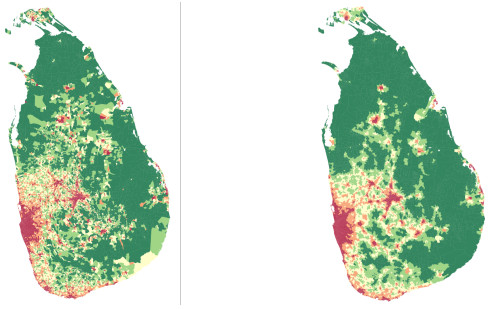
Informing Public Policy with Machine Learning: Mapping Poverty in Sri Lanka Using Mobile Call Detail Records and Remote Sensing Data

Chanuka Algama[1], **Merl Chandana**[1], **Viren Dias**[1, 2], **Kasun Amarasinghe**[1, 3]

[1]LIRNEasia, [2]Calcey, [3]Carnegie Mellon University
chanuka, merl@lirneasia.net
viren@calcey.com
kamarasi@andrew.cmu.edu

Abstract

Accurate estimations of the spatial distribution of poverty are vital to facilitating poverty alleviation initiatives and monitoring poverty over time. While censuses and surveys are the established benchmarks for poverty measurement, they are resource-intensive and time-consuming, making them impractical for continuous monitoring of spatial and temporal poverty disparities and rapid deployment of benefits programs. This gap is particularly pronounced in contexts like Sri Lanka, where even routinely collected proxy data—such as income, employment, and social benefits—are scarce and difficult to access. Our paper makes two key contributions: First, we replicate previous studies that use machine learning (ML) approaches based on mobile call records and remote sensing data to provide new insights into the spatial distribution of poverty for the first time in Sri Lanka. Second, we address a significant gap in the literature by proposing a framework for validating ML outputs on their ability to inform poverty alleviation initiatives. We use the most recent census data to establish a poverty index at the smallest administrative unit (GN division) level, and train ML models using remote sensing and mobile operator aggregated data to estimate the established index. We evaluate the models' ability to identify GN divisions with the highest poverty to report poverty alleviation interventions. An impressive showcase of the models results emerges through its performance 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 poorest 22 DSDs out of 25.



(a) Ground Truth-PC1

(b) ML Model Estimation

(c) Ground Truth-SAE

(d) Ground Truth-SAE: Processed to GND level

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).

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). Ending poverty in all its forms remains one of the most significant challenges and is the first target of the Sustainable Development Goals (SDGs) (Lee et al. 2016). To alleviate poverty effectively, reliable information is essential on where affected populations live. Such data can reveal spatial patterns of poverty, play a crucial role in allocating resources for poverty alleviation programs, and serve as a key component for monitoring poverty rates over time. This knowledge can guide development initiatives such as infrastructure, education, and healthcare improvements, leading to significant and sustained poverty reduction (Akinyemi 2007). Moreover, the spatial poverty data can be used to the effectiveness of both individual and regional in-

terventions can be monitored using this spatial data.

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. Large-scale surveys for targeted social assistance programs, such as cash transfers, provide valuable data but are resource-intensive and infrequent, limiting their utility for ongoing poverty monitoring and development initiatives.

In this work, we utilize spatially and temporally overlapping remote sensing (RS), call detail records (CDR), and census data to conduct the first systematic study using publicly available data sources to gain new insights into the spatial distribution of poverty in Sri Lanka. Building on the work of (Steele et al. 2017), who demonstrated the effectiveness of alternative data sources like mobile operator data and geospatial information for accurately estimating and monitoring poverty rates in Bangladesh. We extend this approach to Sri Lanka, highlighting the importance of evaluating how effectively the method informs real-world decision-making. While most past work focuses on statistical measures such as R^2 error to assess model performance, our approach takes a nuanced perspective. We consider how the model will be applied in practical decision-making scenarios. By working backward from the intended decision-making process, we aim to ensure that the model's outputs are not only statistically robust but also actionable and relevant in real-world contexts.

