“Mapping Wealth Disparities in Luzon: A Comparative Analysis of Regression Models for Predicting Wealth Index.”

Lambert Famorca [1], Alexcious Norlan C. Decena[2], Remo C. Del Rosario[3], Maria Sheena Shield P. Emocling[4], Tomas Antonio H. Henson[5], Aliyah Jenelle B. Javier[6], Lawrence II T. Miguel[7], Kaizer S. Oman[8], Lyvette Claire Y. Tumaliuan[9]

School of Accountancy, Management, Computing, and Information Studies
Saint Louis University

[1]Computer Science, Saint Louis University- Baguio lpfamorca@slu.edu.ph

[2]Computer Science, Saint Louis University- Baguio 2221089@slu.edu.ph

[3]Computer Science, Saint Louis University- Baguio 2227393@slu.edu.ph

[4]Computer Science, Saint Louis University- Baguio 2220197@slu.edu.ph

[5]Computer Science, Saint Louis University- Baguio 2221277@slu.edu.ph

[6]Computer Science, Saint Louis University- Baguio 2223075@slu.edu.ph

[7]Computer Science, Saint Louis University- Baguio 2226633@slu.edu.ph

[8]Computer Science, Saint Louis University- Baguio 2222613@slu.edu.ph

[9]Computer Science, Saint Louis University- Baguio 2222610@slu.edu.ph

ABSTRACT
The study examines wealth disparities across the Philippines, with analyses focusing on the Luzon region, using comparative regression modeling to predict and analyze regional wealth indices. The research addresses previously identified challenges of identifying key socioeconomic drivers of inequality through the integrated analysis of household survey data, satellite imagery, and geospatial reports. Random Forest and XGBoost models (with R^2 scores of 0.61 and 0.59, respectively) reveal household asset ownership and infrastructure quality as primary determinants of wealth distribution, further highlighting the divide between urban and rural areas. Findings exhibit distinct wealth concentration in Metro Manila with significantly lower indices for regions such as Bicol, CAR, and MIMAROPA. Spatial analysis provides data-driven insights for targeted policy intervention aimed at reducing inequalities. The developed framework supports progressive planning aligned with SDGs by enabling the precise identification of local wealth drivers and spatial patterns, aiding in addressing systemic socioeconomic disparities in developing regions.

KEYWORDS:
Regional inequality, machine learning, rural-urban divide, geospatial analysis, socioeconomic indicators

1. INTRODUCTION

A financial wealth index quantifies socioeconomic status by combining monetary and asset-based indicators. It provides insights into disparities in low- to middle-income regions (Howe, 2009). In the Philippine context, these indices have highlighted severe regional inequalities in financial inclusion and access to resources (Mojica & Mapa, 2016). Nonetheless, traditional approaches rely heavily on fragmented and aggregated datasets (Tingzon et al., 2019), limiting their capacity to capture dynamic factors such as infrastructure quality and geospatial accessibility. In response to these limitations, satellite-based poverty mapping from the Asian Development Bank (Sawada et al., 2021) demonstrates the potential of integrating multidimensional data—including education, healthcare, and infrastructure—to refine wealth assessments. These developments underscore the need for holistic frameworks to address regional financial analysis's systemic deficiencies.

However, despite ongoing efforts, existing wealth indices often neglect the complex interplay of socio-economic, infrastructural, and geospatial determinants, as highlighted by Tingzon et al. (2019) and Rondinelli (1980). Aggregated national data conceals localized disparities, while static indices fail to reflect temporal shifts and non-linear relationships—such as the U-shaped correlation between spatial inequality and development (Pagaduan, 2023). Furthermore, data quality issues (Navarro, 2023), the persistent urban-rural divide, and the exclusion of marginalized populations (Clausen, 2006) compound this challenge. As a result, policies risk being misaligned, leading to inefficient resource allocation. To address these gaps, there is a pressing need for robust, integrated models capable of processing dynamic, multi-source datasets to deliver more accurate assessments.

Accordingly, this study responds to these limitations by reviewing over 50 peer-reviewed articles and reports from 2010 to 2024. The focus is on machine learning models for wealth or poverty prediction, particularly those incorporating socio-economic and geospatial data. Seminal works by Tingzon et al. (2019), Salvador (2024), and Sawada et al. (2021) were selected as key references due to their methodological alignment and relevance to regional contexts. Together, these studies informed the methodological framework and underscored the absence of comparative analyses specific to Luzon.

