

# Informing Public Policy with Machine Learning: Mapping Poverty in Sri Lanka Using Mobile Call Detail Records and Remote Sensing Data

Chanuka Algama<sup>1</sup>, Merl Chandana<sup>1</sup>, Viren Dias<sup>1, 2</sup>, Kasun Amarasinghe<sup>1, 3</sup>

<sup>1</sup>LIRNEasia, <sup>2</sup>Calcey, <sup>3</sup>Carnegie Mellon University  
chanuka, merl@lirneasia.net  
viren@calcey.com  
kamarasi@andrew.cmu.edu

## Abstract

Pinpointing where poverty is most severe and tracking its changes over time is crucial for facilitating and monitoring the effects of poverty alleviation initiatives. However, traditional benchmarks like household surveys and national censuses often fall short—they’re expensive, resource-intensive, infrequent, and incapable of reflecting the full spectrum of household well-being. They often fail to account for geographic variation in cost of living, medical needs, or the costs of earning income. This gap is particularly pronounced in contexts like Sri Lanka, leading to a reliance on obsolete information when responding to economic shocks or disasters. On top of that, poverty cannot be determined by income data alone; rather, it’s multidimensional, where factors such as infrastructure, access to services, and economic activity also play a role in determining the well-being of a community.

This work makes two key contributions: First, we adapted existing machine learning (ML) approaches for the first time to Sri Lankan contexts, demonstrating the applicability while providing fresh insights into the spatial distribution of poverty. Second, we address significant gaps in the literature by proposing frameworks for validating model outputs on their ability to inform poverty alleviation initiatives or humanitarian aid.

We train ML models by using satellite imagery, remote sensing, and mobile operator aggregated data. We evaluate the models’ ability to identify Grama Niladhari divisions (GN, the smallest administrative divisions in Sri Lanka) with the highest poverty. An impressive showcase of the model performance emerges against the established poverty metrics using the Household Income and Expenditure Surveys (HIES), and Census data conducted at the second smallest administrative unit level (DS division), the models were capable of correctly identifying the poorest 22 DSDs out of 25.

## 1 Introduction

Poverty casts a long shadow over societies, leading to child mortality, limited access to education, societal instability, and conflict—each of which erodes the quality of life (Cruz et al. 2015). Today poverty rates in low-income countries are higher than before the COVID-19 pandemic resulting in more than 8.5% of the global population falling into the

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

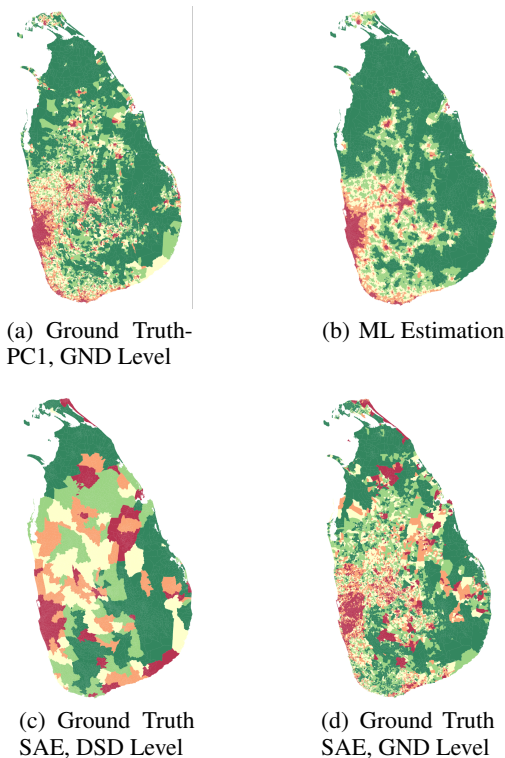


Figure 1: Spatial distribution of poverty, predicted (b) vs ground truth (a), (c) and (d). Red indicates areas of high socioeconomic status while green highlights areas of low socioeconomic status. (a) Ground truth poverty index created using Census data through the first Principle component (PC1), while (c), and (d) are Ground truth of poverty index estimated through small area estimation (SAE).

hands of extreme poverty, despite the efforts of the Sustainable Development Goals (SDGs) as ending poverty in all its forms remains their first target (Lee et al. 2016) (Lawrence and Shipman 2024).

To alleviate poverty effectively, reliable information is essential on where the affected populations live. Such data reveal spatial patterns of poverty, play a crucial role in allocating resources and guiding development initiatives such as in-

frastructure, education, and healthcare improvements, leading to significant and sustained poverty reduction (Akinyemi 2007). Moreover, the spatial distribution of wealth and poverty data can be used to monitor the effectiveness of such interventions and programs.

Decennial censuses, while not primarily designed for poverty measurement, offer most countries' most detailed data on socioeconomic conditions. Developed nations often complement census data with routinely collected information on public benefits, unemployment insurance, and income, which can serve as poverty proxies. Developing countries also conduct Household Income and Expenditure Surveys, but these surveys lack the depth and granularity needed to inform poverty alleviation measures.

