Predicting community-level poverty in Bolivia: Insights from satellite imagery, census data, and spatial modeling

Osmar Bolivar *
Ministerio de Economía y Finanzas Públicas

Carlos Mendez
Nagoya University

December 4, 2024 – Version 0.4
Preliminary

Abstract

This study predicts community-level poverty headcount ratios in Bolivia for 2022, using a combination of machine learning, remote sensing, and spatial modeling techniques. By analyzing Unsatisfied Basic Needs (UBN) poverty in 953 communities between 2012 and 2022, the methodology successfully reveals a general decline in poverty rates, with approximately 50% of communities projected to fall below the 41.8% threshold in 2022. Notably, poverty reductions are more pronounced in communities with lower initial poverty levels, while regional disparities persist, with urban areas consistently exhibiting lower poverty rates. The approach demonstrates the effectiveness of combining machine learning and geospatial data to inform targeted poverty reduction strategies in Bolivia, offering a replicable model for other developing countries facing scarce and outdated high-resolution spatial data. This method provides valuable insights for policymakers seeking to address poverty at a granular level despite data limitations.

Keywords: Poverty, machine learning, spatial analysis, geospatial data
JEL Codes: C8, I32, R10, R11, R58

*Working paper presented at the ECINEQ - Latin America and the Caribbean Chapter (ECINEQ-LAC) Conference. Please do not cite. Corresponding author: osmar.economics@gmail.com

1. Introduction

The accurate measurement of poverty has long been a central concern for economists. Despite significant advances, poverty quantification remains a challenging process, shaped by choices around definitions, welfare indicators, and measurement units.

Pioneering work in this field, such as Orshansky (1969) and Galbraith (1998), introduced the first official poverty thresholds in the United States, employing methods like the cost of a basic food basket and household income comparisons. Building on these early contributions, scholars such as Thon (1979), Foster et al. (1984), Kanbur (1990), and Jenkins and Lambert (1997) have expanded poverty measures, providing deeper insights into the phenomenon.

Yet, as noted by Atkinson (1987), poverty measurement remains complex and contested. Atkinson emphasized the need for multidimensional approaches that incorporate material deprivation, social exclusion, and quality of life. Such frameworks, he argued, should reflect changing living costs and economic conditions, further underscoring the limitations of purely monetary measures.

In response to these limitations, many researchers advocate for multidimensional poverty measures. Bourguignon and Chakravarty (1999), Tsui (2002), Alkire and Foster (2011), and others have developed indicators encompassing aspects such as income, education, health, housing, and access to services. These approaches provide a more comprehensive view of poverty and help identify vulnerable groups, as Alkire and Foster (2011) highlighted, offering critical information for effective poverty reduction policies.

However, in regions where traditional surveys are expensive or difficult to conduct, remote sensing and machine learning offer promising alternatives. Studies by Jean et al. (2016), Blumenstock (2016), Piaggesi et al. (2019), and others have demonstrated the potential of using satellite imagery and algorithms to estimate poverty and assess public policy impacts. While this approach presents challenges, its ability to enhance the accuracy and timeliness of poverty data is a compelling reason for its adoption.

In Bolivia, where community-level poverty data are scarce—only available from

the 2012 census—the use of remote sensing and machine learning provides an opportunity to generate new insights. This limitation, common across many developing countries, restricts high-frequency spatial analyses of poverty, which are crucial for timely policy intervention.

The primary aim of this study is to predict community-level poverty headcount ratios for 2022 using machine learning algorithms and satellite data. By bridging the gap between 2012 census data and predicted 2022 estimates, this research offers a dynamic view of poverty trends in Bolivia. Additionally, this study serves as a replicable model for other developing countries where spatial poverty data are scarce or outdated. The approach leverages modern data sources and machine learning techniques to produce high-resolution, up-to-date poverty estimates, making it a valuable tool for countries facing similar data limitations.

Moreover, an Exploratory Spatial Data Analysis (ESDA) will be conducted to reveal spatial patterns in community-level poverty, offering further insights that can guide policymakers in targeting poverty alleviation efforts. While the focus is on Bolivia, the methodology and findings can inform poverty mapping strategies in other developing contexts, particularly where traditional data collection is costly or logistically challenging.

2. Methods and Data

2.1 Predicting Poverty

In Bolivia, the only source of information that allows for the computation of poverty metrics at the community level is the 2012 census. The Unsatisfied Basic Needs (UBN) poverty headcount ratio can be constructed since the census data does not include monetary aspects such as income.

This study proposes a methodological framework to predict the UBN poverty headcount ratio at the community level in Bolivia for the year 2022, utilizing machine learning algorithms and remote sensing data. This methodology can be summarized in two steps:

- **Data:** Variables describing the characteristics of the communities for the years 2012 and 2022 are generated through the processing of remote-sensed data.
- **Prediction:** Machine learning algorithms are trained to accurately predict the poverty headcount ratios at the community level, using the 2012 census data as a reference. These validated models are then applied to predict community-level poverty in 2022.

2.1.1 Study Communities

The 2012 census records 19,420 communities in Bolivia; however, there is no official source that establishes the geographic boundaries of these communities. Since this study relies on information about the characteristics of the communities derived from satellite image processing, it is necessary to define the spatial extent or concentration of the population in these communities.

Thus, the scope of this study is limited to communities with a population greater than 500 inhabitants, resulting in the prediction of poverty headcount ratios for 953 study communities. Households in smaller communities tend to be dispersed, making it challenging not only to establish the potential extent of the community but also to construct indicators of their characteristics.

The procedure for defining the extent of the study communities is detailed in Appendix A.

2.1.2 Target Variable and Features

The target variable is the UBN poverty headcount ratio. Although data from the 2012 census is available in Bolivia, there is no official publication of poverty metrics by UBN at the community level. Given this limitation, we processed the socioeconomic data from the 2012 census following the methodological guidelines of Unsatisfied Basic Needs (INE, 2015), and constructed a UBN poverty indicator that represents the percentage of the population in a community that is in poverty because they live

below the minimum standards to meet their basic needs.[1]

The UBN poverty measurement evaluates the presence or absence of essential goods and services in households, constituting a direct and observational method for assessing the satisfaction of basic needs. In Bolivia, the incidence of UBN poverty is constructed from the aggregation of deficiencies in four components: Housing, Basic services and supplies, Education, and Health. Due to the lack of community-level information on these aspects, we resorted to collecting and processing satellite imagery and GIS data to generate variables that approximate these determinants of UBN poverty.

These variables are the features used to train and validate the poverty forecasting algorithms. The following paragraphs explain the features according to their information source.

Open Street Map Data

The use of Open Street Map (OSM) data variables as predictors of income and poverty offers an innovative approach to addressing economic and social issues in regions where traditional data availability may be limited (Feldmeyer et al., 2020), as is the case in Bolivia.

Based on OSM databases, variables such as the number of banks, schools, and hospitals within the study communities are generated for both 2012 and 2022. These indicators reflect access to education and health services, as well as economic opportunities, which are fundamental not only for greater access to the mentioned services but also for the overall socioeconomic well-being and development of the communities.

Land Use Coverage

For the years 2012 and 2022, images from the "MODIS Land Cover Type product (MCD12Q1)" were downloaded, providing global land cover maps with a spatial resolution of 500 meters.

[1]This implies that this percentage of the population falls into the categories of moderate poverty, indigence, or marginalization, according to the parameters established in the UBN methodology.

Based on this information, indicators are obtained for the area (in square kilometers) of the extent in the communities corresponding to the coverages of: Urban areas and Croplands. For example, in a community j during year t, the area of croplands is defined as the number of pixels categorized as crops within the community's extent multiplied by 0.25 —since each 500-meter pixel covers an area of 0.25 km^2.

These variables can be related to the housing component of the UBN methodology, as houses with better materials are more likely to be categorized within urban area pixels. Furthermore, in general, changes in these land use coverages tend to be related to changes in economic activity, and therefore in poverty (Liu et al., 2021; Zhou et al., 2021).

Urban Settlements

The classification of urban settlements involves identifying and mapping areas of high human development density, such as buildings and infrastructure. These areas are strong indicators of economic activity, as they often correspond to densely populated areas with high economic productivity, leading to lower levels of poverty.

For this study, we use the urban settlement raster layers from Bolivar (2023). These files are based on the "Global Human Settlement Layer" but focus on providing a binary classification of urban settlements for Bolivia in the years 2012 and 2022.[2] In this framework, for year t in community j, the area of urban settlements is defined as the number of pixels classified as urban settlements within the community's extent multiplied by 0.25.