Building on this foundation, the research investigates how socio-economic and infrastructure-related features influence wealth indices across Luzon, using data from 2017 and incorporating external sources such as Ookla, OpenStreetMap, and VIIRS. The primary objectives are to identify key predictors of wealth disparities, develop and evaluate multiple regression models, and determine which model offers optimal performance. By doing so, the study aims to uncover spatial patterns in wealth distribution by integrating diverse datasets—including nighttime lights (Xu, 2021), vegetation indices (Tang, 2022), and service access indicators. The outcomes are intended to support the design of more targeted, data-informed policy interventions.

The study also produces high-resolution Wealth Index Maps to visualize these findings, illustrating spatial variations in predicted wealth scores across Luzon. These maps are supplemented by a thematic categorization of key predictors grouped under socio-economic assets, infrastructure, environmental indicators, and service accessibility. Each variable's impact is quantified using feature-level analysis. Moreover, four regression models—Random Forest, XGBoost, Gradient Boost, and LightGBM—are evaluated and benchmarked. These analyses aim to identify the most effective predictive models for informing data-driven development strategies.

In line with its objectives, the study answers three core questions: (1) How do the selected regression models perform, and which two are most effective at predicting the wealth index? (2) What are the primary drivers of wealth disparities identified by these models? (3) How does wealth distribution vary across Luzon? These questions emphasize multidimensional, spatially-aware analyses to bridge methodological gaps in traditional wealth assessments.

Ultimately, by incorporating geospatial and infrastructure-related features, the study enhances current wealth index frameworks and contributes a nuanced understanding of regional economic disparities. This approach provides actionable insights for policymakers targeting poverty reduction, infrastructure development, and inclusive growth—aligning with SDG 1 (No Poverty), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 10 (Reduced Inequalities). At the same time, the framework highlights the academic value of integrating non-traditional datasets, such as satellite imagery, into socio-economic modeling.

Nonetheless, the study scope is limited to Luzon and uses 2017 data, restricting temporal generalization. While geospatial datasets like VIIRS offer enhanced granularity, missing subregional metrics and reliance on a single year of data constrain longitudinal analyses. Even so, the framework remains effective for identifying regional disparities and provides a foundation for future, more adaptive studies incorporating newer datasets.

2. LITERATURE REVIEW

Geospatial Data and Remote Sensing in Socio-Economic Analysis

Advancements in geospatial technologies and remote sensing have revolutionized socio-economic analysis. Nighttime light (NTL) data, such as those captured by the Visible Infrared Imaging Radiometer Suite (VIIRS), serve as reliable proxies for economic activity and correlate strongly with GDP and population density (Elvidge et al., 2019). Vegetation indices like the Normalized Difference Index (NDVI) indicate agricultural productivity, a major driver of rural wealth (Jean et al., 2016). Infrastructure data from OpenStreetMap (OSM) reveal access to essential services and road networks, which are vital for economic engagement. Furthermore, internet connectivity metrics from Ookla offer insights into the growing importance of digital infrastructure in economic growth (Szabó et al., 2024). Together, these datasets support a multidimensional approach to wealth assessment.

Machine Learning Applications for Wealth Prediction

Machine learning algorithms, especially ensemble methods such as Random Forest and Gradient Boosting, have shown promise in modeling complex socio-economic relationships (Breiman, 2001). For example, Jean et al. (2016) successfully used convolutional neural networks with satellite imagery to predict poverty levels in Africa. In the Philippines, Tingzon et al. (2019) demonstrated that machine learning models using satellite and crowd-sourced geospatial data can achieve high-resolution poverty mapping. These global and local applications affirm the effectiveness of machine learning in generating detailed and accurate socio-economic forecasts.

Wealth Disparities in the Philippines

Despite advancements in modeling, the Philippines continues to exhibit stark regional wealth disparities. Urban areas such as Metro Manila and Central Luzon consistently outperform rural regions in wealth indicators. Pagaduan (2023) confirmed these inequalities through spatial income analyses using satellite data. Similarly, Sawada et al. (2021) demonstrated the potential of satellite imagery to inform targeted poverty interventions. These findings reinforce the need for localized, data-driven frameworks to inform regional development.