These large-scale surveys provide valuable data when conducting targeted social assistance programs like cash transfers. Still, they are resource-intensive and infrequent, limiting their utility for ongoing poverty monitoring and development initiatives in low and middle-income countries like Sri Lanka. Moreover, they often fail to account for geographic variation in cost of living, medical needs, or the costs of earning income. They are based on outdated assumptions about consumption patterns, resulting in an incomplete picture of a community's well-being. Additionally, these measures are based on outdated consumption patterns, relying on thresholds that no longer reflect how households allocate their spending. As a result, policy evaluations—especially in welfare decentralization can be flawed or misleading (Shrider et al. 2021).

Recent advances in ML and the increasing availability of high-resolution data sources have induced a growing body of research aimed at enhancing traditional poverty estimation methods. Satellite imagery, remote sensing (RS) data, and mobile phone metadata—particularly call detail records (CDRs)—have emerged as promising alternatives to unravel complex patterns of socioeconomic conditions at granular spatial and temporal resolution. RS data captures physical attributes such as rainfall, temperature, vegetation, infrastructure, nighttime lights, settlement patterns, and accessibility, providing insights into environmental and geographic factors influencing poverty (Engstrom, Hersh, and Newhouse 2016; Mitterling et al. 2021). Pioneering studies such as (Jean et al. 2016) and (Steele et al. 2017) have demonstrated that convolutional filters extracted from deep learning models trained on satellite images can effectively predict asset-based wealth indices across sub-Saharan Africa and Bangladesh.

(Blumenstock, Cadamuro, and On 2015) demonstrated that mobile phone usage patterns, such as the call frequency, duration, geographic mobility patterns, top-up behavior, potential economic activity, and social network characteristics, can serve as good proxies for socioeconomic status. They further demonstrated that the diversity and size of a user's or community's social network, and the frequency of mobile recharge amounts were strongly predictive of household or granular-administrative area wealth levels. These behavioral indicators capture difficult-to-observe community well-being, such as disposable income, access to resources, and social capital.

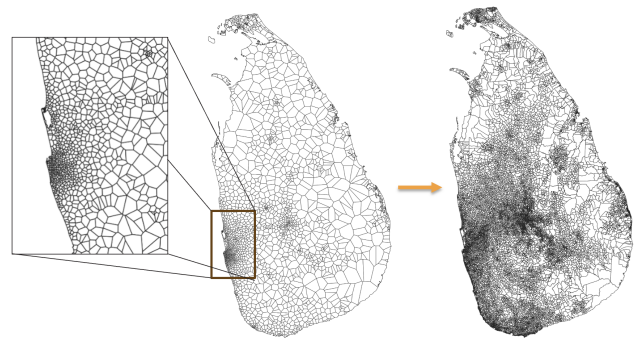


Figure 2: Spatial scaling of CDR data from Voronoi cell-level calculations to GN boundaries, where each GN polygon is assigned CDR values based on the mean, sum, or mode of the data, depending on the feature under analysis.

The more recent studies, such as (Chi et al. 2022), (Espín-Noboa, Kertész, and Karsai 2023), and (Agyemang et al. 2023) have gone further, integrating multiple data modalities including satellite imagery, mobile phone networks, topographic maps, crowd-sourced metadata, de-identified connectivity data from Facebook and administrative data—into unified predictive frameworks. These studies prove that ML models can significantly improve spatial coverage and update frequency for poverty mapping.

However, model validation in almost all of this literature has relied on statistical metrics such as  $R^2$ , RMSE, or cross-validated error, assuming each prediction is equally important. These metrics give an overall fit and performance in predicting the exact index up to the last decimal, but don't show how well the models support real-world decision-making processes.

Our work seeks to address this gap by adopting a decision-centered evaluation framework. Rather than treating model performance as an end in itself, we start by asking how the outputs will be used in practice, such as identifying the most affected communities following an economic shock or natural disaster to inform resource-constrained humanitarian aid. Through this backward design approach, we highlight the need for ranking the administrative units rather than considering the statistical robustness. For instance, in a scenario where limited resources must be allocated quickly, we assess whether the model successfully prioritizes the most severely impacted communities.

On the other hand, one of the main drawbacks in training ML models with spatially adjacent administrative units is the risk of data leakage from testing to training data, resulting in overly optimistic results. This is particularly pronounced in estimating poverty due to the nature of neighboring communities having similar socioeconomic conditions. Our framework overcomes this by incorporating stratified cross-validation techniques while considering the divisional secretariat divisional level splits to select GN divisions to train the ML models. Through these approaches, our framework brings a more meaningful measure of utility to translate predictive models into better-informed policies

and more effective interventions.

## 2 Datasets and Preparation

We use data from five sources: (1) Census data to derive a socio-economic index (SEI) to establish a "ground truth" for poverty distribution, (2) RS data, (3) CDR, and (4) satellite imagery as readily available data sources for building poverty maps in the absence of census data, and (5) to compare and validate the results we used poverty headcount index (HCI) generated by the World Bank in collaboration with the Department of Census and Statistics (DCS) in Sri Lanka.