2 Datasets and Preparation

Recent research highlights the potential of satellite remote sensing data (RS data) and mobile operator call detail records (CDRs) for mapping poverty and socioeconomic well-being (Steele et al. 2017). 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). Conversely, CDRs offer valuable information on financial access, mobile phone usage, and potential economic activity, reflecting household economic conditions and behavior (Aiken et al. 2022). These data sources complement each other, offering a comprehensive perspective on socioeconomic conditions.

We use data from four sources: (1) Census data to derive a socio-economic index (SEI) to establish a "ground truth" for poverty distribution, (2) RS data, (3) CDR as readily available data sources for building poverty maps in the absence of census data, and (4) to compare and validate the results we used poverty headcount index (HCI) generated for Sri Lanka generated by the World Bank in collaboration with the Department of Census and Statistics (DCS). CDR was collected from two of the leading mobile network operators in Sri Lanka for 2013, which 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 offer coarser resolution in urban areas, focusing on land properties, but provide continuous coverage across regions and can be aggregated at desired geographic levels. RS and CDR data also complement each other due to their different spatial scales. HCI was generated using the small area estimation method which is a standard poverty mapping method that has been widely used by both the World Bank and international researchers to estimate poverty at disaggregate administrative levels. It is at the divisional secretariat (DSD) level in Sri Lanka, where the data from the 2012 Census of Population and Housing (CPH) and the 2012/13 Household Income and Expenditure Survey (HIES) are utilized for the generation.

To prepare the data sources, we performed the following preprocessing steps: (1) deriving the poverty index from census data and (2) aligning the spatial scale of RS data and CDRs with the geographical boundaries of the Grama Niladhari Divisions (GNDs).

2.1 Deriving the poverty index from Census data

Since Sri Lanka does not have open poverty data at the division level of the GN for 2013 (the year was determined by the CDR data period used in the study), we use principal component analysis techniques (PCA) (Dias et al. 2020) to derive a socio-economic index for each GND in Sri Lanka. Recently, PCA-based spatial poverty analysis using census data has emerged as a popular and reliable measure of socioeconomic well-being (Krishnan, 2010). We selected 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 the demographic characteristics of the household and used a variable elimination process to reduce the number of variables to 61. Then, we normalized and standardized each variable and performed PCA on the dataset. We multiplied the weights of the resulting first principal component with the standardized data set and summed each row to produce a score for each GN Division. This score was to serve as the socioeconomic index. We denote the first principal component of the dataset by PC1.

2.2 Aligning the spatial scales of RS and CDRs

Although the census data contained information at the GND level, RS and CDR were collected at different spatial resolutions and did not align directly with the GND boundaries. All RS data were processed to spatially align with the polygon boundaries of the GN divisions. The CDR metrics were first calculated at Voronoi cell levels and then mapped to the GN division limits (see Figure 2). Depending on the feature being considered, each polygon was assigned RS and CDR values that represent the mean, sum, or mode of the corresponding data.

The 37 GN divisions that recorded zero population in the census were removed from the study. For two main reasons,

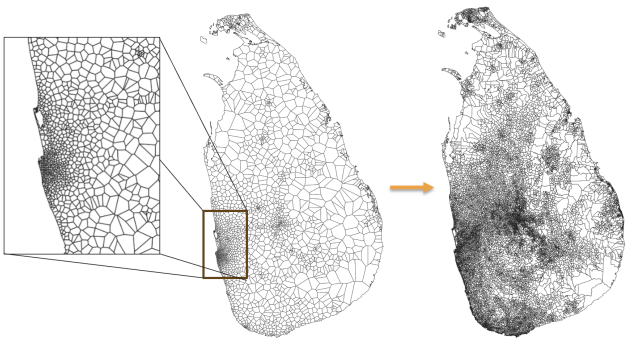


Figure 2: Spatial scaling of CDR data from Voronoi cell-level calculations to GN boundaries, where each GN polygon is assigned CDR values based on the mean, sum, or mode of the data, depending on the feature under analysis.

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
	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

Table 1: Features derived from CDR data

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.

