poverty[1]

Chanuka Algama, Merl Chandana (LIRNEasia)

Draft as of 6 September 2023

Abstract

Ending all forms of poverty remains a primary global challenge and a key objective of the Sustainable Development Goals (SDGs). To effectively combat poverty, access to accurate information about the locations of affected populations is vital. This data enhances our comprehension of poverty's root causes, facilitates optimized resource allocation for poverty alleviation initiatives, and is a fundamental element in monitoring poverty rates over time. While censuses and surveys are the established benchmarks for measurement and the foundation of targeted social assistance programs like cash transfers, they are often resource-intensive and time-consuming. This makes them impractical for rapid deployment during emergencies and continuous monitoring of spatial and temporal poverty disparities, which is essential for comprehensive development strategies. To bridge this gap, this study examines the potential of utilizing alternate data sources, both publicly available data and those from private data providers, to gain fresh insights into poverty's spatial distribution. By employing unsupervised and supervised machine learning techniques, we explore the feasibility of utilizing mobile call detail records (CDRs) as well as geographic information system (GIS) and remote sensing (RS) data to map poverty spatially. Despite the limitation of lacking comprehensive, precise ground truth data for full validation, our initial findings suggest the approach holds promise in achieving higher spatial and temporal resolutions for poverty mapping.

1. Background

Poverty casts a long shadow over societies, leading to undesirable consequences like child mortality, limited access to education, societal instability & conflict, all chipping away at the quality of life (Cruz, et al. 2015). Countries worldwide prioritize the alleviation of poverty in its diverse forms, be it the deeply ingrained structural kind or the temporary setbacks that afflict certain regions, households, and individuals. Ending poverty in all its forms remains a major challenge and it remains the first target of the Sustainable Development Goals (SDGs) (Transforming our world: the 2030 Agenda for Sustainable Development 2015). To alleviate poverty, it is crucial that information is available on where affected people live. These data enhance our understanding of the causes of poverty and play a crucial role in better allocating resources for poverty alleviation programs and serve as a critical component for monitoring poverty rates over time (Transforming our world: the 2030 Agenda for Sustainable Development 2015).

In economic development, poverty data serves a dual purpose. Firstly, it's crucial for assessing poverty at an individual level, which is vital for targeted assistance programs like cash transfers (Handa, et al. 2012). This involves analyzing factors such as income, consumption, and relevant indicators to effectively identify those in urgent need of support. Secondly, poverty data helps grasp the spatial distribution of poverty, enabling governments and organizations to strategically identify regions with

[1] This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada. The views expressed herein do not necessarily represent those of IDRC or its Board of Governors. The authors wish to acknowledge the contributions of: (1) Viren Dias, whose work was instrumental in developing the ground truth poverty maps; (2) Rohan Samarajiva, who provided feedback on an earlier draft of the paper.

high poverty rates. This knowledge informs the implementation of comprehensive development initiatives, including infrastructure projects and improvements in education and healthcare, all contributing to substantial and lasting poverty reduction (Akinyemi 2007).

The high-resolution indicators of poverty used in targeted social assistance programs such as cash transfers, often rely on well-designed large-scale surveys. However, given the resource-intensive nature of such surveys, it is difficult to conduct them regularly or on demand. As such, even when, approximate estimates are needed for providing social assistance during times of crisis (e.g., COVID-19 lockdowns), surveys become a prohibitive tool. Small area estimation (SAE), on the other hand, is a technique deployed by statistical offices around the world to obtain sub-national spatial estimates of poverty. SAE methods usually leverage statistical techniques to estimate poverty related parameters for sub-populations using census data (Rao and Molina 2015) (Albacea 2020). However, census data are typically collected every 10 years and often released with a delay of one or more years, making the updating of poverty estimates challenging in most instances. However more recent research points to novel sources of high-resolution data that can help estimate the level of socio-economic well-being for smaller geographical regions (AIken and Ohlenburg 2023).