Recently, PCA-based spatial poverty analysis using census data has emerged as a popular and reliable measure of socioeconomic well-being (Senna, Maia, and Medeiros 2019). We used the 2012 national census, the most recent census data available for Sri Lanka, to produce a score to serve as the socioeconomic index (SEI) for each GND. Starting with 109 variables corresponding to the demographic characteristics of the households, we employed variable elimination and the standard principal component analysis techniques. We denote the first principal component of the dataset by PC1.

CDR was collected from two of the leading mobile network operators in Sri Lanka for 2013. It is aggregated at the level of cell towers and provides spatial resolution determined by tower coverage, which varies between rural and urban areas. Within CDR data, Voronoi cells represent polygonal regions around towers, which encompass locations closest to each tower and do not correspond to administrative boundaries.

In contrast, RS data, which we collected from existing sources, satellites, and open maps, offers coarser resolution in urban areas, focusing on land properties, but it provides continuous coverage across regions and can be aggregated at desired geographic levels. Satellite images were collected from the Landsat 8 Earth observation satellite for the year 2013, and 8 different ranges of frequencies (bands) were selected along the electromagnetic spectrum. Despite census data being available at the GND level, both RS and CDR data required spatial harmonization due to their mismatched resolutions (see Figure 2).

To validate our results, we used the poverty headcount index (HCI), which serves as a benchmark generated using small area estimation methods, a widely accepted method among researchers around the globe to estimate poverty at disaggregated administrative levels. This index, developed by the World Bank in collaboration with Sri Lanka's Department of Census and Statistics, combines the 2012 Census of Population and Housing (CPH) and the 2012/13 Household Income and Expenditure Survey (HIES). It provided poverty estimates at the Divisional Secretariat (DSD) level (Department of Census and Statistics, Sri Lanka and World Bank 2015).

Metric Type	Feature
Phone usage	Call count
	Average call duration
	Geometry
	Nighttime call count
	Incoming call count
	Avg nighttime call duration
	Avg incoming call duration
Location/Mobility	Avg outgoing call duration
	Radius of gyration
	Home location
Social Network	Spatial entropy
	Avg call count per contact
Handset type	Smart/feature/basic phone

Table 1: Features derived from CDR data

## 3 Model Building and Validation

### 3.1 Feature Generation from CDR, RS data and satellite images

Using CDR data, we created features that capture basic phone usage, social networks, user mobility, and handset usage. They include various parameters of the corresponding distributions, e.g., weekly or monthly median, mean, and variance. Table 1 gives the complete list of the CDR features generated.

To capture environmental and physical attributes that are likely to be associated with human welfare, such as vegetation indices, nighttime lights, climatic conditions, and accessibility information like distance to roads or major urban areas, A full list of RS features calculated and used is given in Table 2.

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

Table 2: Features derived from GIS & Remote Sensing Data for the year 2013

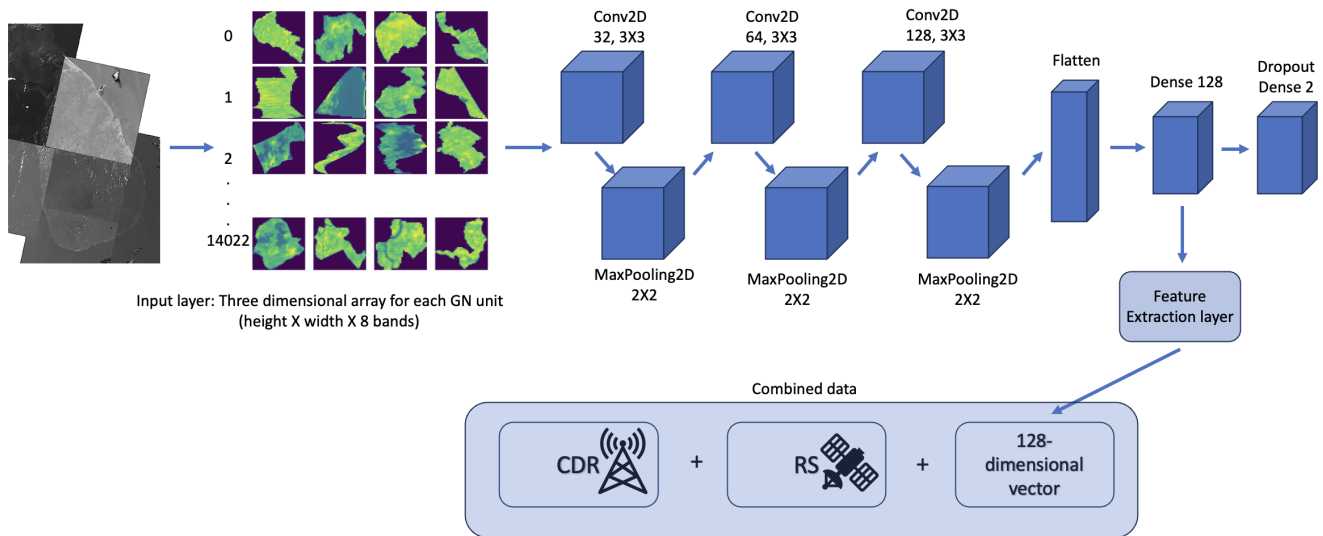


Figure 3: Process of complementing RS, and CDR with a 128-dimensional vector learned from a CNN architecture described.