Research Gaps

Specific Focus on Luzon

While national-level studies on wealth mapping exist, research specifically addressing Luzon remains sparse. Luzon's diverse economy—from metropolitan centers to remote rural communities—requires focused study. Pagaduan (2023) and Sawada et al. (2021) provided useful insights, but a Luzon-specific framework addressing its unique socio-economic disparities is still lacking.

Integration of Diverse Data Sources

Existing studies tend to focus on NTL and satellite data, often neglecting internet speed metrics critical to evaluating digital infrastructure. For instance, while Tingzon et al. (2019) employed geospatial data, internet connectivity was not emphasized as a feature. Incorporating such data could significantly enhance the accuracy of wealth prediction models for Luzon.

Comparative Evaluation of Machine Learning Models

Most prior research, including those by Tingzon et al. (2019) and Salvador (2024), applied individual machine-learning models for poverty or wealth prediction. However, direct comparative studies evaluating Random Forest, HistGradientBoosting, XGBoost, and LightGBM in the Philippine context, especially in Luzon, are absent. Addressing this gap, the current study evaluates and compares these four models for their effectiveness in wealth index prediction.

Identification of Key Predictors

Another gap lies in the underuse of interpretability tools like SHAP in identifying key predictors. While NTL has been established as important (Tingzon et al., 2019), other features, such as road accessibility and digital connectivity, require deeper analysis to inform policy development.

Theoretical and Conceptual Framework

This study adopts a conceptual framework that integrates spatial, socio-economic, and institutional determinants of wealth. The Relative Wealth Index (RWI) serves as the dependent variable, influenced by spatial metrics such as road density, urban proximity, and digital connectivity, as outlined in Spatial Economics Theory (Smith, 2007). Socio-economic indicators, including education, health, and employment, are grounded in the New Economic Geography Theory (Krugman, 1991). Institutional factors such as governance and policy are treated as moderating variables, consistent with Institutional Economic Theory (North, 1990).

Analytically, machine learning models—including Random Forest Regressor—and geospatial analysis techniques were applied to datasets from Ookla, OpenStreetMap (OSM), Earth Observation Group (EOG), and the Demographic and Health Survey (DHS). These tools enabled the identification of non-linear, multidimensional relationships across Luzon's regional landscape.

Figure 1

Conceptual Framework

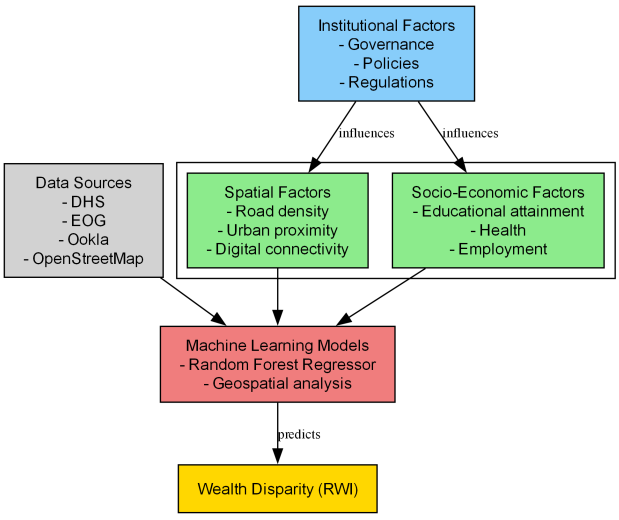


Figure 1 illustrates the conceptual framework, which builds upon prior research utilizing machine learning approaches and satellite imagery to identify economic outcomes and predict wealth disparities in the Philippines. Integrating spatial, socio-economic, and institutional variables enhances the understanding of wealth distribution patterns and supports evidence-based policy recommendations.

Table 1 outlines key studies that informed the conceptual and methodological foundation of this research. Each entry highlights the methodological approach and primary contribution of the cited work. Howe et al. (2009) established the use of wealth indices as proxies for socioeconomic status, forming a basis for subsequent measurement techniques. Tingzon et al. (2019) demonstrated the application of machine learning and satellite data for poverty prediction, achieving strong model performance. Sawada et al. (2021) advanced the field by producing granular, satellite-based poverty maps, while Pagaduan (2023) emphasized the importance of spatial analysis in revealing income inequality across regions. Collectively, these studies guided the selection of features, modeling strategies, and evaluation metrics in the present analysis.