In particular, recent work illustrates the potential of features derived from satellite remote sensing data (hereafter called RS data) and mobile operator call detail records (hereafter called CDRs) for spatial mapping poverty and socio-economic well-being. RS data captures physical attributes such as rainfall, temperature, vegetation, infrastructure, settlement patterns, and proximity to roads and markets (Engstrom, Hersh and Newhouse 2016) (Martillan and Martinez Jr 2021). On the other hand, CDRs provide insights into household access to financial resources through monthly credit consumption on mobile phones and the prevalence of mobile phone usage in an area, which can be indicative of remittance flows and economic opportunities (Aiken, et al. 2022). These data sources offer distinct and complementary information for understanding various aspects of socioeconomic conditions (Soto, et al. 2011).

RS and CDR data also complement each other due to their different spatial scales. CDR data, aggregated at the level of cell towers, provides spatial resolution determined by tower coverage, which varies between rural and urban areas. Within CDR data, Voronoi cells represent polygonal regions around towers, encompassing locations closest to each tower and they do not correspond to administrative boundaries. In contrast, RS data offers coarser resolution in urban areas, focusing on land properties, but provides continuous coverage across regions and can be aggregated at desired geographic levels.

In this study, we employ a combination of overlapping RS data and CDRs to assess their effectiveness in accurately estimating spatial poverty in Sri Lanka. Here, poverty is defined based on a socio-economic index that we derived from the 2012 Census data. To estimate spatial distribution of poverty, we utilize principal component analysis (PCA) and spatial regression techniques – presenting two distinct approaches for integrating the two data sources. The discussion section compares our results, outlines the limitations of our data and methods, and suggests future research directions to build upon this study.

2. Data Collection, Materials, and Methods

2.1. Spatial Scale and Data Processing

The two main types of data used in this study were of two different spatial scales. In order to make sure that the spatial resolutions and extents matched, the data had to be further processed. The final spatial scale was of the analysis was chosen to be Grama Niladhari (GN) divisions of Sri Lanka, which is the most granular administrative division in the country. All downloaded RS data was processed and summarized to spatially align with the polygon boundaries of GN division boundaries. The CDRs metrics were first calculated at Voronoi cell levels and then mapped on to the GN boundaries. Depending on the feature being considered (see 2.3 below), each polygon was assigned RS and CDR values representing the mean, sum or mode of the corresponding data. The GN divisions with zero population were removed for two main reasons. First, these divisions did not contribute any relevant data for the analysis of poverty levels, as there were no households to consider. Second, including these divisions could introduce unnecessary complexity and error into the model. Therefore, 37 such divisions were excluded for the sake of accuracy and relevance in the poverty mapping analysis. The socioeconomic index which is used to train and validate the models was based on the census and was calculated at the GN level (see 2.2 below).

2.2. Poverty Data

Given that Sri Lanka doesn't have open poverty data at GN division level for the year 2013 (the year was determined by the time period of the CDR data that was used in the study) we had to develop our own socio-economic index using principal component analysis (PCA) techniques (Dias, et al. 2020). In recent times, PCA based spatial poverty analysis using census data has emerged as a popular and reliable estimate of measuring socio-economic well-being (Krishnan 2010).We selected, as our input, the 2012 national census, which is available as a summary of counts at the Grama Niladhari Division (GND) level. We started with 109 variables corresponding to household and demographic characteristics of households and employed a variable elimination process to reduce the variable number to 61. Then we normalized and standardized each variable and ran PCA on the dataset. We multiplied the weights of the resulting first principal component with the standardized dataset and summed each row to produce a score for each GN Division. This score was to serve as the socioeconomic index. This method hinges on the assumption that that the first principal component resulting from the application of PCA on a dataset of socioeconomic indicators is the socioeconomic index.

2.3. CDR & RS Data

CDR features were generated from mobile CDR data collected from two of the leading mobile network operators in Sri Lanka for the year 2013. CDR features range from metrics such as basic phone usage, social network to metrics of user mobility and handset usage. These features are easily calculable metrics that do not rely on complex algorithms. They include various parameters of the corresponding distributions such as weekly or monthly median, mean and variance. The full list of CDR features generated is given in the table below.