In addition to these tabular features, we incorporated spatial patterns extracted from raw satellite images into our modeling pipeline. We selected eight bands across the electromagnetic spectrum—including visible, near-infrared, and shortwave infrared wavelengths from the Landsat 8 Earth observation satellite, as they provide rich information about land use, vegetation health, water bodies, and built-up structures.

Each image was clipped to the administrative boundaries of the GN divisions, resulting in a three-dimensional array for each GN unit: width  $\times$  height  $\times$  8 bands. These image cubes were then fed into a Convolutional Neural Network (CNN), designed to learn abstract spatial features that might be predictive of poverty. The CNN architecture we implemented consisted of three convolutional layers with increasing filter depths (32, 64, and 128), each followed by max pooling operations to reduce dimensionality and capture hierarchical spatial patterns. We extracted a 128-dimensional vector from the penultimate dense layer. These learned feature vectors were used as input alongside CDR and RS to the downstream models (See Figure 3).

All CDR and RS features were log-transformed for normality. Then, the Bivariate Pearson’s correlations were computed for each pair of features to assess multicollinearity, and for high correlations ( $r > 0.70$ ), we eliminated covariates that were less generalizable across countries/regions. Spatial weights were calculated and assigned to each GND considering its neighbors. However, Sri Lanka has 62 islands that do not share boundaries with any other region, making it impossible to assign weights to them using a traditional contiguity-based spatial weight matrix. To address this, we used a K-Nearest Neighbors (KNN) approach, such that each observation had a fixed number of spatial neighbors, to effectively capture spatial relationships based on proximity. This is particularly pertinent in spatial poverty mapping, where geographic closeness often implies similar

poverty conditions.

### 3.2 Model Development

Even though the primary objective of this study is to develop a predictive model that can estimate both continuous and discretized versions of the Socio-Economic Index (SEI) at the GND and DSD levels, our focus is on ranking administrative units from the poorest to the least poor rather than predicting the SEI values precisely. In addition to prediction, we also use inference models to examine which covariates are associated with poverty the most. For example, we investigate how well variables such as nighttime lights or population density correlate with SEI.

**Predicting continuous SEI:** We experimented with frequentist, Bayesian, and ensemble approaches such as Random Forest and XGBoost. We selected three frequentist spatial regression models: the Spatial Error Model (SEM), the Spatial Lag Model (SLX), and the Spatial Durbin Error Model (SDEM). The Spatial Error Model was chosen to account for spatial autocorrelation in the error terms (Anselin 2009) (Gao, Asami, and Chung 2006) and to test whether the spatial dependency is based heavily on the unobserved factors or the omitted variables ( $y = X\beta + \epsilon$ ).

The Spatial Lag Model captures the influence of neighboring observations on the dependent variable. It is ideal when the outcome of interest, such as poverty levels, is directly influenced by the outcomes in nearby areas (F. Dormann et al. 2007). By including a spatially lagged dependent variable ( $\rho Wy$ ), the model tries to incorporate situations where the value of the dependent variable in one location depends on values in adjacent locations ( $y = \rho Wy + X\beta + \epsilon$ ).

To capture more complex spatial dependencies, we utilized the Spatial Durbin Error Model to integrate the features of both the SEM and SLX models, to account for spatial spillover effects. SDEM is particularly valuable when