Table 1

Key Studies and Their Contributions

Study	Methodology	Key Contribution
Howe	Wealth index construction	Established wealth indices for socio-economic status (Howe et al., 2009)
Tingzon et al.	Machine learning, satellite imagery	Mapped poverty with R^2 of 0.63 (Tingzon et al., 2019)
Sawada et al.	Satellite-based poverty mapping	Provided granular poverty maps (Sawada et al., 2021)
Pagaduan	Spatial income inequality analysis	Highlighted regional disparities (Pagaduan, 2023)

3. METHODOLOGY

The methodology section outlines the sequential steps followed in the data collection, preparation, model development, and evaluation processes.

3.1 Data Collection

Table 2 shows the group's initial dataset. Furthermore, the wealth index dataset used in this research to train different experimental models was obtained from the 2017 Philippine Demographic and Health Survey (DHS). This nationally representative survey provides comprehensive information on health, demographics, and household characteristics. Using a 2.4 km tile size, a subset of 1,247 GPS-tagged sample locations was extracted, each accompanied by 22 attributes, including essential geospatial and socio-economic indicators such as wealth index, latitude, and longitude coordinates. Each of these 1,247 cluster points corresponds to a group of individual respondents and was stored in GeoJSON format, which is widely used for spatial data due to its ease of use and compatibility with mapping and geospatial tools.

Table 2

Data Collection Overview

Demographic and Health Survey (DHS)	**OpenStreetMap (OSM)**
Year Collected: 2017 Number of Features: 36	Year Collected: 2017 Number of Features: 78
Ookla	**Visible Infrared Imaging Radiometer Suite (VIIRS)**
Year Collected: 2017 Number of Features: 10	Year Collected: 2017 Number of Features: 5

The 2017 dataset is the only available year for model training and analysis, as the subsequent iteration of comparable national-level data was conducted in 2022, which was used to test the prediction capabilities of the models.

To enrich the dataset by authenticating credentials on EOG (Earth Observation Group), incorporating spatial and infrastructure-related features from multiple external sources, and aligning them with the original DHS clusters based on geographic coordinates. The data sources used are as follows:

Geospatial enrichment (EOG (Earth Observation Group) credential authentication, multiple externally sourced spatial and infrastructure-related features aligned with DHS clusters based on geographic coordinates) included data from:

- Ookla: Provided data on internet speed within the vicinity of the DHS cluster points. Features include average download/upload speeds in kilobits per second (kbps), representing the level of digital infrastructure.
- OpenStreetMap (OSM): Contributed location-based data on various services and infrastructure. Extracted counts and proximity metrics for facilities such as banks, police stations, fast food restaurants, and supermarkets offer insight into urban development and public access to services.
- Visible Infrared Imaging Radiometer Suite (VIIRS): Supplied night-time light radiance data, which is widely used as a proxy for economic activity and electrification in spatial analysis.

These data sources were merged with the original DHS sample at the group level. This resulted in an expanded dataset with 93 new external features per cluster point, yielding a primary dataset of 1,247 rows by 115 columns. Furthermore, 14 additional person-level features from DHS were retrieved, including indicators such as house quality, access to electricity, and ownership of durable goods.

These features were merged into the DHS dataset, having expanded the dataset with 93 new external features per cluster point, resulting in a dataset of 1,247 rows and 115 columns. Fourteen additional individual-level indicators were aggregated to the cluster level.

3.2 Data Preparation

The dataset underwent initial preprocessing by removing the "Wealth Index" column, which represents an average of individual-level wealth scores per tile. Only numeric features were retained to allow for meaningful aggregation, while the "DHSCLUST" identifier was preserved for group-level analysis. Aggregation by cluster followed, producing group-level features aligned with the main data structure.

During integration, missing values were imputed using feature-wise means to ensure completeness. A variance thresholding technique filtered out near-zero variance features, reducing dimensionality and noise. Principal Component Analysis (PCA) was then applied to transform the data into 10 principal components while preserving key patterns. Finally, correlation-based clustering grouped components with similar variance profiles using the elbow method, resulting in the clustered PCA dataset.