Table 1: Features derived from CDR data.

Feature category	Individual feature description
Phone usage	Call count, average call duration, nighttime call count, incoming call count, avg nighttime call duration, avg incoming call duration, avg outgoing call duration

Location/mobility	Radius of gyration, home location
Social Network	Spatial entropy, avg call count per contact
Handset type	Smart/feature/basic phone

We further identified, assembled, and processed 25 raster and vector datasets into a set of for the whole of Sri Lanka at GN division level. These data were obtained from existing sources (maps produced by other researchers and agencies) and produced ad hoc for this study to include environmental and physical metrics likely to be associated with human welfare such as vegetation indices, night-time lights, climatic conditions, and distance to roads or major urban areas. A full summary of assembled covariates is provided in the table below. Where the data did not perfectly overlap with the time period being considered (the calendar year of 2013), RS data sources that were either partially overlapping with or closest to the time period being considered were selected.

Table 2: Features derived from GIS & Remote Sensing Data

Feature Category	**Feature Description**
Accessibility	Accessibility to populated places with more than 50k people
Population	Population count
Population	Population density
Climate	Mean aridity index
Climate	Average annual potential evapotranspiration [mm]
Night-time lights	VIIRS satellite night-time lights intensity
Elevation	Elevation in meters
Vegetation	Vegetation Index
Distance	Distance to roads
Distance	Distance to waterways
Urban/rural	MODIS satellite -based global urban extent
Protected area	Protected areas
Land cover	European Space Agency land cover map values
Demographic	Pregnancies
Demographic	Births
Ethnicity	Georeferenced ethic groups
Climate	Mean annual precipitation
Climate	Mean annual temperature

2.4. Feature Selection

Prior to statistical analyses, all CDR and RS features were log transformed for normality. Bivariate Pearson's correlations were computed for each pair of covariates to assess multicollinearity, and for high correlations ($r > 0.70$), we eliminated covariates that were less generalizable across countries/regions.

2.5. Modeling

Given the poverty data used for the study was derived by running PCA run on census data, we ran principal component analysis using features using CDR and RS data to allow for better comparable

results. We ran three types of models. We ran one model each using CDR and RS data separately and a third model combining CDR-RS data. We selected the first principal component from each of the models to be the measure representing the socioeconomic status of each of the GN divisions.

In addition to PCA, we also used spatial regression techniques to predict the socio-economic index using the CDR and RS data. First spatial weights were assigned to the data using the nearest neighbor method. This method was employed as it defines the neighbor set of an observation as its nearest 'k' observations, effectively capturing spatial relationships based on proximity. This is particularly pertinent in spatial poverty mapping where geographic closeness often implies similar poverty conditions.

Upon assigning weights, we applied various spatial models to the data. The initial model applied was the Ordinary Least Squares (OLS) model. However, the presence of clustering in the errors suggested that the OLS model may not be the most suitable for our analysis.

In response to this, we ran a Spatial Error Model (SEM), a Spatial Lag Model (SLX), and a Spatial Durbin Error Model (SDEM). These models account for spatial dependencies in the data that the OLS model does not consider. Subsequently, we compared these models using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) which are.

3. Results

PCA Models

Since PCA is an unsupervised machine learning technique, it is not appropriate to evaluate the derived models using traditional metrics like R-squared value or RMSE (root mean squared error). Instead, we focus on comparing the alignment between the final outputs of the models and the ground truth poverty index. To do this, we examine the overlap between each model and the poverty dataset by considering the number of common poorest GN divisions. This intersection analysis allows us to assess the similarity between the models and the poverty data.