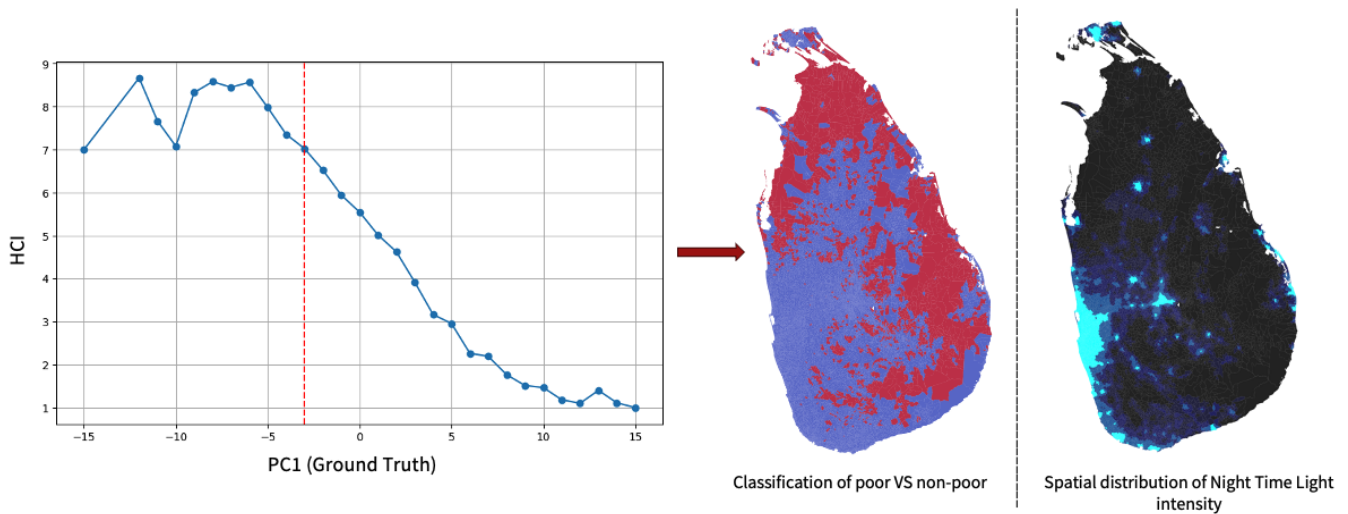


Figure 4: Poverty Headcount Index vs. PC1: A Threshold Analysis for Poverty Classification. The threshold of  $PC1 < -3$  was selected to identify poverty, as it yields above the 70th percentile in the mean Poverty headcount index, providing a clear demarcation for poverty classification.

both the error terms and some of the independent variables exhibit spatial correlation ( $y = X\beta + WX\theta + \epsilon$ ).

We also tested a Bayesian Geostatistical Model, considering its ability to incorporate prior information, handle complex spatial structures, and offer an assessment of uncertainty (Gelfand and Banerjee 2017). Additionally, we tested ML approaches such as Random Forest Regression, Decision Tree Regression, and Boosting Regression (XGBoost) to explore different predictive frameworks. We built Random Forest Regressors with 500, 1000, and 2000 trees to capture non-linear relationships and interactions between variables that might not be fully addressed by the spatial models (Belgiu and Drăguț 2016). Considering the importance of interpretability, we implemented a Decision Tree Regressor as well with depths of 3, 5, and 10 (Friedl and Brodley 1997) (Ramraj et al. 2016).

**Predicting discretized SEI:** As discussed the utility of the built model is not to accurately predict the exact index value, but to obtain a relative ordinal mapping of the GNDs based on their poverty level, we explored building a binary classification model where the predicted score of the model can be used to map the GNDs to a ranked list ordered by the estimated likelihood of poverty’s prevalence in the GND.

To obtain a binary label — where the positive class would indicate that a GND is poor — from the PC1, we discretized the space using a fifth data source: the poverty headcount index (HCI) at the Divisional Secretary’s Division (DSD) level, the second smallest administrative district in Sri Lanka. We used the percentile HCI to identify the PC1 threshold (-3) that yields above 70<sup>th</sup> percentile (see Figure 4). Any PC1 values below -3 were marked as positive, and the rest as negative (Department of Census and Statistics, Sri Lanka and World Bank 2015).

### 3.3 Validating Model Performance

Ideally, evaluation would require data from two distinct time points: one for training and another for validation. Lacking this, we employ two methods to simulate future generalization: bootstrapping and stratified cross-validation.

**Bootstrapping:** Randomly sampling GNDs into train and validation sets can result in leaking spatial information from the validation set to the train set, e.g., consider two adjacent GNDs where one is included in the train set and the other in the validation set. To reduce the risk of data leakage, we split the data based on Divisional Secretariat Divisions (DSDs). We randomly selected 60% of the DSDs to train and randomly selected 50% of the rest (20% of the dataset) to validate the models, and to maintain consistency in validation sample size across iterations. We repeat this process 1,000 times, generating different splits to minimize bias due to specific data configurations.

**Stratified Cross Validation** To account for the spatial heterogeneity of the data, we stratify the train-validation sample based on the 9 provinces in Sri Lanka. At each iteration, one province was held out as the test set, while the remaining provinces were used to train the model, to allow us to examine how well the model generalizes across different geographic regions, reflecting the variability in feature distributions and patterns (Wang, Zhang, and Fu 2016).

**Metrics:** To evaluate the performance of our models, we conducted a series of comparisons. While we initially assessed overall model accuracy using traditional statistical measures like  $R^2$ , mean squared error (MSE) (see Table 3), our primary focus was on evaluating the ability to rank the administrative divisions based on poverty. Therefore, we calculated precision and recall. We also looked into the Pearson correlation, overlap, and Jaccard similarity to account

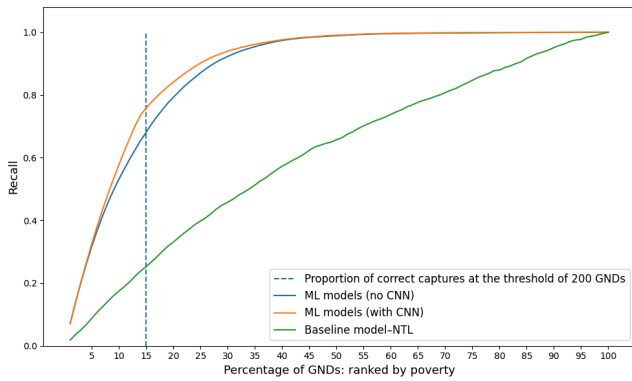


Figure 5: Recall Comparison: ML models With and Without CNN Features VS Baseline model NTL.

for how well the rankings correlate between the ground truth and model rankings.