Nighttime Light

Nighttime light is also employed as a potential predictor of poverty, as variations in luminosity have proven to be adequate indicators for explaining differences at the cross-sectional and temporal levels in income levels, and consequently, in poverty

[2]Bolivar (2023) predicts the classifications of urban settlements for the years 2012 and 2022 by training a random forest algorithm. This algorithm uses 2015 GHSL images to classify 1 km pixels as urban settlements, using data from the blue, green, red, near-infrared, and shortwave infrared 1 and 2 bands of the Landsat-8 satellite, along with nighttime light images at the pixel level.

levels. Specifically, from the collection "VIIRS Lunar Gap Filled BRDF Nighttime Lights Daily L3 Global 500m", luminosity images for the years 2012 and 2022, covering all of Bolivia, are obtained. Subsequently, for a community j in the year t, the average luminosity intensity indicator is constructed, representing the average of the luminosity values in the pixels located within the community boundaries.

Electricity

Various studies have highlighted the close relationship between access to electricity and the socioeconomic well-being of communities (Foster and Rosenzweig, 2010; Dinkelman, 2011). Likewise, in the UBN methodology for Bolivia, access to electricity is directly included as one of its determinants. In this regard, vector layers of the medium voltage electricity grid in Bolivia for 2012 and 2022 were obtained.

The density of this grid by community can be considered a proxy for access to electricity, as it reflects the infrastructure and coverage of the electricity grid in a given area. A higher density of the medium voltage grid may suggest that households and businesses are more likely to have access to reliable electricity services. With this information, for a community j in the year t, the medium voltage electricity grid density indicator is constructed as the ratio between the total length (in km) of this grid within the community's extent and the community's total area.

Main Roads

The effect of main roads and highways on income and poverty has been studied extensively due to their close relationship with economic development and access to opportunities (Banerjee et al., 2020; Bolivar, 2022). A road network facilitates the transport of goods and services, connects rural areas with urban centers, and improves people's mobility, having a positive impact on economic growth and poverty reduction. Using vector layers of the main road network in Bolivia for 2012 and 2022,[3] the variable of primary road density is generated as the ratio between the total length of these roads within the community and the community's total area.

[3]This network is referred to as the "Red Vial Fundamental"

Time-Invariant Characteristics

Additionally, four time-invariant characteristics of the study communities are generated: the area of the delimited community extent; and dummy variables identifying whether the community is a population center, intermediate city, or capital city. The use of these variables as predictors of poverty levels at the community level is based on the idea that these initial characteristics determine heterogeneous poverty levels.

2.1.3 Training, Validation, and Production Sets

The dataset comprises 13 features derived from remote-sensed and GIS data, as well as time-invariant characteristics at the community level. These variables are available for the 953 study communities in both 2012 and 2022. The target variable, UBN poverty headcount ratio, is available for 2012. The objective is to predict community-level poverty for 2022 using machine learning techniques.

Data from 2012, including both the target variable and features, are used to form the training and validation sets. Seventy percent of the 2012 data is randomly assigned to the training set (667 observations), while the remaining thirty percent is allocated to the validation set (286 observations).

The production set for 2022 consists of the features of the 953 study communities, which are used to predict the UBN poverty headcount ratios. To ensure comparability and avoid bias, all variables in the training, validation, and production sets are normalized using z-score normalization.

2.1.4 Machine Learning Algorithms

The following algorithms were selected for training and validation, with the aim of predicting the 2022 UBN poverty. These algorithms were chosen due to their prevalence in forecasting literature and their capability to model both linear and non-linear relationships:

1. *Ridge Regression* (L2 regularization) adds a penalty proportional to the square of

the coefficients' L2 norm:

$$\text{minimize } J(\beta) = \text{SSE} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

2. *Lasso Regression* (L1 regularization) adds a penalty proportional to the coefficients' L1 norm:

$$\text{minimize } J(\beta) = \text{SSE} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

3. *ElasticNet* combines L1 and L2 penalties:

$$\text{minimize } J(\beta) = \text{SSE} + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \tag{3}$$

4. *Decision Tree Regressor* recursively splits the feature space into smaller regions, selecting at each step the feature and threshold that best splits the data to minimize mean squared error (MSE):

$$\text{MSE}(R) = \frac{1}{|R|} \sum_{i \in R} (y_i - \bar{y}_R)^2 \tag{4}$$

   The process continues until a stopping criterion, such as maximum depth or minimum samples per leaf, is met.

5. *AdaBoost Regressor* trains a series of weak learners, adjusting sample weights to focus on previously mispredicted data. The final model is a weighted sum of these learners:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{5}$$

6. *Gradient Boosting Regressor* builds an ensemble of weak learners, where each learner corrects the errors of the previous ones. The final prediction is the

weighted sum of individual learners:

$$\hat{y}_i = \sum_{m=1}^{M} \gamma_m f_m(x_i) \tag{6}$$

7. *Random Forest Regression* is an ensemble method that builds multiple decision trees on random subsets of features. The final prediction is the average of all trees' predictions:

$$y' = \frac{1}{T} \sum_{t=1}^{T} f_t(x') \tag{7}$$

8. *Extra Trees Regressor* is similar to Random Forest but introduces more randomness in selecting splits, resulting in diversified trees. The final prediction is the average of all trees' outputs:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{8}$$

Hyperparameter tuning, primarily via k-fold cross-validation, is crucial for optimizing model performance. This process involves iteratively selecting hyperparameter values to maximize model performance, ensuring robustness and avoiding overfitting. The aforementioned algorithms undergo a rigorous fine-tuning process based on a 10-fold cross-validation method.[4]

2.1.5 Poverty Prediction

After training the algorithms, the Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) are calculated for the validation set. Based on these evaluation metrics, the best-performing algorithms for predicting UBN poverty headcount ratios are identified.

[4]The k-fold cross-validation process is a technique used in machine learning to evaluate the performance and generalization ability of a model. It involves dividing the dataset into k equal-sized subsets or folds. The model is trained and evaluated k times, each time using a different fold as the validation set and the remaining folds as the training set.

The following procedure is implemented to predict the UBN poverty incidence for each community in 2022:

1. For each of the trained and validated algorithms (i, $\forall\ i = 1, \ldots, 8$), forecasts of UBN poverty headcount ratios are obtained for both 2012 ($\hat{y}_{i,2012}$) and 2022 ($\hat{y}_{i,2022}$). The 2022 forecast utilizes the production dataset.

2. Independently for the years 2012 and 2022 ($t = 2012, 2022$), the weighted geometric mean of the forecasts from all algorithms is computed. The weights correspond to the inverse of the MSE values ($\alpha_i = \frac{1}{MSE_i}$). Averaging individual forecasts is an effective strategy supported by the literature to improve the quality and reliability of predictions in regression problems (Wolpert, 1992; Breiman, 1996; Dietterich, 2000).[5]

$$\bar{\hat{y}}_{i,t} = \left(\prod_{i=1}^{B} \hat{y}_{i,t}^{\alpha_i} \right)^{\frac{1}{\sum_i \alpha_i}} \tag{9}$$

3. The difference (Δ) in percentage points between the forecasts (weighted geometric means) for 2022 and 2012 is calculated for each community.

$$\Delta_i = \bar{\hat{y}}_{i,2022} - \bar{\hat{y}}_{i,2012} \tag{10}$$

4. The final forecast of UBN poverty headcount ratios for 2022 ($\Upsilon_{i,2022}$) is determined by applying the difference calculated in step (3) to the observed poverty data for 2012 ($y_{i,2012}$). This approach ensures better comparability between the UBN poverty headcount ratios of 2012 and 2022.

$$\Upsilon_{i,2022} = y_{i,2012} + \Delta_i \tag{11}$$

[5]The geometric mean is calculated because it is less sensitive to outliers compared to the arithmetic mean.

2.2 Spatial Modeling

Exploratory Spatial Data Analysis (ESDA) functions as a preliminary method for examining spatial patterns in georeferenced data. This approach allows researchers to identify the geographic distributions of socioeconomic phenomena prior to performing advanced statistical modeling. ESDA incorporates various techniques for visualizing spatial distributions, identifying atypical locations, and detecting patterns of spatial association (Anselin, 1999).

A key aspect of ESDA is the investigation of spatial clustering and the identification of statistically significant spatial clusters. The detection of spatial clusters constitutes a central component of ESDA. Spatial dependence analysis in ESDA integrates the concepts of attribute similarity and locational proximity. This integration is founded on Tobler First Law of Geography, which states that ”everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). Spatial clusters may indicate the presence of agglomeration economies, knowledge spillovers, or other spatially bounded economic processes.