3 Model Building and Validation

3.1 Feature Generation from RS and CDR Data

Using CDR data, we create 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 distance to roads or major urban areas, We obtained data from existing sources (e.g., maps produced by other researchers and agencies). A full list of the RS features is provided in Table 2.

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, ensuring that each observation had a fixed number of spatial neighbors, and effectively capturing spatial relationships based on proximity. This is particularly pertinent in spatial poverty mapping where geographic closeness often implies similar poverty conditions.

Information type	Feature
Accessibility	Accessibility to populated places with more than 50k people
Population	Population count
	Population 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 ethic groups

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

3.2 Data Stratification

To cater to the different modeling approaches employed in this study, we adopted various data stratification strategies to ensure representative sampling across administrative divisions. The first strategy involved partitioning the dataset such that 60% of the DSDs were allocated for training, while 20% each were designated for testing and validation. The second strategy focused on stratifying the data geographically by province: in each of the nine rounds, data from eight provinces were used for training, and the remaining province was reserved for testing, while ensuring that the model's performance was evaluated across all regions, accounting for potential spatial heterogeneity in the data.

3.3 Model Development

We explored using machine learning to predict the continuous socio-economic index and a discretized version to further simplify the prediction process.

Prediction VS Inference: The primary goal of this study is to develop a model capable of accurately predicting

poverty at the Divisional Secretariat Division (DSD) and Grama Niladhari Division (GND) levels, using satellite imagery and call detail records (CDR) data. While traditional econometric models focus on inference—determining which covariates most strongly correlate with poverty such as understanding how night-time lights or population density directly influence poverty levels. The objective here is not to understand the causal mechanisms behind poverty. Instead, our focus is on building a predictive framework that can provide timely and actionable poverty estimates.

Our approach aims to leverage readily available data sources to estimate poverty levels with sufficient accuracy, independent of the need for periodic, labor-intensive census data, as they are only conducted approximately once every ten years. This infrequency limits their utility in responding to dynamic socio-economic conditions. For example, if a poverty alleviation initiative needs to be implemented quickly, relying solely on census data would be insufficient due to its outdated nature.

Furthermore, Our goal is not to estimate the precise socio-economic index (SEI). Instead, we aim to rank administrative divisions from the poorest to the least poor and evaluate how well our model replicates this ground truth ranking, assessing the accuracy of the predicted rankings relative to the SEI, rather than achieving exact numerical predictions.

Regression Approach: We initially applied a diverse set of methods, including frequentist and Bayesian approaches, and ensemble methods such as Random Forest and XGBoost, allowing us to rigorously model the spatial patterns and relationships present in our dataset.

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. 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). To capture more complex spatial dependencies, we utilized SDEM to integrate the features of both the SEM and SLX models, allowing it to account for spatial spillover effects. SDEM is particularly valuable when both the error terms and some of the independent variables exhibit spatial correlation. 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 Drăguţ 2016). Considering its interpretability, we implemented a Decision Tree Regressor with depths of 3, 5, and 10 (Friedl and Brodley 1997). (Ramraj et al. 2016).

Classification Approach: Since 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 fourth 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^{th} percentile (see Figure 3). Any PC1 values below -3 were marked as positive, and the rest as negative. (DCS 2024)

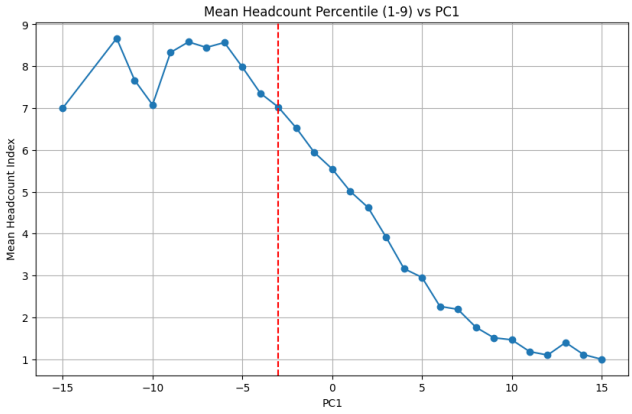


Figure 3: Mean Headcount Index vs. PC1: A Threshold Analysis for Poverty Classification. The threshold of $PC1 < -3$ was selected to identify poverty, as it yields above the 70th percentile in the mean headcount index, providing a clear demarcation for poverty classification.