3.3 Model Development

Four regression algorithms were developed to evaluate model performance: Random Forest, Gradient Boosting, XGBoost, and LightGBM. Each model was configured with optimized hyperparameters, including 100 estimators and a learning rate of 0.1, along with a fixed random state (42) for reproducibility. This setup ensured uniform conditions for comparison. The focus was to identify the top two performing models that best predict the wealth index for further analysis and interpretation.

3.4 Model Evaluation

Model performance was assessed using four metrics: R^2, MAE, MSE, and RMSE. R^2 measured explained variance, MAE calculated average absolute error, MSE penalized larger deviations, and RMSE reported errors in original units. Cross-validation involves computing the mean and standard deviation of each metric to allow consistent model comparisons.

SHAP analysis was used to interpret model behavior and identify the most influential features. A subsequent visualization phase included wealth index maps and clustered bar charts from the top two models, offering spatial insights into wealth distribution across Luzon for the year 2022.

4. RESULTS AND DISCUSSION

5.1 Data Overview and Preprocessing Outcomes

The analysis used an integrated dataset focused on Luzon, derived from the 2017 Philippine Demographic and Health Survey (DHS). This included 1,247 GPS-tagged clusters with 22 group-level indicators, among them the Relative Wealth Index (RWI). To enrich the dataset, additional features were sourced from Ookla (internet speed, 10 indicators), OpenStreetMap (infrastructure metrics, 78 indicators), and VIIRS (nighttime light radiance, five indicators), resulting in 93 external features.

In addition, 14 person-level features (e.g., electricity access and durable goods ownership) were aggregated to the cluster level, yielding a total of 129 features per cluster. During preprocessing, only numeric features were retained for aggregation; missing values were imputed using mean substitution, and sparse categorical classes were consolidated. Variance thresholding was applied to eliminate low-informative features, after which principal component analysis (PCA) reduced the feature space to 10 components that captured essential socio-economic and geospatial variation.

Table 3 provides a summary of the final retained features per dataset after preprocessing, showing that 11 features remained from DHS, 40 from OpenStreetMap, 10 from Ookla, and five from VIIRS. This overview highlights the dominance of geospatial and infrastructure-related variables in the final modeling dataset.

Table 3

Preprocessed Data Overview

Demographic and Health Survey (DHS)	**OpenStreetMap (OSM)**
Remaining Number of Features: 11	Remaining Number of Features: 40
Ookla	**Visible Infrared Imaging Radiometer Suite (VIIRS)**
Remaining Number of Features: 10	Remaining Number of Features: 5

For SHAP (Shapley Additive exPlanations) analysis, household assets identified, such as refrigerators and televisions, as the most significant predictors of RWI, alongside infrastructure metrics like internet speed and night-time light radiance—a proxy for economic activity. These findings underscored the model's emphasis on urbanization, infrastructure access, and rural productivity.

5.2 Model Performance and Evaluation

To assess predictive performance, four models—Random Forest, XGBoost, LightGBM, and Gradient Boosting—were evaluated using three data treatments: baseline, PCA-transformed, and clustered PCA. Random Forest emerged as the most robust performer, maintaining high accuracy (R^2: 0.605–0.610) with only minor degradation under PCA (~8% relative drop). XGBoost followed as a close second with raw features (R^2: 0.590), demonstrating better resilience to PCA (11% drop) and partial recovery with clustered PCA (R^2: 0.5773) compared to other boosters. LightGBM and Gradient Boosting performed comparably to XGBoost using baseline-processed features (R^2: 0.581 and 0.584, respectively) but suffered more significant PCA-induced declines (~10–12% drops). Gradient Boosting exhibited the largest sensitivity to PCA (12% R^2 reduction), while clustered PCA only partially mitigated this effect. Despite these variations, the models' relative rankings remained consistent (Random Forest > XGBoost > LightGBM > Gradient Boosting) across all treatments, underscoring Random Forest's superiority for this task regardless of feature transformation. These rankings remained consistent across all treatments, confirming Random Forest's overall robustness. Table 4 presents the detailed results.