In the context of the ESDA framework, a global spatial dependence analysis examines spatial randomness and clustering strength. This analysis primarily utilizes Moran’s I statistic, which ranges from -1 to +1. Moran’s I measures the association between values at a given location and those of neighboring areas. A statistically significant Moran’s I value indicates the presence of spatial autocorrelation. Positive values suggest clustering, while negative values imply a checkerboard pattern. Local spatial dependence analysis, often derived from global statistics, identifies specific spatial clusters and outliers. This analysis classifies regions into four distinct groups: high-high clusters, low-low clusters, high-low outliers, and low-high outliers.

3. Results on Predicting Poverty

3.1 Hyperparameter Fine-tuning

The paragraphs below detail the hyperparameter tuning using 10-fold cross-validation for the algorithms in Section 2.1.4.

1. **Ridge**: The regularization parameter λ was tuned over 1,000 values from 10^{-5} to 10^{5} on a logarithmic scale. The optimal λ minimized the MSE.

2. **Lasso**: Similarly, λ was tuned over 1,000 values from 10^{-5} to 10^{5}, with the optimal λ minimizing the MSE.

3. **ElasticNet**: Two hyperparameters were tuned: λ (1,000 values from 10^{-5} to 10^{5}) and α (values from 0.05 to 0.95 in increments of 0.01). The optimal values minimized the MSE.

4. **Decision Tree Regressor**: Hyperparameters tuned included maximum depth (d), minimum samples to split (mss), and minimum samples at leaf (msl), using ranges from prior research.

5. **AdaBoost Regressor**: Hyperparameters tuned were maximum depth (d) from 3 to 10, number of estimators (T) from 50 to 200, learning rate (α_t) from 0.01 to 3, and loss function (linear, squared, exponential). The optimal values minimized the MSE.

6. **Gradient Boosting Regressor**: Hyperparameters tuned included learning rate (γ_m) from 0.01 to 2, maximum depth (d) from 3 to 10, number of estimators (T) from 100 to 500, and minimum samples to split (mss) from 2 to 20. Optimal values minimized the MSE.

7. **Random Forest**: Hyperparameters included number of estimators (T) from 100 to 300, splitting criterion (MSE, MAE, Friedman's score), minimum samples to split (mss) from 2 to 10, and use of out-of-bag samples. Optimal values minimized the MSE.

8. **Extra Tree Regressor**: Hyperparameters tuned included number of trees (T) from 100 to 500, and parameters d, mss, and msl with out-of-bag samples. Optimal values minimized the MSE.

Table 1 presents the performance metrics (MSE, MAE, R^2) before and after hyperparameter tuning. Linear models (Ridge, Lasso, ElasticNet) showed weaker predictive power compared to non-linear models (Decision Tree, AdaBoost, Gradient Boosting, Random Forest, Extra Trees).

Extra Trees and Gradient Boosting performed best post-tuning, with the lowest MSE (0.125, 0.127) and MAE (0.273, 0.278), and highest R^2 (0.885, 0.884). Nonetheless, Decision Tree, AdaBoost and Random Forest also performed well.

These results highlight the strengths of boosting techniques in managing complex data relationships and improving prediction accuracy through iterative error correction.

3.2 Community-Level Poverty in 2022

Before analyzing the UBN poverty forecasts for 2022 and comparing them with the 2012 data, it is important to recall that this study focuses on a subset of 953 communities. These communities were selected due to their population sizes exceeding 500 individuals in 2012, which facilitated precise geographical delineation. Beyond this specific sample, the results aim to demonstrate the effectiveness of the proposed methodology in predicting poverty rates at a fine geographical scale.

Figure 1 displays violin plots representing the distribution of community-level UBN poverty headcount ratios, using observed data for 2012 and predicted values for 2022. The results indicate a significant shift in the distribution of UBN poverty between 2012 and 2022 among the studied communities. In particular, the median UBN poverty headcount ratio decreased from 56.7% in 2012 to 41.8% in 2022, a notable improvement of 14.9 percentage points.

The violin plots illustrate key changes in the distribution of community-level UBN poverty headcount ratios. In 2022, the distribution is narrower, particularly at

Algorithm	Pre-tuning			Post-tuning (10-fold cross-validation)			
	MSE	MAE	R^2	MSE	MAE	R^2	Hyperparameters
Ridge	10.772	0.745	-8.870	0.856	0.727	0.215	$\lambda = 20.541$
Lasso	1.092	0.883	-0.000	0.847	0.736	0.224	$\lambda = 0.05$
ElasticNet	1.092	0.883	-0.000	0.853	0.728	0.218	$\lambda = 0.05$ $\alpha = 0.05$
DT	0.224	0.377	0.776	0.132	0.288	0.879	$d = 3$ $mss = 13$ $msl = 4$
ADA	0.169	0.336	0.845	0.135	0.284	0.877	$d = 7$ $\alpha_t = 2.6$ $loss =$ Exponential $T = 185$
GBR	0.128	0.276	0.883	0.127	0.278	0.884	$\gamma_m = 0.07$ $mss = 10$
RF	0.137	0.285	0.874	0.137	0.285	0.875	$T = 225$ *out-of-bag samples* = True *criterion* = Friedman MSE
ET	0.145	0.289	0.860	0.125	0.273	0.885	$d = 14$ *out-of-bag samples* = True

Table 1: Forecast Evaluation and Fine-tuned hyperparameters for the validation set

Note: Not all the hyperparameters described in the preceding paragraphs are included. Those excluded were assigned the default values of the scikit-learn functions, which proved to be the most suitable.

lower poverty levels, suggesting a reduction in the spread of poverty rates. This change indicates a convergence towards lower poverty levels among a greater number of communities. Furthermore, the interquartile range has contracted in 2022, with a more pronounced reduction in the first quartile compared to 2012. This implies a significant improvement at the lower end of the poverty distribution, with fewer communities experiencing extremely high poverty levels. Additionally, the 2022 distribution appears more symmetric, suggesting a more balanced distribution of poverty levels across communities.

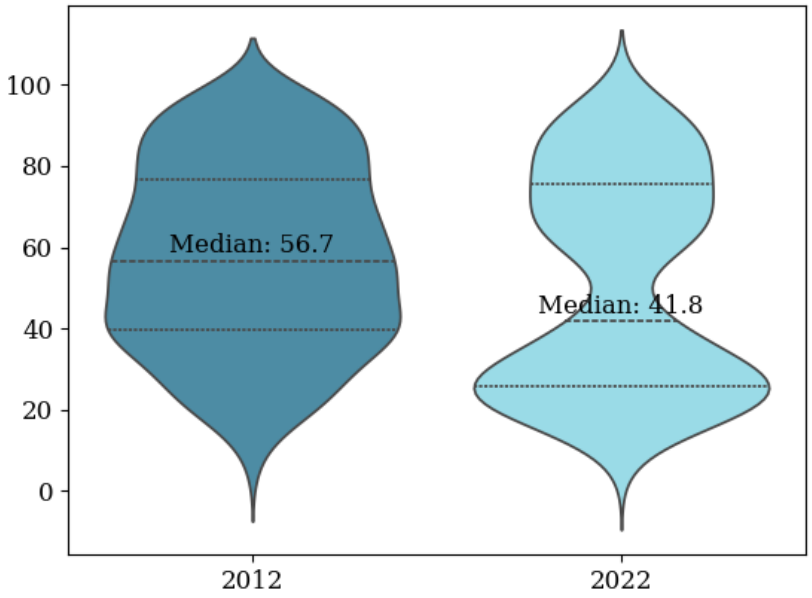


Figure 1: Distribution of UBN Poverty Headcount Ratios, 2012–2022

Figure 2 shows that the majority of the studied communities have a lower UBN poverty headcount ratio in 2022 compared to 2012. Specifically, the forecasts for 2022 indicate that 80% of the communities (763 out of 953) reduced their poverty levels, while 20% (190 communities) experienced an increase in poverty levels relative to 2012.