Table 3: Ability to correctly identify top N poorest Grama Niladhari Divisions

Top N% poorest GN Divisions	% of GNs accurately identified - model output compared against ground truth (PCA) poverty data		
	CDR-only	CDR-RS	RS-only
Top 25%	6.1%	3.3%	45%
Top 40%	17%	13%	61%
Top 50%	30%	26%	71%

An examination of the results obtained from the unsupervised technique of PCA indicate that, RS-only model does a reasonable job at correctly identifying the poorest Grama Niladhari divisions while both the CDR-only model and the combined CDR-RS model perform very poorly at correctly identifying the poorest GN divisions.

Regression Models

To assess the performance of our regression models, we used two common statistical methods for comparing spatial models: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) (Kuha 2004). AIC helps in model selection by weighing how well a model predicts data against its complexity. Lower AIC scores indicate simpler models that remain effective at prediction, making them

preferable for data analysis. BIC, on the other hand, also balances prediction quality and model complexity but has a stronger inclination toward simpler models. Lower BIC values suggest simpler models that are often better suited for data analysis. In general, when all other factors are equal, opting for a model with a lower AIC or BIC value tends to yield better results by striking a balance between accurate prediction and model simplicity.

We have also used two additional metrics typically used for evaluating regression models - r^2 value, which measures model fit or the proportion of the variance in the dependent variable that is explained by the independent variables and the MSE (Mean Squared Error) value which quantifies the average difference between predicted and observed values. Upon considering AIC & BIC criteria in concert with the r^2 value and the MSE value, we concluded that the SLX (spatial lag) model was the optimal choice for our analysis. As evidenced in the tables below the SLX outperformed the SDEM model and the SLX model across all four-evaluation metrics across different configurations of the data sources.

Table 3: Evaluation metrics for the combined RS + CDR regression models (evaluated using 20% of ground truth data)

Model type (RS+CDR)	**r^2 value**	**MSE**	**AIC**	**BIC**
SDEM Model (Spatial Durbin Error Model) – CDR-RS models	0.7460	2.76	53980.93	60541.07
SEM Model (Spatial Error Model)	0.6263	4.41	60541.07	60880.63
SLX Model (Spatial Lag Model)	0.8155	2.0	49525.64	49865.20

Table 4: Evaluation metrics for the RS only regression models (evaluated using 20% of ground truth data)

Model type (RS only)	**r^2 value**	**MSE**	**AIC**	**BIC**
SDEM Model (Spatial Durbin Error Model) – CDR-RS models	0.7022	3.241	56198.99	56440.45
SEM Model (Spatial Error Model)	0.4240	16.126	78636.82	78878.29
SLX Model (Spatial Lag Model)	0.8092	2.077	49974.79	50216.26

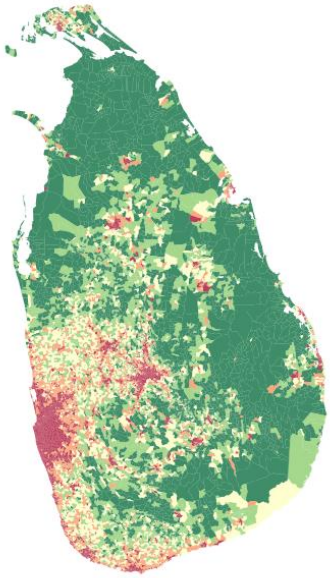
Table 5: Evaluation metrics for the CDR only regression models (evaluated using 20% of ground truth data)

Model type (CDR only)	**r^2 value**	**MSE**	**AIC**	**BIC**
SDEM Model (Spatial Durbin Error Model) – CDR-RS models	0.4570	6.7653	66454.45	66567.64
SEM Model (Spatial Error Model)	0.4579	6.7288	66378.75	66491.93
SLX Model (Spatial Lag Model)	0.7844	2.370	51786.79	51899.98