3.4 Validating Model Performance

While using ML models built on alternative data sources like RS and CDR data to map poverty is well-established, a key gap lies in validating these models for their ability to inform policy decisions. Most studies train ML models on household surveys to estimate poverty at an administrative level but rarely validate how well these models can guide resource allocation, given data limitations.

In this work, we address this gap. Ideally, evaluation would require data from two distinct time points: one for training and another for validation. Lacking this, we employ three methods to simulate future generalization: bootstrapping, random cross-validation, and stratified cross-validation. We assess model robustness in terms of its effectiveness and efficiency in informing resource allocation.

Bootsrapping: 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 this risk of data leakage, we split 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, to further reduce the risk of spatial data leakage. 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 Sri Lanka's 9 provinces. At each iteration, one province was held out as the test set, while the remaining provinces were used to train the model, allowing 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).

Random Cross Validation For random cross-validation, the data was split by DSDs into 9 equal folds. In each of the 9 iterations, one fold was isolated as the test set, and the model was trained on the remaining 8 folds. Metrics such as recall, precision, rank correlation, and overlap were computed for each iteration across all three methods.

As we mentioned above, we evaluate all models based on the ranked list of the GNDs they produce. One particular challenge in this work is the absence of a second set of data where we can evaluate the generalizability of the models. Given this constraint, we build models using the data from the whole country and attempt to overfit the models to the census data-derived targets (i.e., the poverty index and the binary labels). We assume that if the model can capture the relative rankings of the GNDs accurately correlating RS and CDR data to the census data, the RS and CDR data can provide an approximate estimate of relative rankings in the absence of census data.

Metrics: To evaluate the performance of our models, we conducted a series of comparisons using both statistical and policy-relevant metrics. While we initially assessed model accuracy using traditional metrics like R^2, mean squared error (MSE), precision, and recall, our primary focus was on evaluating the models on their utility to inform policy decisions. Specifically, we prioritized the models' ability to accurately identify the poorest 25, 50, and 100 Divisional Secretariat Divisions (DSDs), as well as the poorest 100, 500, 1000, and 3000 Grama Niladhari Divisions (GNDs). This allowed us to assess how well each model performed in identifying the most economically disadvantaged areas, which is crucial for informing targeted policy interventions.

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.