A key insight from Figure 2 is that the communities with the most significant reductions in poverty headcount ratios between 2012 and 2022 were those with

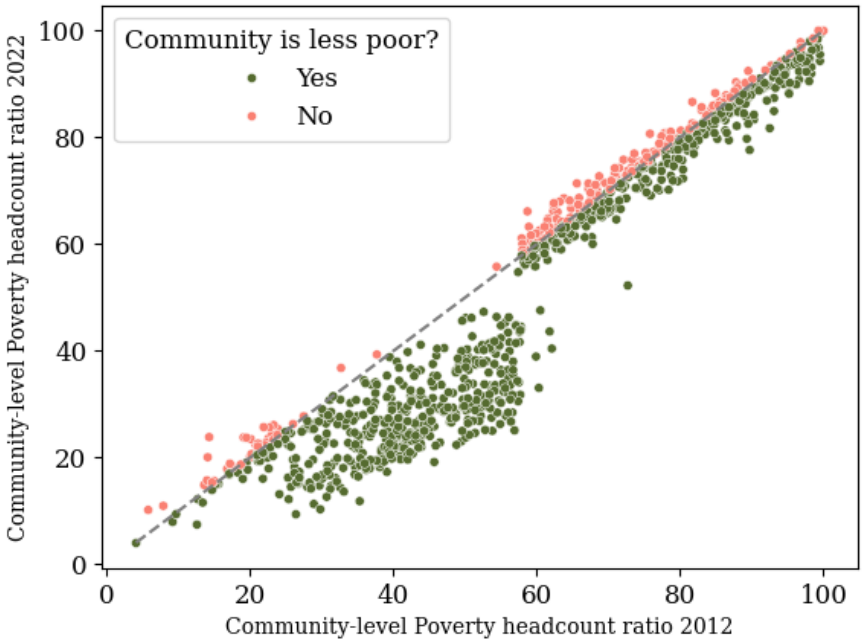


Figure 2: Scatter Plot of UBN Poverty Headcount Ratios, 2012–2022

poverty rates —approximately— below 60% in 2012. Conversely, communities that were more disadvantaged in 2012, with poverty incidences above 60%, exhibited a more moderate reduction in poverty levels.

As one of the study's objectives is to map the 2022 UBN poverty headcount ratios, Figure 3 spatially displays these ratios across the studied communities. This geospatial analysis reveals distinct patterns in the distribution of community-level poverty in Bolivia. The communities with the lowest poverty rates are predominantly located in urban areas, particularly in capital cities and their surrounding regions. For example, La Paz (14.9%), Tarija (17.9%), and Sucre (20.3%) are notable for having the lowest poverty headcount ratios among the capital cities in 2022.

In contrast, the most impoverished communities are primarily found in dispersed rural areas, especially in the highland regions of the departments of La Paz and Oruro, as well as in the department of Cochabamba.[6] There are 69

[6]Departments are the Level 2 Administrative Regions; they are equivalent to states.

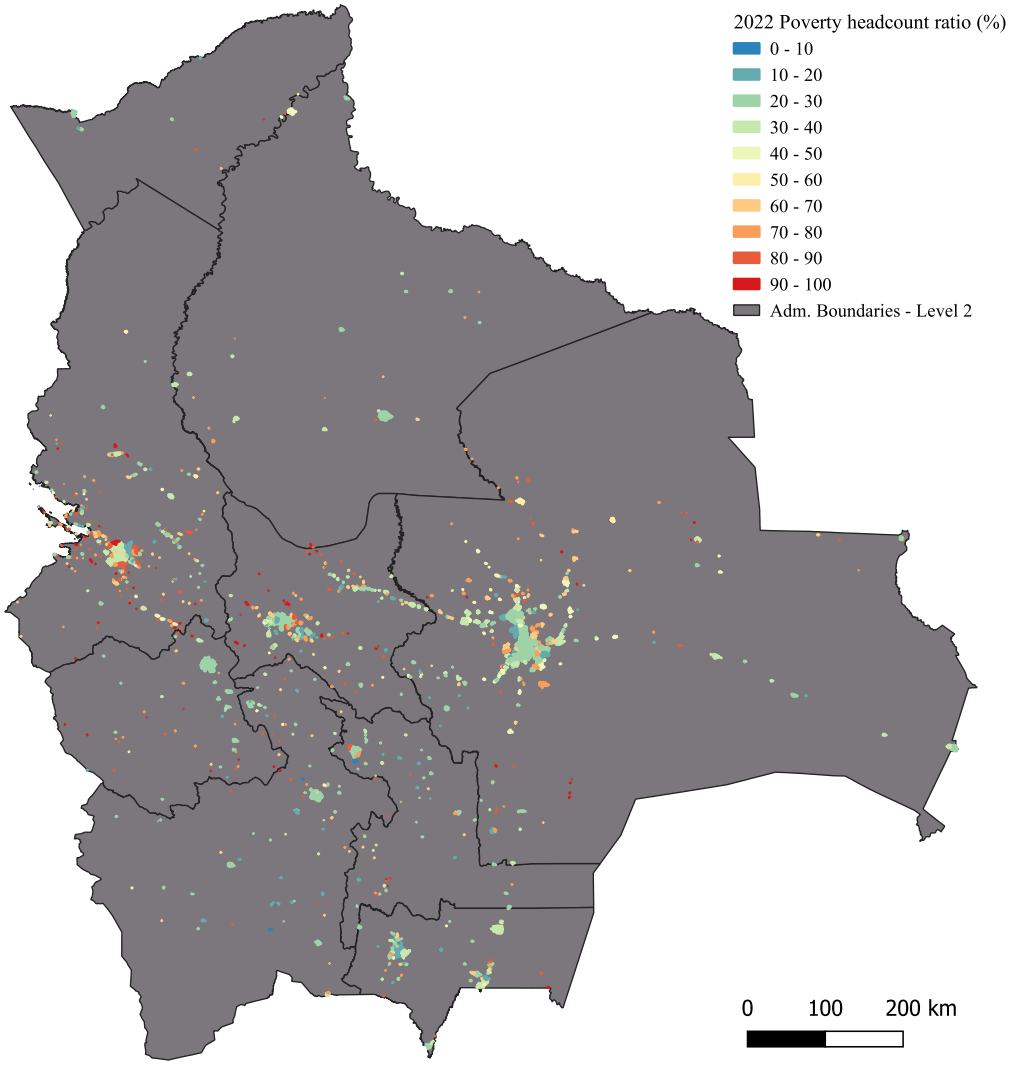


Figure 3: 2022 Community-level Poverty Headcount Ratios

communities where more than 90% of the population lives in poverty, with 50 of these communities located in the departments of La Paz, Oruro, and Cochabamba.

In 2022, the five communities most affected by poverty are Rancho Nuevo (Department of Santa Cruz), Los Yuquis (Department of Santa Cruz), Sacari (Department of Chuquisaca), Chullchungani (Department of Cochabamba), and Ichocollo (Department of Cochabamba), all with a UBN poverty headcount ratio exceeding 98%. These findings underscore the importance of the methodology in identifying the most vulnerable communities.

Although the map provides a cross-sectional view of poverty levels in 2022, a more valuable perspective is the change in poverty headcount ratios between 2012 and 2022. Figure 4 presents a visual representation of these changes across the 953 communities, using a color gradient to indicate the percentage point change in poverty levels. Notably, communities in capital cities have experienced moderate reductions in poverty, typically ranging from -1 to -3 percentage points.

However, several communities, particularly in the department of Santa Cruz and the northeastern part of the department of Cochabamba, have seen substantial decreases in their UBN poverty headcount ratios between 2012 and 2022.

A remarkable example is the community of "Puerto Perez" in the department of La Paz, which achieved a significant reduction of -31.8 percentage points in UBN poverty incidence, decreasing from 56.9% in 2012 to 25.1% in 2022. Conversely, the community of "Huanuni" in the department of Oruro experienced a significant increase in poverty levels, rising by 9.6 percentage points compared to 2012.

The map also highlights that communities near capital cities or economic hubs tend to show more significant progress in poverty reduction. For instance, urban and peri-urban communities around the cities of Santa Cruz and La Paz have performed above average in reducing poverty. For the rest of the study communities, the spatial distribution of changes in UBN poverty headcount ratios reveals some regional and intra-departmental heterogeneity.