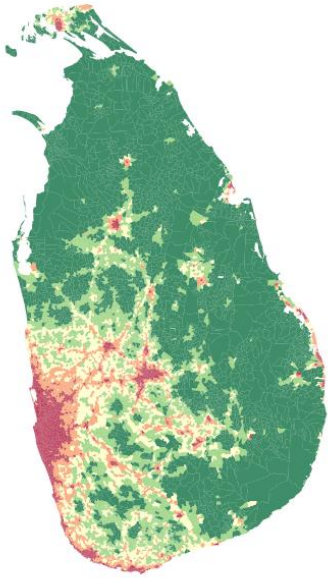
Visual comparison of model outputs

SDEM Model

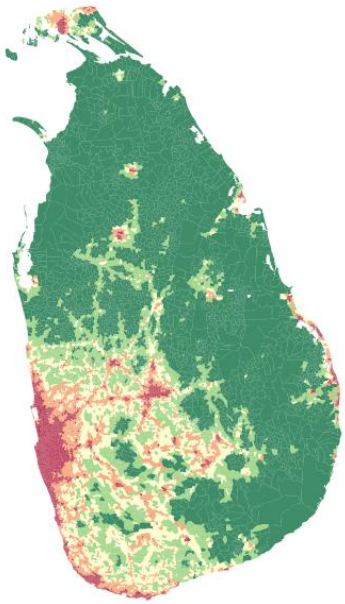
Ground Truth Map



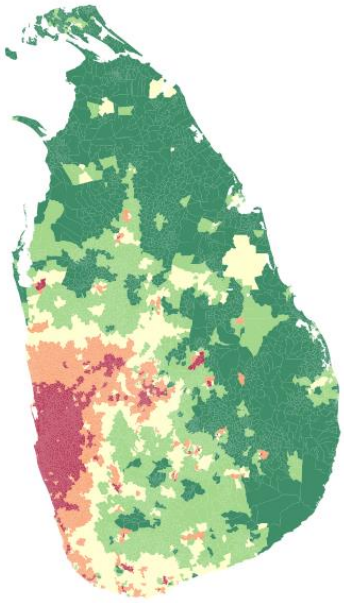
RS + CDR Map



RS Only Map

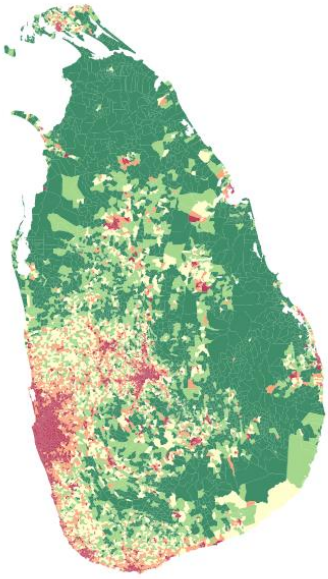


CDR Only Map

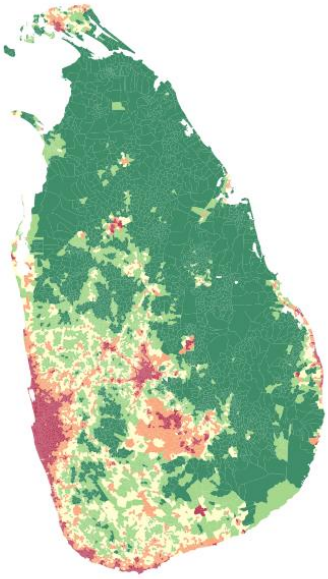


SEM Model

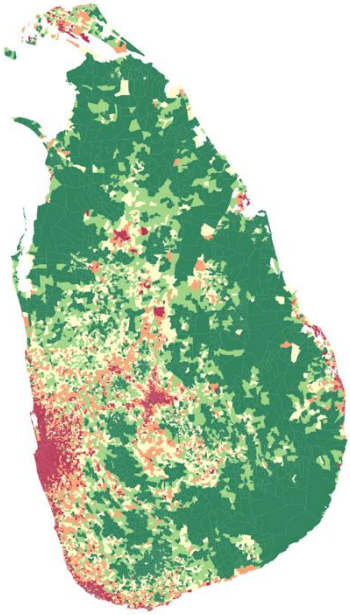
Ground Truth Map



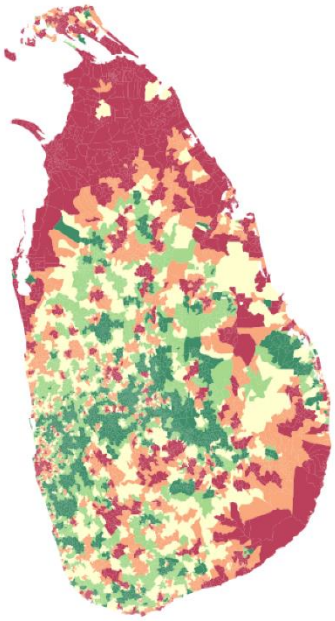
RS+CDR Map



RS Only Map

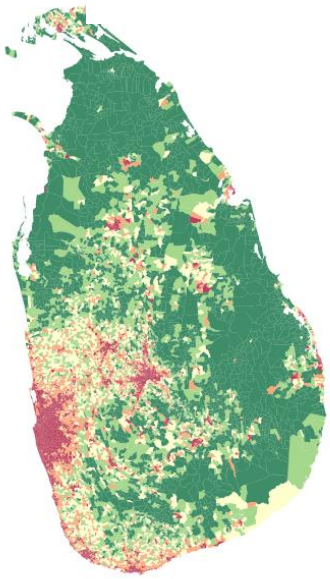


CDR Only Map

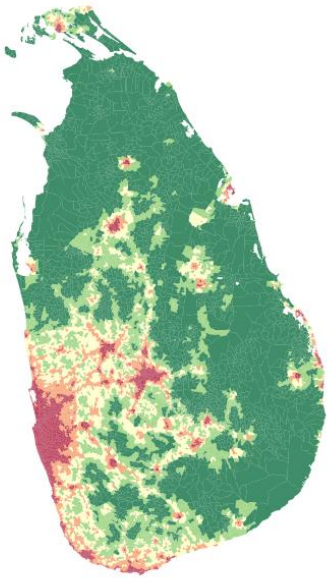


SLX Model

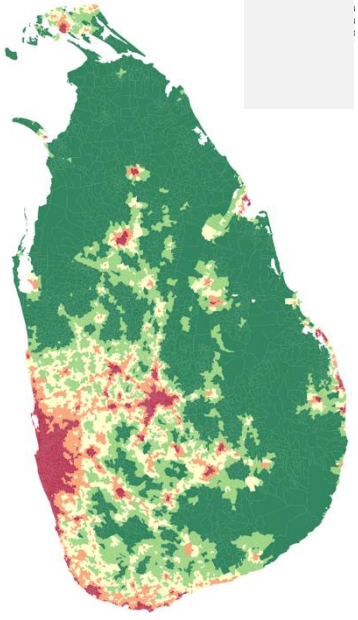
Ground Truth Map



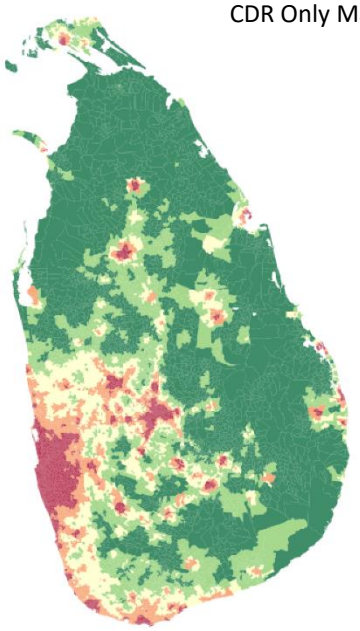
RS+CDR Map



RS Only Map



CDR Only Map



The maps display a color gradient from red to green, with red representing high socio-economic status and green representing low socio-economic status. The four maps presented for each type of model offer a visual sense of accuracy of our attempt in mapping poverty using remote sensing data and call detail records. The maps combining RS and CDR consistently appear to show a more faithful representation of the ground truth, with clusters and connections between those clusters clearly visible. This suggests that the combination of these two methods can potentially enhance the accuracy of the map. The RS-only maps are also reasonably accurate, as evidenced both by the evaluation metrics and the visual representation. This demonstrates the effectiveness of remote sensing in capturing spatial details. Lastly, the CDR-only map is less accurate. While CDRs provide valuable insights into socioeconomic behavior, including consumption and mobility patterns, CDR features used in this study may not be as reliable for directly predicting spatial poverty levels in absence of other complementary features.