**Baselines:** To compare our more expensive-to-build ML models, we built two simpler ranking approaches that yield an ordinal mapping of GNDs: (1) Ranking based on the VIIRS satellite night-time lights intensity (NTL), and (2) the population density (PD). With NTL, we assume that the higher intensities correspond with lower poverty levels, and with PD, we assume that higher density values correlate with higher degrees of poverty.

## 4 Results

**Models excelled at rapidly identifying the poorest regions**, while the baseline model–Night-time-light intensity—one of the most widely recognized poverty proxies, demonstrated a state of random guessing. Figure 5 shows recall comparisons with a hypothetical poverty alleviation initiative denoted by the vertical dotted line that can serve up to 200 GNDs. The ultimate goal of policymakers in such a scenario would be to make sure that they have selected the poorest 200 GNDs for the program. The recall curves illustrate that the ML models would correctly identify more than twice the poorest GNDs as the best-performing baseline model.

**Combination of the three data sources yielded better results:** Our analysis revealed that models combining CDR, RS, and satellite data consistently outperformed those based on either data source alone (See Table 3). Figure 5 further shows that, even though not significant, the features derived from the CNN architecture played a role in capturing more poor GNDs than without.

However, models using only RS data or only CDR data also performed comparably well in certain contexts. The combined CDR–RS model exhibited strong performance across both urban and rural areas, as well as at the national level. In rural areas, models relying solely on RS data performed reasonably well, though they fell short in urban environments. Conversely, CDR-only models showcased superior performance in urban areas but were less effective in

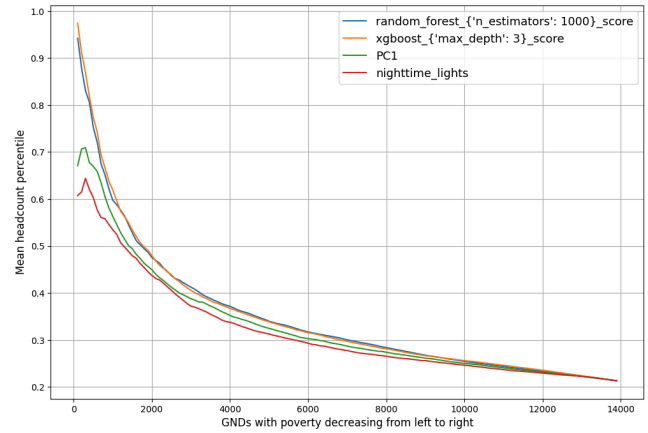


Figure 6: How the trajectories of ML models align strongly with the percentile GND-level headcount index.

rural settings. These results suggested the importance of selecting and utilizing different data types depending on the context.

**Baselines align well with ground-truth only in less critical stages:** Figure 6 shows how the mean headcount percentile (percentile mapping of HCI, which was processed to GND level from DSD, depending on the GND population) variations of the prediction scores of ML models for whole Sri Lanka and the scores of PC1 and baseline-NTL. The scores are indexed on the X-axis, with poverty decreasing from left to right. The ML model prediction scores strongly align with the mean headcount percentiles, while the baseline model aligns well with ground truth PC1 at less critical stages.

**ML models shows better alignment with PC1 ranking at both GND and DSD level:** The Pearson correlation values indicated a strong alignment between the rankings produced by the Random Forest, Decision Tree, and XGBoost models when compared to the Ground Truth PC1 rankings than the rankings of baselines and Bayesian Geostatistical models, particularly in the topmost ranks. Additionally, the overlap percentage was also high among the ML models (see Figure 7), showcasing the ability to pinpoint the most disadvantaged administrative regions.

**Context-specific data importance** For the whole country, features like night-time light intensity, population density, travel time to major cities, and elevation were crucial in decision-making for the ML models. In rural areas, it was the climate variables such as temperature and vegetation that played a significant role. While in urban areas, features derived from CDR data played a key role in the models to help decision-making.