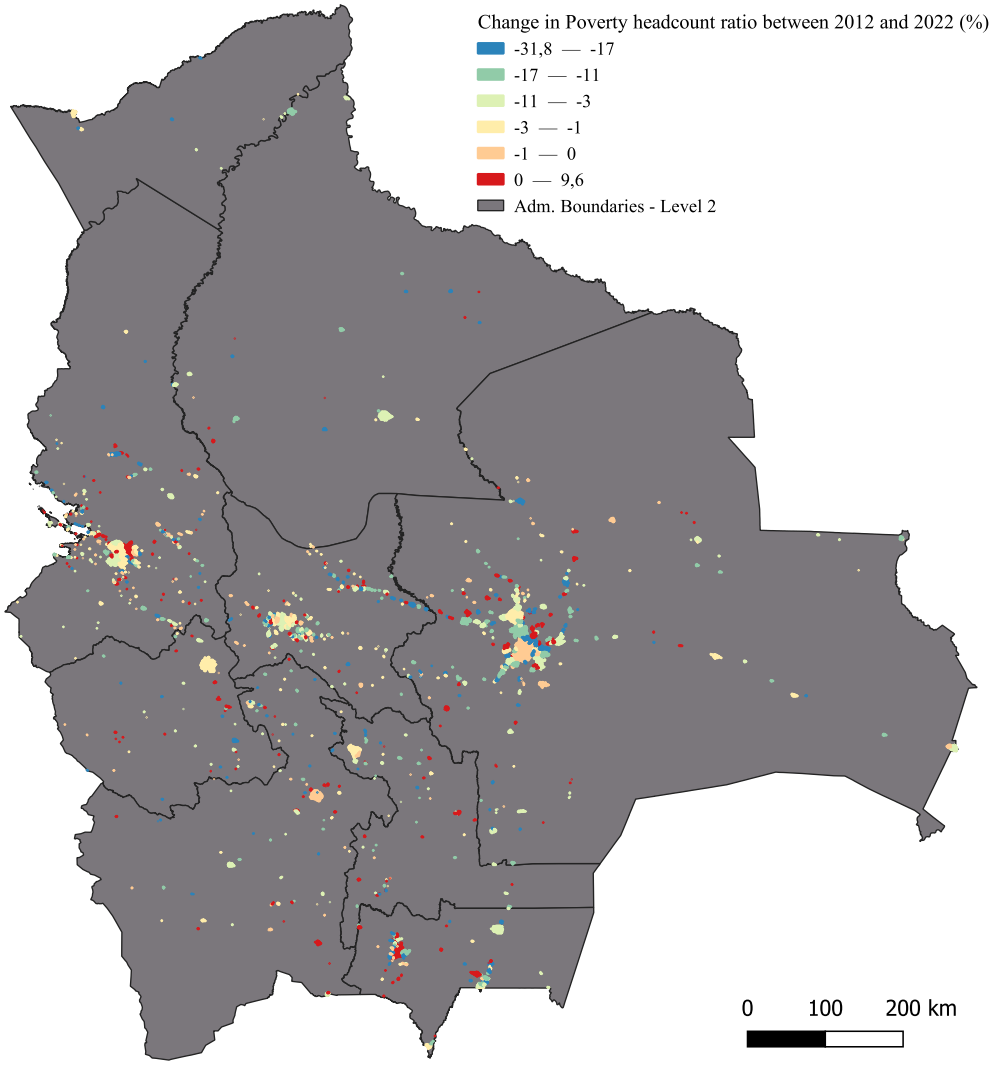


Figure 4: Change in Community-level Poverty Headcount Ratios (2022 vs. 2012)

3.3 Prediction Accuracy

The previous section analyzed the results of the 2022 UBN poverty forecasts for the studied Bolivian communities. We now delve deeper into evaluating the accuracy of these predictions. Figure 5 illustrates the fit between the observed communal poverty headcount ratios for 2012 and their predicted values ($\Upsilon_{i,2022}$), following the procedure outlined in subsection 2.1.5.

As is customary in the forecasting literature, there is an remarkable fit for the training set predictions ($R^2 = 0.954$). Importantly, a significant level of fit is also achieved in the validation set ($R^2 = 0.887$), indicating only a minor discrepancy compared to the training data.

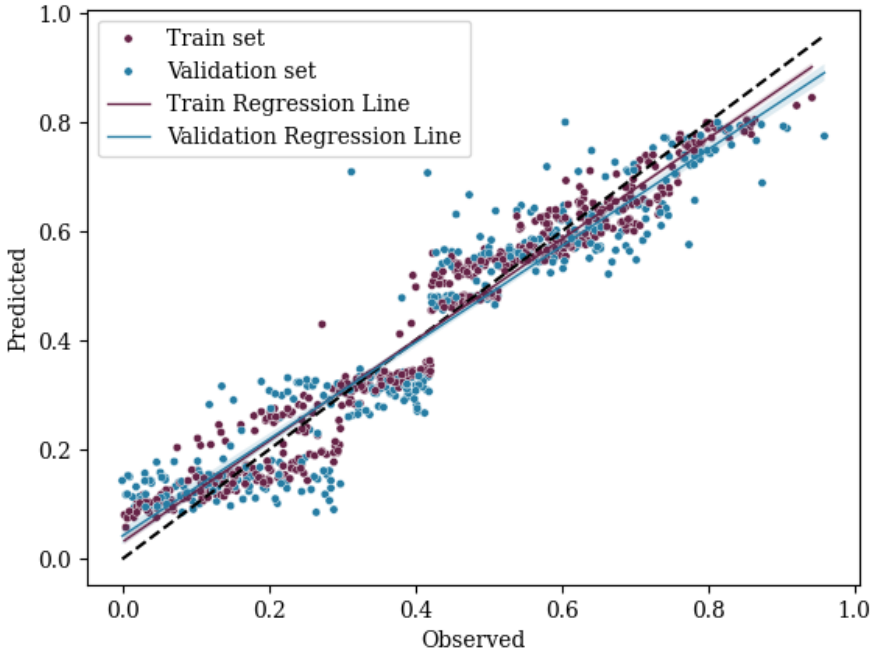


Figure 5: Observed vs. Predicted Data

Note: The dashed black line represents a 45° diagonal line indicating a perfect fit.

The plot also reveals an interesting pattern in the predictions. Generally, for communities with relatively low poverty rates, the forecasts tend to be slightly overestimated, as indicated by the regression lines lying above the 45° diagonal line. Conversely, for communities with higher poverty levels, the forecasts tend to be

slightly underestimated, as shown by the regression lines below the 45° diagonal line in this part of the plot.

In recent years, the use of machine learning and remote sensing data to predict poverty has attracted significant attention in academia. The current study, which applies machine learning techniques to satellite images, Open Street Maps, and other geospatial data, achieved a remarkable R^2 of 0.887 in predicting poverty incidence.[7] This result not only aligns with but also surpasses the performance of several other studies employing similar methodologies.

For example, Chi et al. (2022) reported R^2 values ranging from 0.54 to 0.96 for wealth index predictions using a combination of satellite, phone, and Facebook data. Yeh et al. (2020) achieved R^2 values between 0.75 and 0.83 by applying deep learning to satellite imagery for asset wealth prediction. In contrast, Hersh et al. (2020) reported lower R^2 values between 0.13 and 0.36 when predicting poverty incidence using satellite imagery, indicating variability in performance depending on context and data sources.

Further comparisons include studies by Martinez Jr (2020) and Steele et al. (2017), which explored poverty incidence and wealth indices using a combination of machine learning, deep learning, and hierarchical Bayesian geostatistical models, yielding R^2 ranges of 0.42 to 0.53 and 0.64 to 0.78, respectively. Notably, Engstrom et al. (2017) used linear regression with satellite imagery, achieving R^2 values between 0.60 and 0.61, while Jean et al. (2016) reported R^2 values between 0.55 and 0.75 using deep learning techniques. The study by Blumenstock et al. (2015), which utilized machine learning with mobile phone metadata, stands out with an R^2 of 0.916, demonstrating the high predictive power of mobile data for socio-economic indicators.

Overall, the current study's forecast accuracy is notable, particularly given the comparative analysis, underscoring the strength of integrating diverse geospatial data and machine learning techniques in poverty prediction.

[7] It should be noted that this predictive performance may not necessarily extrapolate to other communities, either within or outside Bolivia.

3.4 Feature Importance

The proposed methodology allows for analyzing the significance of various features in predicting community-level UBN poverty headcount ratios. Table 2 presents the feature importance scores, with a focus on the results from the Extra Trees Regressor (ET), the algorithm with the highest predictive accuracy among those evaluated.

Feature	Ridge	Lasso	ENet	DT	ADA	GBR	RF	ET
Number of Banks	0.0	0.0	0.0	84.3	75.8	78.9	73.7	39.4
Number of Schools	21.2	19.5	20.7	15.2	6.9	15.5	13.7	27.1
Population Center	40.4	35.7	39.5	0.0	0.8	0.4	0.6	11.7
Intermediate City	18.6	15.2	18.0	0.0	0.2	0.0	0.1	4.2
Average Luminosity	6.7	4.4	6.8	0.2	4.5	1.0	3.4	3.8
Area	0.0	0.0	0.0	0.0	3.6	0.5	2.3	2.9
Primary Road Density	9.4	8.4	9.6	0.0	2.8	1.6	1.9	2.6
Number of Hospitals	0.0	0.0	0.0	0.0	0.7	0.2	0.4	2.1
Electricity Grid Density	0.0	0.0	0.0	0.0	2.8	1.0	2.3	1.8
Area of Urban Coverage	0.0	0.0	0.0	0.3	0.9	0.4	0.8	1.6
Area of Urban Settlements	0.0	0.0	0.0	0.0	0.5	0.4	0.4	1.2
Area of Croplands Coverage	3.3	0.0	3.1	0.0	0.7	0.2	0.5	1.0
Capital City	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6

Table 2: Feature Importance

The number of banks is the most significant feature, with an importance score of 39.4. This highlights the crucial role of financial services in reducing poverty. Banks provide access to essential financial resources, such as credit and savings, which are vital for economic growth and improving living standards. Communities with more banks are likely to experience lower poverty levels due to better financial access.