4. Discussion

This work represents an attempt to replicate work done by Steele et al., (2017) to build predictive maps of poverty using a combination of CDR and RS data in the Sri Lankan setting. While it shows promising early results, further work needs to be done to further evaluate its efficacy in the Sri Lankan study. This study is also constrained by limitations of poverty data the challenges and limitations around the use of CDR data. The following points about the study are particularly worth highlighting.

Spatial lag models (SLX) outperforming SDEM models & SEM models in mapping poverty

Spatial lag models (SLX) could be outperforming Spatial Durbin Error Models (SDEM) and Spatial Error Models (SEM) in mapping poverty for several reasons. Poverty often exhibits spatial autocorrelation, with neighboring areas sharing similar poverty rates, making SLX models, which explicitly consider this spatial dependency, more effective. SDEM introduces complexity with both a spatial lag and a spatial error term, while SEM primarily focuses on modeling spatial errors, potentially missing direct spatial dependencies. In contrast, spatial lag models excel in capturing the direct influence of neighboring areas, providing a more accurate representation of how poverty spreads across regions, making them a preferred choice for poverty mapping.

The need for better poverty data

There are many approaches to measuring poverty and common measures include asset, consumption, and income-based measures of well-being. However, given the absence of granular publicly available poverty data at GN division level, we utilized our own socio-economic index derived using the 2012 census data. This index was composed using principal component analysis technique that leveraged household and demographic characteristics of households. This relied the assumption that the first principal component resulting from the application of PCA on a dataset of socioeconomic indicators is the socioeconomic index. While this represents a reasonable approximation of poverty in data poor, resource-constrained settings, it is only considered to be a moderately accurate approximation.

Extension of this work could include generating ground truth using small area estimation (SAE) techniques on census data available at GN division level. While SAE involves considerable modeling it's use of more advanced statistical approaches such as area level and hierarchical models, is believed to be capable of generating better spatial approximations of poverty compared to PCA methods.

Using more advanced modeling techniques

The choice of modeling techniques for this work was largely guided by the type of poverty data that was available. Given that the poverty data (the socio-economic index calculated using the census data) was derived using principal component analysis (PCA) techniques, the first set of models also leveraged PCA techniques to enable straight forward comparisons of output. We generated 3 models (mobile only, CDR only, & mobile & CDR) using PCA techniques and regression techniques. However, the availability of better poverty data (such as those generated through SEA techniques) could enable more advanced modeling techniques. For example, Steele et al (2019), used hierarchical Bayesian geostatistical models (BGMs) to predict geographical distribution of poverty in Bangladesh. BGMs offer several advantages for addressing the limitations and constraints associated with modelling geolocated survey data. These include straightforwardly imputing missing data, allowing for the specification of prior distributions in model parameters and spatial covariance, and estimating uncertainty in the predictions as a distribution around each estimate.

Limitations of CDR data & alternatives

This study leveraged mobile network CDR data obtained from multiple operators for the period in consideration. However, the data was only obtained for a specific period and did not contain all the features leveraged by Steele et al., (2017) in their original study. Further, continued access to CDR data requires technical procedures to ensure the privacy of individuals included in the dataset and legal expertise to ensure that the use of data does not violate data protection laws other laws that govern the use of personal data in countries. These factors make the use of CDR data for poverty mapping prohibitive; especially when it comes to replication and extension of this work by other researchers in Sri Lanka as well as in other countries.