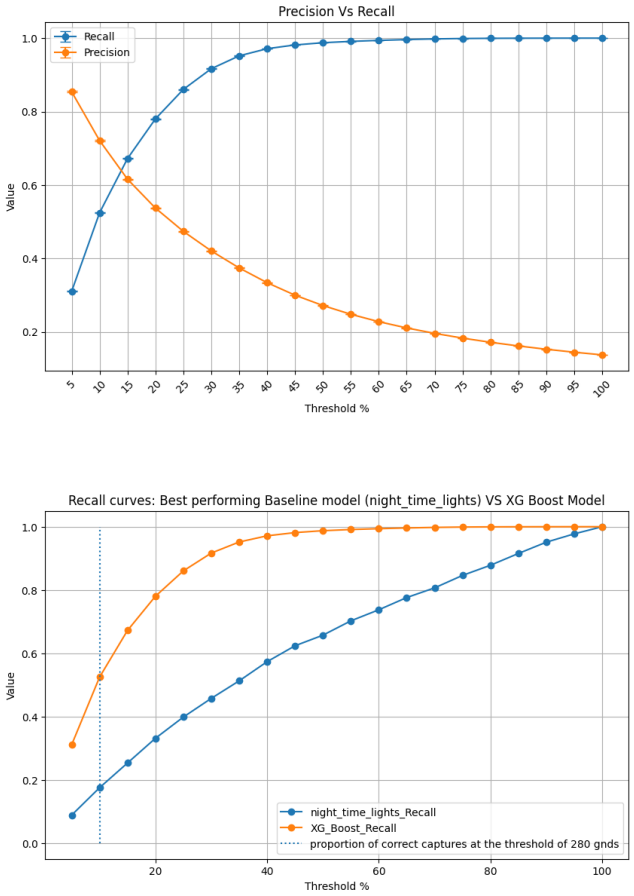


Figure 4: First Figure depicting the Precision and Recall curves for each threshold of Gnds in the Bootstrapping Approach, while the second figure illustrates the Recall curve comparison with the baseline model: night-time lights. The vertical dotted line on the x-axis corresponding to 10% threshold shows the superior performance of the model 'XG Boost' over the baseline, in correctly capturing the poorest regions.

4 Results

Combination of the two data sources yielded better results: Our analysis revealed that models combining CDR and RS data consistently outperformed those based on either data source alone (see Table 3). 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 rural settings. These results suggested the importance of selecting and utilizing different data types depending on the context. The variation in performance across different geographic areas highlights the necessity of tailoring data sources to the specific charac-

teristics of the region being studied, thereby enhancing the accuracy and relevance of the models.

Baselines align well only in less critical stages: Figure 5 shows how the mean headcount percentile varies with different thresholds across various prediction models, the Baseline and Ground Truth PC1 follow similar trajectories, especially in the middle and lower threshold ranges. But doesn't align well at the top which is the area of interest in identifying the poorest administrative division. Since the comparison doesn't provide a complete picture of how well the rankings correlate. We looked into the Pearson correlation, overlap, and Jaccard similarity among the models and PC1-based rankings.

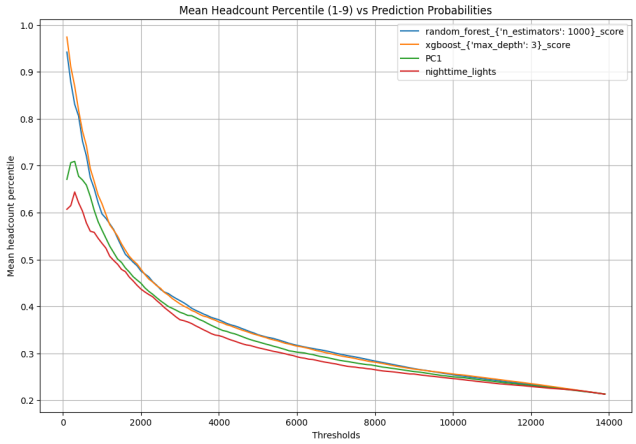


Figure 5: How the percentile DSD-level headcount index varies with the different ranking approaches. The trajectories of the Random Forest and XGBoost align well with the mean headcount index at the most critical stage of correctly and quickly capturing the poorest regions.

ML models shows impressive 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, particularly in the topmost ranks. Additionally, the overlap percentage was high (see Figure 6), signifying the reliability of the models in pinpointing the most disadvantaged administrative regions.

Models use a combination of the CDR and RS features Different factors emerged as significant depending on the context. For the whole country, features like night-time lights, population density, travel time to major cities, and elevation were crucial. In rural areas, climate variables such as temperature and vegetation played a more significant role. While for CDR data, unique tower count and night-time call count were identified as key features.

5 Discussion

This work extends the work done by (Steele et al. 2017) to build predictive maps of poverty using a combination of

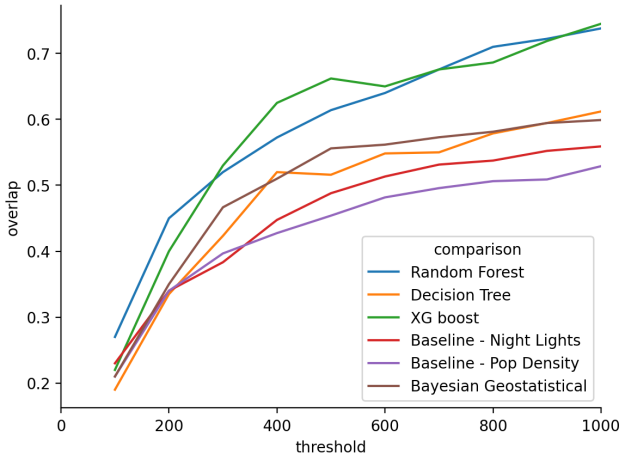


Figure 6: Overlap Of GNDs at different thresholds : ML-based VS Bayesian Geostatistical Model VS Baseline Models