The number of schools is the second most important feature, with a score of 27.1. Education is a fundamental factor in poverty reduction, as it enhances individuals' skills and employment prospects. A higher number of schools indicates better educational infrastructure, contributing to lower poverty rates by equipping individuals with the necessary skills for economic participation.

The population center feature, with an importance score of 11.7, indicates the impact of urbanization on poverty levels. Population centers typically have better access to infrastructure, services, and economic opportunities, leading to improved

living conditions. Similarly, intermediate city, with a score of 4.2, reflects the importance of urban areas as hubs of economic and social activity. These areas provide better access to markets and services, fostering regional development and reducing poverty.

Average luminosity, with a score of 3.8, serves as a proxy for economic activity and infrastructure. Higher luminosity levels suggest greater economic development and access to electricity, which are associated with lower poverty levels.

Other features, such as area, primary road density, number of hospitals, electricity grid density, and areas of urban coverage, urban settlements and croplands coverage, have lower importance scores but still contribute to the model's predictions. These features represent various aspects of infrastructure, public services, and land use, which influence the socioeconomic conditions of a community and, consequently, its poverty levels.

4. Results on Spatial Modeling

Figure 6 shows the spatial connectivity structure of the regions in this study. The network is constructed using a Thiessen polygon diagram. Thiessen polygons connect spatial points such that each location within a polygon is closest to its central point. Nodes represent spatial units, with lines indicating adjacency relationships. The queen contiguity approach, derived from this structure, defines neighbors as units sharing borders or corners. From this structure, one can observe that Bolivia's central and southern regions exhibit higher network density. The spatial connectivity structure of Figure 6 enables the identification of neighbors for each location, which is essential for spatial statistics and autocorrelation analysis.

Figure 7 presents a spatial analysis of poverty in Bolivia using Moran's I statistic and local indicators of spatial association (LISA). The left panel (a) shows a Moran scatterplot with a Moran's I value of 0.27, indicating positive spatial autocorrelation of poverty across regions. The scatterplot is divided into four quadrants representing different spatial relationships: high-high (HH), low-low (LL), low-high (LH), and high-low (HL). The right panel (b) maps these relationships onto Bolivia's

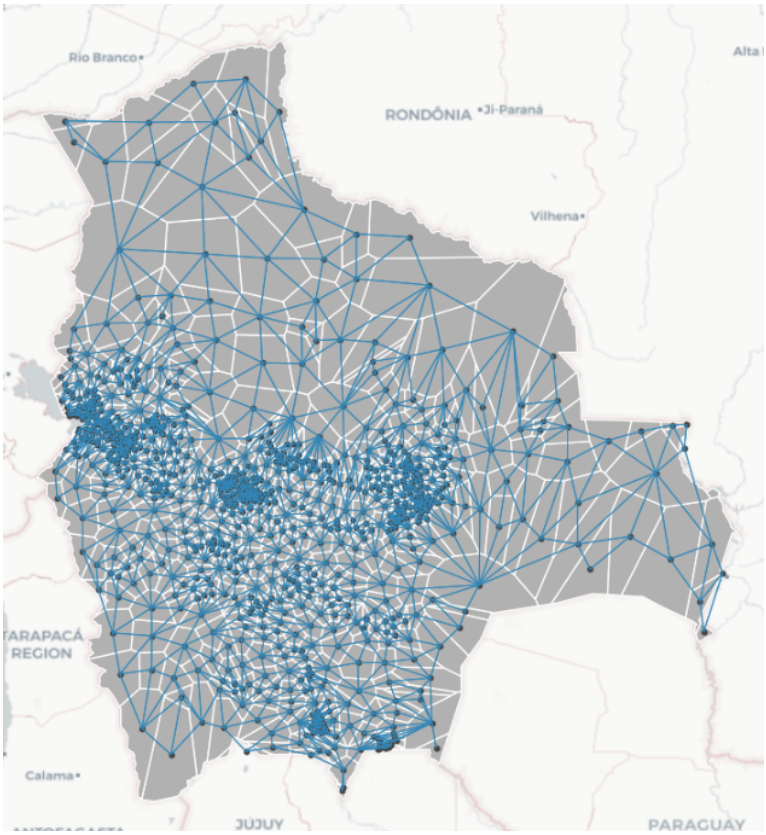


Figure 6: Identification of regional neighbors based on Thiessen polygons

geography, highlighting the location of spatial clusters and outliers. Red areas represent high-poverty clusters (HH), dark blue areas show low-poverty clusters (LL), light blue indicates low-poverty areas surrounded by high-poverty neighbors (LH), and orange represents high-poverty areas surrounded by low-poverty neighbors (HL). The map reveals that poverty in Bolivia has distinct spatial patterns, with clusters of high poverty particularly visible in parts of Santa Cruz, Cochabamba, and the south of La Paz. This analysis helps identify areas for targeted poverty reduction interventions and understand the spatial distribution of poverty in Bolivia in 2012.

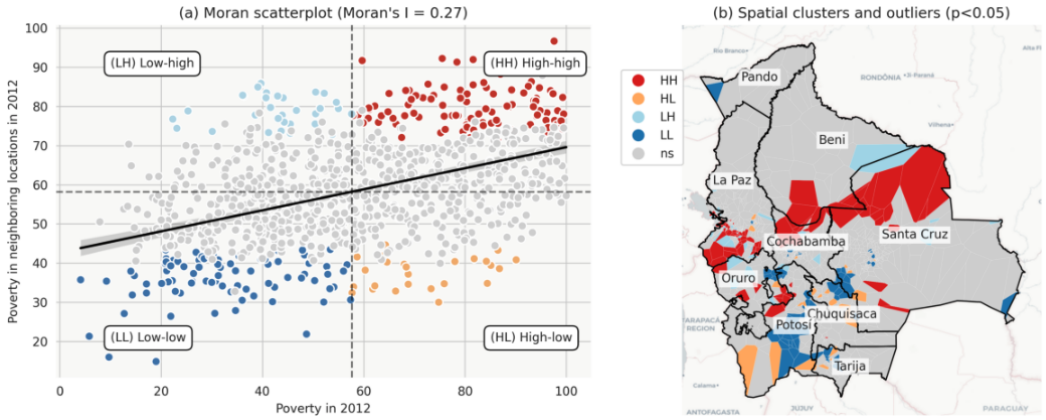


Figure 7: Patterns of spatial dependence in 2012

Figure 8 depicts Bolivia's spatial poverty patterns in 2022 using a Moran scatterplot and a spatial clusters map. The Moran's I value of 0.24 signifies positive spatial autocorrelation, indicating clustering of areas with similar poverty levels. Comparison of 2022 and 2012 poverty patterns reveals subtle yet significant changes over the decade. The Moran's I value decreased from 0.27 to 0.24, suggesting a slight reduction in overall spatial autocorrelation of poverty. While the general poverty cluster distribution remains similar, notable shifts include an expanded high-poverty cluster in Santa Cruz and reduced low-poverty areas around La Paz. The 2022 map shows more scattered and numerous low-high and high-low outliers, suggesting increased local variability in poverty levels. These changes reflect a

complex evolution of Bolivia's poverty landscape, with varying regional progress and challenges.

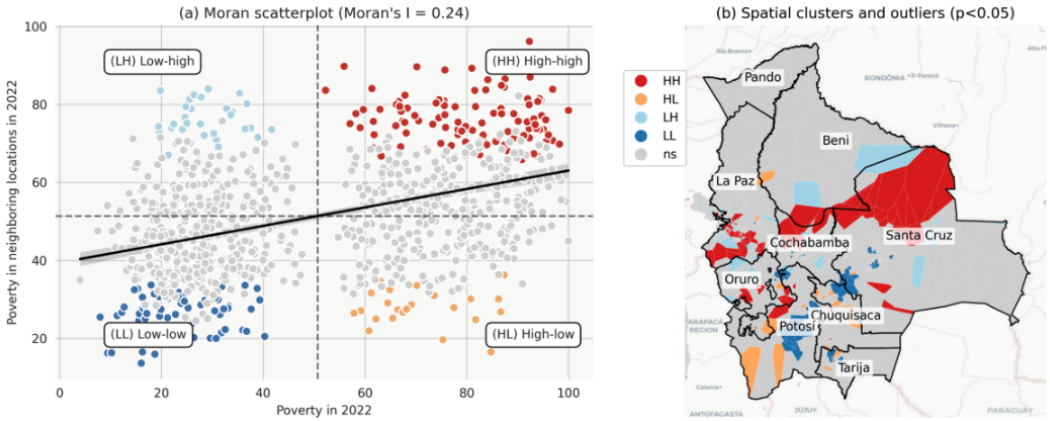


Figure 8: Patterns of spatial dependence in 2022

Figure 9 depicts spatial dependence patterns in Bolivia's poverty changes from 2012 to 2022. The Moran scatterplot exhibits a Moran's I value of 0.04, indicating minimal positive spatial autocorrelation in poverty change. This weak correlation suggests that poverty level changes are not strongly clustered. The spatial dependence map reveals dispersed cluster patterns across Bolivia, with notable High-High areas in parts of Santa Cruz, Tarija, and Potosi. A significant Low-Low cluster is observed in Beni, alongside various outliers throughout the country. Overall, this fragmented pattern indicates geographically diverse outcomes of poverty reduction efforts between 2012 and 2022. Some regions demonstrate improvement, while others experience increased poverty, underscoring the complex and localized nature of Bolivia's economic development during this period.