More recent work has shown that remote sensing indicators alone can be used to effectively map spatial distribution of poverty (Martillan and Martinez Jr 2021). As such it might be advisable to start with a wider array of remote sensing data, such as optical satellite data, night-light data & radar data and calculator features that might be indicative of poverty. Then, through a combination of desk research & feature selection techniques, the optimal combination of features could be used in spatial poverty estimation models.

Making poverty maps usable

An effective poverty map should exhibit certain essential characteristics. Firstly, it should encompass the temporal dynamics of poverty, recognizing its dynamic nature influenced by economic fluctuations and seasonal variations. Additionally, the poverty map should offer high-resolution representations at local or sub-national levels, facilitating the identification of poverty hotspots, prioritizing interventions, and enabling effective monitoring of progress. Lastly, it should be cost-effective, scalable, and readily accessible, supporting evidence-based decision-making for policymakers and development practitioners. Therefore, further work in this line should prioritize using widely available ground truth (poverty) data, features derived from regularly updated, openly accessible datasets and flexible modeling approaches that meet the different needs of policy makers faced having varying priorities.

References

Alken, Emily, and Tim Ohlenburg. 2023. *Novel Digital Data Sources for Social Protection: Opportunities and Challenges.* Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH .

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine learning and phone data can improve targeting of humanitarian aid." *Nature* 603: 864-870.

Akinyemi, Felicia O. 2007. "Spatial data needs for poverty management." *Research and theory in advancing spatial data infrastructure concepts* 33-54.

Albacea, Zita. 2020. *Introduction to Small Area Estimation Techniques.* Manila: Asian Development Bank.

Assembly, United Nations General. 2015. "Transforming our world: the 2030 Agenda for Sustainable Development."

Atkinson, Tony. 2017. *Monitoring Global Poverty. Report of the Commission on Global Poverty.* Washington, DC: World Bank Group. Accessed 07 17, 2023. https://blogs.worldbank.org/developmenttalk/why-world-bank-adding-new-ways-measure-poverty.

Cruz, Mario, James Foster, Bryce Quillin, and Schellekens. 2015. *Ending Extreme Poverty and Sharing Prosperity: Progress and Policies.* The World Bank.

Dias, Viren, Lasantha Fernando, Tharaka Amarasinghe, and Yudhanjaya Wijeratne. 2020. *Mapping Poverty and Wealth: an Alternative Socioeconomic Index for Sri Lanka.* Colombo, Sri Lanka: LIRNEasia.

Engstrom, Ryan', Jonathan Hersh, and David Newhouse. 2016. *Poverty in HD: What Does High Resolution Satellite Imagery Reveal about Economic Welfare.* `The World Bank Group.

Handa, Sudhanshu, Carolyn Huang, Nicola Hypher, Clarissa Teixeira, Fabio V. Soares, and Benjamin & Davis. 2012. "Targeting effectiveness of social cash transfer programmes in three African countries." *Journal of Development Effectiveness* 4 (1): 78-108 .

Jessica E. Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M. Iqbal, Khandakar N. Hadiuzzaman, Xin Lu, Erik Wetter, Andrew J. Tatem and Li. 2017. "Mapping poverty using mobile phone and satellite data." *Journal of the Royal Society Interface* 14 (127).

Krishnan, Vijaya. 2010. "Constructing an Area-based Socioeconomic Index:A Principal Components Analysis Approach." University of Alberta, Edmonton.

Kuha, Jouni. 2004. "AIC and BIC: Comparisons of assumptions and performance." *Sociological methods & research 33.*

Martillan, Marymell, and Arturo Martinez Jr. 2021. *Mapping the Spatial Distribution of Poverty Using Satellite Imagery in Thailand.* Manila: Asian Development Bank.

Rao, N.K.., and Isabel Molina. 2015. *Small Area Estimation.* Wiley.

Soto, Victor, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. 2011. "Prediction of Socioeconomic Levels Using Cell Phone Records." *User Modeling, Adaptation and Personalization* . Girona.