Table 4

Models' Performance

Model	R^2 Score	MAE	MSE	RMSE	Data Treatment	Notes
Random Forest	0.61	0.087	0.012	0.109	Processed	Most Robust Performer: Maintains high accuracy across all treatments (R^2: 0.605–0.610), with only minor degradation under PCA (~8% relative drop).
	0.56	0.091	0.014	0.116	PCA	
	0.6051	0.0869	0.0121	0.1099	Clustered PCA	
Gradient Boosting	0.584	0.090	0.013	0.113	Processed	Suffers the largest performance drop with PCA (12% R^2 relative drop). Clustered PCA partially mitigates (R^2 0.5710 vs. 0.517 for PCA) but still underperforms baseline-processed features (R^2 0.584).
	0.517	0.096	0.015	0.121	PCA	
	0.571	0.091	0.013	0.115	Clustered PCA	
XGBoost	0.59	0.089	0.013	0.112	Processed	A close second to RF with raw features (R^2 0.590). While PCA

	0.523	0.095	0.015	0.121	PCA	reduces performance (11% R^2 relative drop), clustered PCA shows better recovery (R^2 0.5773) than other boosters.
	0.577	0.090	0.013	0.114	Clustered PCA	
LightGBM	0.581	0.090	0.013	0.113	Processed	Comparable to XGBoost with raw features (R^2 0.581) but similar PCA sensitivity (~10% relative drop). Clustered PCA offers marginal improvement (R^2 0.5703).
	0.521	0.096	0.015	0.121	PCA	
	0.570	0.091	0.013	0.115	Clustered PCA	

Note: Data Treatment Legend:

- Processed: Dataset that underwent comprehensive data cleaning, preserving original feature space, serving as the baseline dataset.
- PCA: Applied Principal Component Analysis to reduce the dimensionality of the processed dataset.
- Clustered PCA: further grouped components using correlation-based clustering.

5.3 Identification of Key Drivers

SHAP analysis confirmed the dominant role of household asset ownership and infrastructure access in predicting RWI as identified by the top models (Random Forest & XGBoost).

The Random Forest Regression model analysis identifies household asset ownership and infrastructure quality as the dominant drivers of wealth disparities in Luzon, though their prominence varies across methodologies. Processed feature importance highlights refrigerator ownership (0.646), television ownership (0.160), and floor type (0.066) as the strongest predictors, reflecting their direct role in traditional asset-based wealth indices. These variables align with PCA3 (Household Assets), where refrigerator, television, and floor type load prominently (loadings: 0.25–0.24), and Clustered PCA (C29), which groups these indicators thematically. This consistency underscores their universal relevance as micro-level wealth proxies. However, PCA1 (General Infrastructure and Network Activity), the most influential component (55% importance), aggregates infrastructure variables like road density and nighttime light radiance, which rank lower individually in processed features (road_count: 0.003, avg_rad_median: 0.0057). This divergence arises because PCA1 captures their collective impact on regional economic vibrancy, while processed features prioritize standalone asset indicators. Similarly, PCA2 (Service Proximity) highlights urban service access but fragments across clusters due to context-dependent urban-rural dynamics—urban households prioritize services over assets, whereas rural areas exhibit inverse trends.

The XGBoost model highlights household asset ownership and infrastructure quality as Luzon's primary wealth drivers, with strong cross-method consistency. Similar to the top key drivers affecting the prediction result of the Random Forest Regression Model, the XGBoost

model revealed refrigerator ownership (0.568), television ownership (0.122), and floor type (0.039) as the dominant processed features, aligning with PCA3 (Household Assets)—where these variables load prominently (refrigerator: 0.255, television: 0.251, floor: 0.238)—and Clustered PCA (C29), which groups them thematically. This tripartite coherence underscores their universal role as micro-level wealth proxies, consistent with methodologies like the DHS. Infrastructure variables, however, exhibit methodological divergence: nighttime light radiance (avg_rad_median: 0.049) ranks third in processed features but anchors PCA1 (General Infrastructure), which synthesizes road density, internet connectivity, and luminosity into a composite regional development indicator (55% importance). While infrastructure variables lack a dedicated cluster, their distributed influence across clusters reflects systemic spatial inequalities. Service proximity metrics, like fast-food access (0.0098 in processed features), are thematically grouped in Clustered PCA (C3) and PCA2 but rank lower individually, suggesting their contextual relevance in urban hubs. Notably, car/truck ownership (0.021) appears as a standalone predictor and within PCA3, illustrating its dual role as a wealth marker and component of broader asset portfolios. These findings emphasize that Luzon's disparities stem from material household assets and infrastructure-driven spatial divides, necessitating policies that bridge rural-urban gaps in durable goods and connectivity (Siatan et al., 2024; Nicoletti et al., 2022).