5. Concluding Remarks

This study set out to predict community-level UBN poverty headcount ratios for Bolivia in 2022, using machine learning algorithms and remote sensing data, in the absence of recent census data. By employing a novel combination of satellite

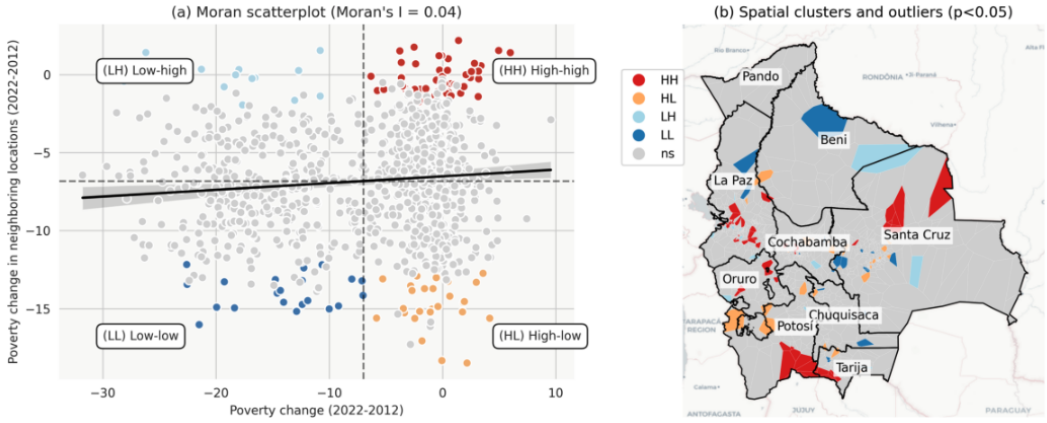


Figure 9: Patterns of spatial dependence over the 2012-2022 period

imagery, Open Street Map data, and advanced machine learning techniques, this research successfully provides high-resolution, community-specific poverty estimates that offer both spatial and temporal insights into poverty dynamics across Bolivia.

The results demonstrate the robustness of machine learning algorithms—particularly the Extra Trees Regressor—in accurately predicting poverty headcount ratios. The importance of features such as the number of banks, schools, and population centers highlights the critical role of infrastructure, education, and access to financial services in shaping poverty outcomes. These findings align with established theoretical perspectives on the multidimensional determinants of poverty, confirming the relevance of these variables as proxies for economic well-being at the community level.

Moreover, the analysis reveals significant spatial disparities in poverty reduction between 2012 and 2022, with urban communities and those near economic centers generally experiencing more pronounced declines in poverty levels compared to rural or remote areas. This spatial heterogeneity underscores the persistent challenges of rural poverty in Bolivia and highlights the need for targeted policies aimed at addressing the specific needs of the most vulnerable communities.

One of the key contributions of this study is the development of a replicable

framework for predicting poverty in contexts where traditional survey data are limited or outdated. The approach not only bridges the data gap between official census years but also offers a scalable model that can be applied to other developing countries facing similar challenges. By utilizing widely accessible remote sensing data and machine learning tools, this research provides a cost-effective and efficient means of producing up-to-date poverty estimates at a high geographical resolution.

In addition to meeting its primary objective of forecasting community-level poverty, this study also advances the literature on spatial poverty analysis by integrating Exploratory Spatial Data Analysis (ESDA). The spatial patterns uncovered in this study, particularly the concentration of poverty in rural and highland regions, offer valuable insights for policymakers. These findings highlight the importance of spatially disaggregated poverty data in designing more effective and equitable poverty alleviation programs.

In conclusion, this study contributes to both the academic discourse on poverty measurement and the practical domain of policy formulation. By leveraging machine learning and remote sensing, the study not only addresses the specific data challenges in Bolivia but also provides a template for future poverty research in other contexts. The insights gained from this analysis can serve as a foundation for evidence-based policy interventions aimed at reducing poverty and improving living standards across communities, particularly in underdeveloped regions where traditional data sources are scarce.

References

Alkire, S. and Foster, J. (2011). Understandings and misunderstandings of multidimensional poverty measurement. *The Journal of Economic Inequality*, 9:289–314.

Atkinson, A. B. (1987). On the measurement of poverty. *Econometrica: Journal of the Econometric Society*, pages 749–764.

Banerjee, A., Duflo, E., and Qian, N. (2020). On the road: Access to transportation infrastructure and economic growth in china. *Journal of Development Economics*, 145:102442.

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

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, 353(6301):753–754.

Bolivar, O. (2022). Roads illuminate development: Using nightlight luminosity to assess the impact of transport infrastructure. Technical report, CAF Development Bank Of Latinamerica.

Bolivar, O. (2023). Bolivia's built-up areas. Accessed on June 15, 2023.

Bourguignon, F. and Chakravarty, S. R. (1999). A family of multidimensional poverty measures. In *Advances in econometrics, income distribution and scientific methodology: Essays in honor of Camilo Dagum*, pages 331–344. Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

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.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from south africa. *American Economic Review*, 101(7):3078–3108.

Engstrom, R., Hersh, J. S., and Newhouse, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. *World Bank Policy Research Working Paper*, (8284).

Feldmeyer, D., Meisch, C., Sauter, H., and Birkmann, J. (2020). Using openstreetmap data and machine learning to generate socio-economic indicators. *ISPRS International Journal of Geo-Information*, 9(9):498.

Foster, A. and Rosenzweig, M. (2010). Microeconomics of technology adoption. *Annu. Rev. Econ.*, 2(1):395–424.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: journal of the econometric society*, pages 761–766.

Galbraith, J. K. (1998). *The affluent society*. Houghton Mifflin Harcourt.

Hersh, J., Martín Rivero, L., Engstrom, R., Mann, M., and Mejía, A. (2020). Mapping income poverty in belize using satellite features and machine learning. *Inter-American Development Bank-Felipe Herrera Library2020*.

INE (2015). Metodología de las necesidades básicas insatisfechas. *Instituto Nacional de Estadística*.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jenkins, S. P. and Lambert, P. J. (1997). Three 'i's of poverty curves, with an analysis of uk poverty trends. *Oxford economic papers*, 49(3):317–327.

Kanbur, R. (1990). *Poverty and development: The human development report and the world development report, 1990*, volume 103. World Bank Publications.

Liu, Y., Zuo, R., and Dong, Y. (2021). Analysis of temporal and spatial characteristics of urban expansion in xiaonan district from 1990 to 2020 using time series landsat imagery. *Remote Sensing*, 13(21):4299.

Martinez Jr, A. M. (2020). Mapping poverty through data integration and artificial intelligence. *UNESCAP: United Nations Economic and Social Commission for Asia and the Pacific*.

Orshansky, M. (1969). How poverty is measured. *Monthly Lab. Rev.*, 92:37.

Piaggesi, S., Gauvin, L., Tizzoni, M., Cattuto, C., Adler, N., Verhulst, S., Young, A., Price, R., Ferres, L., and Panisson, A. (2019). Predicting city poverty using satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–96.

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.

Thon, D. (1979). On measuring poverty. *Review of Income and Wealth*, 25(4):429–439.

Tsui, K.-y. (2002). Multidimensional poverty indices. *Social choice and welfare*, 19:69–93.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583.

Zhou, M., Lu, L., Guo, H., Weng, Q., Cao, S., Zhang, S., and Li, Q. (2021). Urban sprawl and changes in land-use efficiency in the beijing–tianjin–hebei region, china from 2000 to 2020: A spatiotemporal analysis using earth observation data. *Remote Sensing*, 13(15):2850.