CDR and RS data. While it shows promising early results, the validity 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 approximations 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

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

Data Type	Model	R^2	Poorest 25 DSDs	Poorest 50 DSDs	Poorest 100 GNDs	Poorest 500 GNDs
Whole Country						
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%	74%	85%
	XGBoost Regressor	0.81	84%	66%	76%	85%
	BGM	0.77	80%	62%	31%	59%
CDR only	SDEM Model	0.50	8%	26%	0%	60%
	SEM Model	0.50	76%	70%	0%	61%
	SLX Model	0.74	70%	60%	34%	41%
	Random Forest (1000 trees)	0.67	60%	58%	52%	58%
	XGBoost Regressor	0.70	80%	57%	56%	71%
	BGM	0.72	72%	66%	11%	42%
RS only	SDEM Model	0.71	0%	20%	0%	4%
	SEM Model	0.71	72%	60%	24%	54%
	SLX Model	0.80	0%	4%	0%	0%
	Random Forest (1000 trees)	0.71	76%	64%	68%	56%
	XGBoost Regressor	0.82	76%	66%	73%	78%
	BGM	0.73	72%	62%	22%	54%
Rural						
CDR + RS	SDEM Model	0.04	12%	28%	10%	14%
	SEM Model	0.68	36%	48%	24%	55%
	SLX Model	0.73	36%	50%	36%	65%
	Random Forest with 1000 trees	0.75	44%	48%	73%	85%
	XGBoost Regressor	0.75	44%	48%	77%	84%
	BGM	0.70	36%	48%	22%	59%
CDR only	SDEM Model	0.00	40%	42%	11%	30%
	SEM Model	0.49	40%	42%	12%	31%
	SLX Model	0.52	40%	46%	33%	62%
	Random Forest with 1000 trees	0.63	44%	48%	64%	75%
	XGBoost Regressor	0.65	44%	48%	63%	72%
	BGM	0.49	44%	42%	13%	46%
RS only	SDEM Model	0.65	36%	44%	24%	53%
	SEM Model	0.64	36%	44%	21%	51%
	SLX Model	0.73	36%	48%	38%	64%
	Random Forest with 1000 trees	0.75	40%	36%	66%	74%
	XGBoost Regressor	0.76	40%	44%	64%	66%
	BGM	0.66	36%	48%	24%	54%
Urban						
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%	90%
	XGBoost Regressor	0.77	60%	64%	87%	95%
	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	64%	66%	73%	89%
	XGBoost Regressor	0.59	64%	64%	72%	77%
	BGM	0.68	64%	60%	45%	66%
RS only	SDEM Model	0.12	44%	50%	17%	54%
	SEM Model	0.11	60%	54%	25%	55%
	SLX Model	0.55	57%	60%	41%	61%
	Random Forest with 1000 trees	0.65	55%	60%	61%	80%
	XGBoost Regressor	0.66	55%	61%	77%	83%
	BGM	0.59	72%	63%	47%	70%

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

Aiken, E.; Bellue, S.; Karlan, D.; Udry, C.; and Blumenstock, J. E. 2022. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903): 864–870.

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 Drăguţ, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114: 24–31.

Cruz, M.; Foster, J.; Quillin, B.; and Schellekens, P. 2015. Ending extreme poverty and sharing prosperity: Progress and policies. *Policy Research Note*, 15(03).

DCS. 2024. Poverty Indicators. Accessed: 2024-08-16.

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/607414661817437 96/Poverty-in-HD-draft-v2-75. pdf (accessed on 1 December 2016)*.

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.

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.; Addawe, 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.

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.