## 5 Discussion

This work extends the previous studies that build predictive maps of poverty using a combination of CDR, RS, and satellite data. While it shows promising early results, the validity

Table 3: Regression Model Performance Across Different Poverty Metrics and Data Types

Data Type	Model	R <sup>2</sup>	Poorest 25 DSDs	Poorest 50 DSDs	Poorest 100 GNDs	Poorest 500 GNDs
<b>Whole Country</b>						
CDR + RS	SDEM Model	0.69	80%	60%	21%	46%
	SEM Model	0.74	76%	64%	23%	53%
	SLX Model	0.80	76%	64%	33%	64%
	Random Forest (1000 trees)	0.80	84%	64%	68%	75%
	XGBoost Regressor	0.81	88%	76%	70%	78%
BGM	0.77	80%	62%	31%	37%	
CDR only	SDEM Model	0.50	8%	26%	0%	30%
	SEM Model	0.50	72%	70%	0%	31%
	SLX Model	0.74	70%	62%	34%	41%
	Random Forest (1000 trees)	0.67	60%	58%	52%	58%
	XGBoost Regressor	0.70	77%	57%	56%	71%
BGM	0.72	72%	66%	11%	23%	
RS only	SDEM Model	0.71	0%	10%	0%	4%
	SEM Model	0.71	72%	60%	24%	54%
	SLX Model	0.74	0%	4%	0%	0%
	Random Forest (1000 trees)	0.71	76%	64%	68%	56%
	XGBoost Regressor	0.82	76%	66%	73%	68%
BGM	0.73	72%	62%	22%	54%	
<b>Rural</b>						
CDR + RS	SDEM Model	0.04	12%	18%	10%	14%
	SEM Model	0.68	36%	48%	24%	33%
	SLX Model	0.73	36%	50%	36%	45%
	Random Forest with 1000 trees	0.75	44%	48%	70%	72%
	XGBoost Regressor	0.75	44%	40%	52%	54%
BGM	0.70	36%	48%	22%	20%	
CDR only	SDEM Model	0.00	28%	33%	11%	26%
	SEM Model	0.49	40%	42%	12%	25%
	SLX Model	0.52	40%	46%	33%	22%
	Random Forest with 1000 trees	0.63	44%	44%	52%	50%
	XGBoost Regressor	0.65	44%	40%	44%	48%
BGM	0.49	44%	42%	13%	18%	
RS only	SDEM Model	0.65	36%	44%	24%	38%
	SEM Model	0.64	36%	44%	21%	37%
	SLX Model	0.73	36%	48%	66%	64%
	Random Forest with 1000 trees	0.75	68%	64%	66%	74%
	XGBoost Regressor	0.76	66%	64%	64%	66%
BGM	0.66	66%	48%	54%	52%	
<b>Urban</b>						
CDR + RS	SDEM Model	0.04	38%	50%	23%	58%
	SEM Model	0.56	56%	57%	47%	73%
	SLX Model	0.69	56%	62%	55%	76%
	Random Forest with 1000 trees	0.73	64%	64%	81%	77%
	XGBoost Regressor	0.77	60%	64%	87%	76%
BGM	0.71	68%	60%	59%	79%	
CDR only	SDEM Model	0.45	72%	62%	46%	67%
	SEM Model	0.46	72%	63%	46%	68%
	SLX Model	0.73	68%	66%	61%	79%
	Random Forest with 1000 trees	0.66	75%	67%	83%	83%
	XGBoost Regressor	0.59	74%	72%	73%	80%
BGM	0.68	64%	60%	55%	57%	
RS only	SDEM Model	0.12	44%	50%	17%	24%
	SEM Model	0.11	60%	54%	25%	32%
	SLX Model	0.55	57%	60%	41%	40%
	Random Forest with 1000 trees	0.65	55%	57%	61%	71%
	XGBoost Regressor	0.66	55%	61%	73%	77%
BGM	0.59	72%	63%	47%	55%	

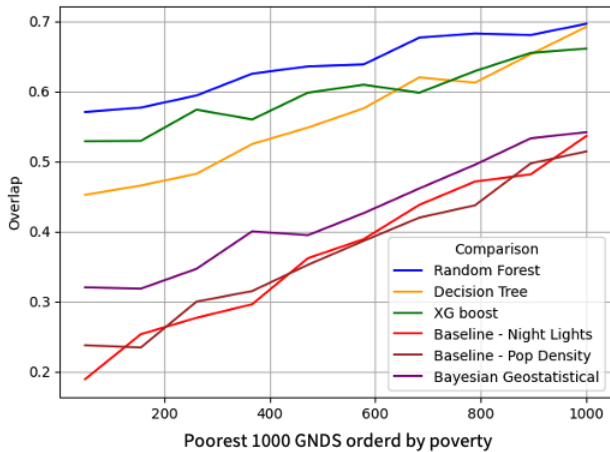


Figure 7: Overlap Of GNDs at different thresholds: ML-based VS Bayesian Geostatistical Model VS Baseline Models

of its approach is constrained by the limitations of poverty data and the challenges and limitations around the use of CDR 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 the GN division level, we utilized our own socio-economic index derived using the 2012 census data. This index was composed using a principal component analysis technique that leveraged household and demographic characteristics of households. This relied on 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 approxima-

tions of poverty compared to PCA methods.