A. Extent of Study Communities

The procedure to define the extent of the study communities is as follows:

1. A GIS file with the point location (latitude and longitude) of the communities across the Bolivia's territory was obtained. This data includes a community identifier that facilitates the matching with other databases such us the census data.
2. Only communities with a population greater than 500 inhabitants are kept.
3. A vector file with a grid of 500-meter by 500-meter cells covering the entire Bolivian territory was created.
4. A vector layer with the polygonal delimitation of the communities in Bolivia for the year 2001 was obtained as an additional reference on the potential extent of these geographic units[8].
5. In the vector file of the grid from step 3, cells that do not meet the following requirements were removed:
    a. Being within one of the municipalities containing the 953 selected communities.
    b. Being within one of the polygons from the 2001 community vector file, specifically those corresponding to the 953 selected communities.
6. Raster layers with satellite imagery information were downloaded to map the possible extent or spatial concentration of the population in the 953 selected communities more accurately. These layers include:
    a. *Nighttime luminosity:* A raster file with average luminosity values in 500-meter pixels for the years 2012 and 2022 was used. These files were processed from the "VIIRS Lunar Gap Filled BRDF Nighttime Lights Daily L3 Global 500m" collection from NASA's Land Processes

[8]This vector layer is available on the GeoBolivia portal.

Distributed Active Archive Center (LP DAAC), providing luminosity images corrected for moonlight and atmospheric effects.

b. *Urban and built-up areas:* Raster files from the "MODIS Land Cover Type Yearly Global 500m" product were obtained, offering global maps of land cover usage with annual frequency and a spatial resolution of 500 meters. The MCD12Q1 legend of the International Geosphere-Biosphere Programme (IGBP) was used to define 17 land use categories, including urban and built-up areas.

c. *Global Human Settlement Layer:* A raster file from the Global Human Settlement Layer (GHSL) for 2015 was used, providing information on urbanized areas, population density, and other urban characteristics with a spatial resolution of 1,000 meters.

7. Remaining cells in the vector file grid (step 5) that do not meet at least two of the following three requirements were removed:

   a. Having a luminosity intensity of at least 0.5.

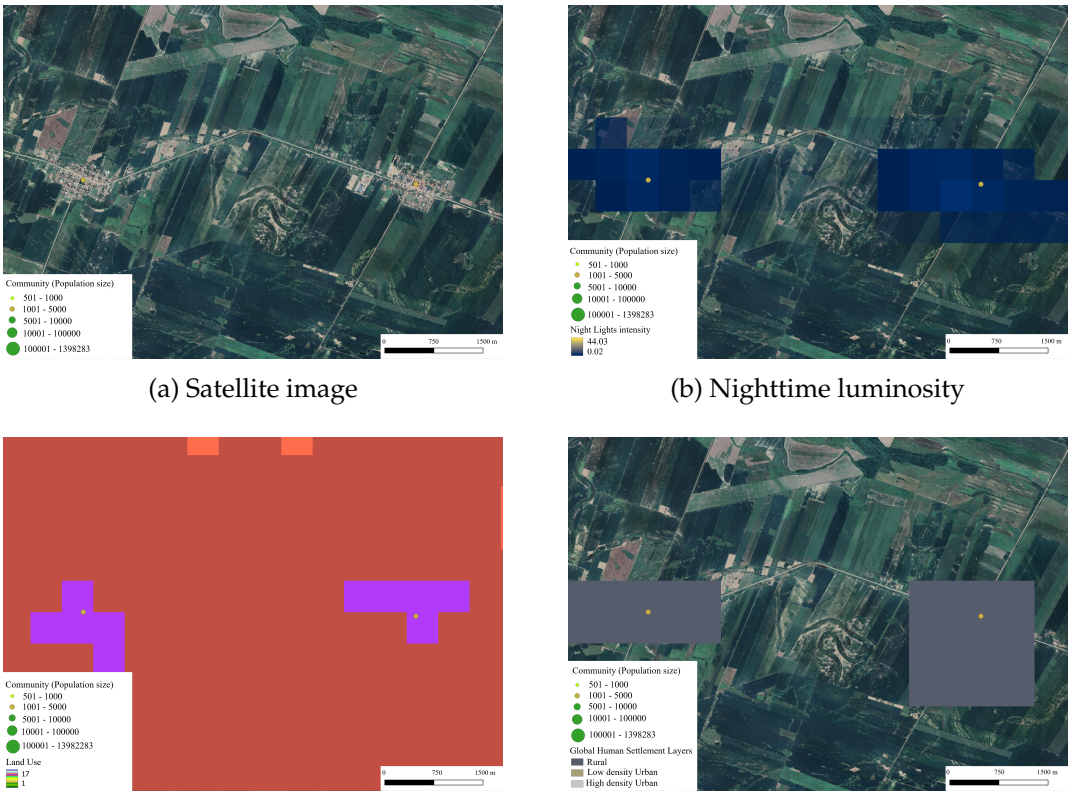
   b. Being classified as "urban and built-up areas" in the "MODIS Land Cover Type Yearly Global 500m" raster layer.

   c. Being classified as "low or high-density urban clusters" in the "Global Human Settlement Layer" raster layer.

   For example, in Figure A.1, some pixels from the nighttime luminosity raster images (A.1-b), land use classified as urban/built-up area (A.1-c, in purple), and human settlements (A.1-d) overlap with areas that would cover 2 communities with populations between 1,000 and 5,000 inhabitants.

8. Each remaining cell was assigned a community identifier according to the municipality in which it is located and its proximity to the georeferenced points of the 2012 census communities and the 2001 community polygons.

9. Finally, for each community, its extent is defined as the area covered by the cells with the corresponding geographic identifier. For example, in Figure A.2,



(a) Satellite image

(b) Nighttime luminosity

(c) Land use

(d) Global Human Settlement Layer

Figure A.1: Raster Layers to Define Community Extent

Source: Own elaboration with data from Google, NASA, MODIS, and European Commission.

the extent for the community Corani Pampa, belonging to the municipality of Colomi in Cochabamba, is shown. This community had a population of 940 inhabitants in 2012. Although some houses and infrastructures in this community are dispersed, the implemented methodology seems to effectively captures these aspects.

It is essential to highlight that the methodology used to define the extents of the 953 study communities captures the concentration of population, infrastructure, and economic activity in these geographic units. This information is fundamental for generating the community characterization variables explained in the next section.

It is important to emphasize that defining the extents of the study communities

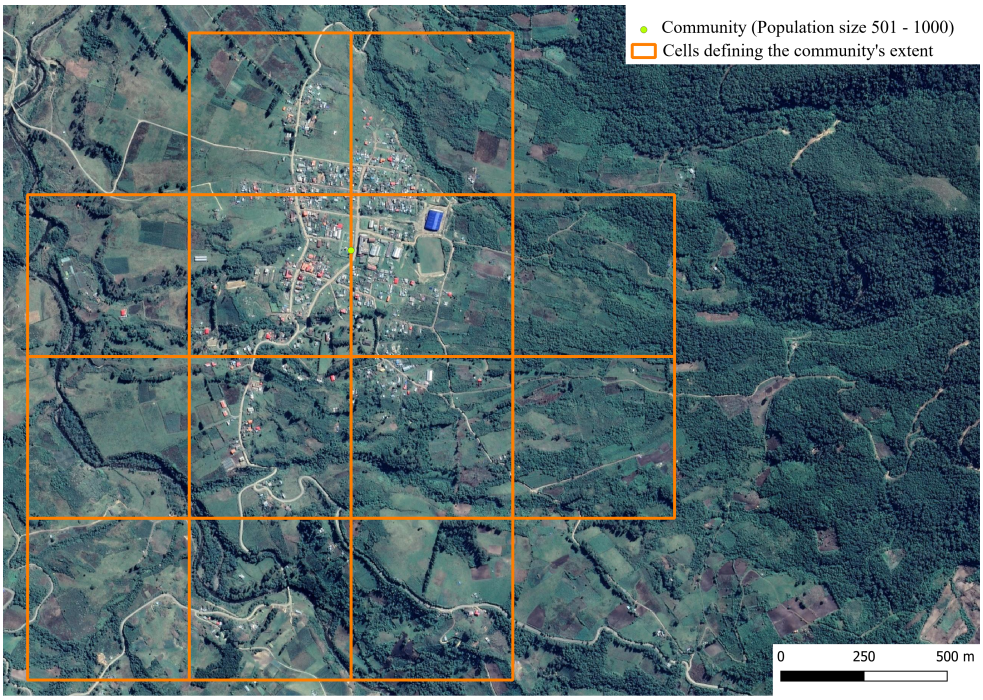


Figure A.2: Extent of the Corani Pampa community

Own elaboration.

does not aim to provide references about the legal boundaries of the communities in Bolivia.