### 5.1 Modeling choices

The modeling choices in this paper were guided by the practical policy problem of needing to find a given number of poorest administrative units. We have experimented with both a regression approach and a classification approach, given that the ground truth data we worked with was the output of a Principal Component Analysis. A significant challenge in this study is the lack of independent data to assess model generalizability. To address this, we developed models using nationwide data and intentionally overfitted them to census-derived poverty measures (both index and binary classifications). This approach assumes a stable relationship between poverty and the selected RS and CDR data over the census period, allowing us to infer spatial poverty distributions for intervening years using the same data sources.

### 5.2 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 (Mitterling et al. 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.

### 5.3 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 policymakers faced having varying priorities.

## References

- Agyemang, F. S.; Memon, R.; Wolf, L. J.; and Fox, S. 2023. High-resolution rural poverty mapping in Pakistan with ensemble deep learning. *Plos One*, 18(4): e0283938.
- Akinyemi, F. O. 2007. Spatial data needs for poverty management. *Research and theory in advancing spatial data infrastructure concepts*, 5(1): 261–277.
- Anselin, L. 2009. Spatial regression. *The SAGE handbook of spatial analysis*, 1: 255–276.
- Belgiu, M.; and Drăguț, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114: 24–31.
- Blumenstock, J.; Cadamuro, G.; and On, R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264): 1073–1076.
- Chi, G.; Fang, H.; Chatterjee, S.; and Blumenstock, J. E. 2022. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3): e2113658119.
- Cruz, M.; Foster, J.; Quillin, B.; and Schellekens, P. 2015. Ending extreme poverty and sharing prosperity: Progress and policies. *Policy Research Note*, 15(03).
- Department of Census and Statistics, Sri Lanka; and World Bank. 2015. The Spatial Distribution of Poverty in Sri Lanka. Technical report, Department of Census and Statistics, Sri Lanka, Battaramulla, Sri Lanka. Authored by Dung Doan under the guidance of David Newhouse and a team of DCS staff headed by Ms. Dilhanie Deepawansa, under the guidance of Dr. Amara J. Satharasinghe.
- Engstrom, R.; Hersh, J.; and Newhouse, D. 2016. Poverty in HD: What does high resolution satellite imagery reveal about economic welfare. Available online: [Pubdocs.worldbank.org/en/60741466181743796/Poverty-in-HD-draft-v2-75.pdf](https://pubdocs.worldbank.org/en/60741466181743796/Poverty-in-HD-draft-v2-75.pdf) (accessed on 1 December 2016).
- Espín-Noboa, L.; Kertész, J.; and Karsai, M. 2023. Interpreting wealth distribution via poverty map inference using multimodal data. In *Proceedings of the ACM Web Conference 2023*, 4029–4040.
- F. Dormann, C.; M. McPherson, J.; B. Araújo, M.; Bivand, R.; Bolliger, J.; Carl, G.; G. Davies, R.; Hirzel, A.; Jetz, W.; Daniel Kissling, W.; et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5): 609–628.
- Friedl, M. A.; and Brodley, C. E. 1997. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3): 399–409.
- Gao, X.; Asami, Y.; and Chung, C.-J. F. 2006. An empirical evaluation of spatial regression models. *Computers & Geosciences*, 32(8): 1040–1051.
- Gelfand, A. E.; and Banerjee, S. 2017. Bayesian modeling and analysis of geostatistical data. *Annual review of statistics and its application*, 4(1): 245–266.
- Jean, N.; Burke, M.; Xie, M.; Alampay Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Lawrence, M.; and Shipman, M. 2024. Positive Pathways through Polycrisis. *Victoria: Cascade institute*. Retrieved July, 6: 2024.

Lee, B. X.; Kjaerulf, F.; Turner, S.; Cohen, L.; Donnelly, P. D.; Muggah, R.; Davis, R.; Realini, A.; Kieselbach, B.; MacGregor, L. S.; et al. 2016. Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37: 13–31.

Mitterling, T.; Fenz, K.; Martinez Jr, A.; Bulan, J.; Ad-dawe, M.; Durante, R. L.; and Martillan, M. 2021. Compiling Granular Population Data Using Geospatial Information. *Asian Development Bank Economics Working Paper Series*, (643).

Ramraj, S.; Uzir, N.; Sunil, R.; and Banerjee, S. 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40): 651–662.

Senna, L. D. d.; Maia, A. G.; and Medeiros, J. D. F. d. 2019. The use of principal component analysis for the construction of the Water Poverty Index. *RBRH*, 24: e19.

Shrider, E. A.; Kollar, M.; Chen, F.; Semega, J.; et al. 2021. Income and poverty in the United States: 2020. *US Census Bureau, Current Population Reports*, (P60-273).

Steele, J. E.; Sundsøy, P. R.; Pezzulo, C.; Alegana, V. A.; Bird, T. J.; Blumenstock, J.; Bjelland, J.; Engø-Monsen, K.; De Montjoye, Y.-A.; Iqbal, A. M.; et al. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127): 20160690.

Wang, J.-F.; Zhang, T.-L.; and Fu, B.-J. 2016. A measure of spatial stratified heterogeneity. *Ecological indicators*, 67: 250–256.