The Random Forest Regression model and XGBoost identify household asset ownership (refrigerators and televisions) and infrastructure quality (nighttime lights and road density) as Luzon's primary wealth drivers. This consistency validates traditional wealth indices (ADB, 2021) like the DHS, where durable goods and housing quality are core metrics. However, the models diverge in emphasis: RFR assigns higher standalone importance to household assets (refrigerator: 0.646 vs. XGBoost: 0.568), reflecting its sensitivity to categorical variables, while XGBoost prioritizes infrastructure (avg_rad_median: 0.049 vs. RFR: 0.0057), leveraging its gradient-boosting architecture to better capture nonlinear interactions. Both recognize the systemic role of infrastructure through PCA1 (55% importance). Service proximity metrics reveal urban-rural divides: Random Forest Regression fragments these variables across clusters, underscoring spatial disparities, while XGBoost groups them in Clustered PCA (C3), emphasizing urban commercialization's role. Car/truck ownership illustrates methodological nuances—it functions as a standalone predictor in RFR but also integrates into XGBoost's broader asset portfolios (PCA3). These differences stem from model architectures: RFR's bagging approach spreads importance across correlated variables, whereas XGBoost boosting sharpens focus on high-impact predictors.

Policy implications remain consistent—dual strategies addressing material needs (rural asset access) and infrastructure gaps (roads, electrification) are critical. However, XGBoost's stronger infrastructure focus suggests prioritizing connectivity could accelerate equity gains. Researchers should consider hybrid frameworks, combining RFR's systemic insights with XGBoost's granular precision, to design robust, context-sensitive interventions for Luzon's spatial and socioeconomic inequalities.

5.4 Spatial Analysis of Wealth Disparities

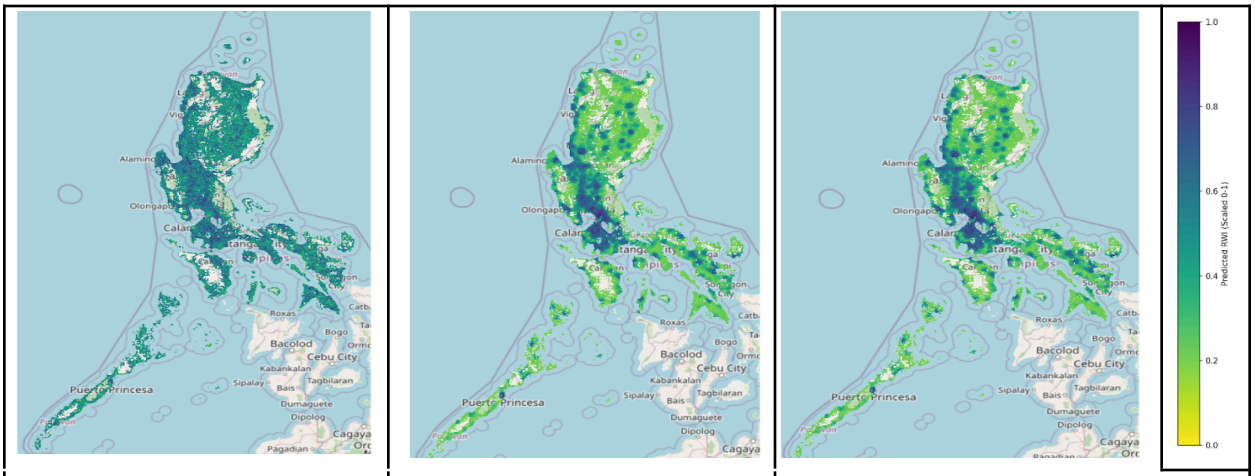
This subsection examines the spatial variation of predicted wealth across Luzon based on the results of the best-performing Random Forest model and the comparison with XGBoost.

Using both models' predictions, the spatial distribution of wealth index values was mapped, and regional patterns were evaluated using bar graphs, heatmaps, and clustered PCA outputs.

Figures 2 and 3 illustrate regional wealth distribution as predicted by Random Forest. Wealth is highly concentrated in urban regions such as Metro Manila and adjacent areas, while rural regions like MIMAROPA, Bicol, and Cagayan Valley exhibit lower predicted wealth.

Figure 2

Random Forest predicted wealth maps: (a) Processed, (b) PCA, (c) Clustered PCA.

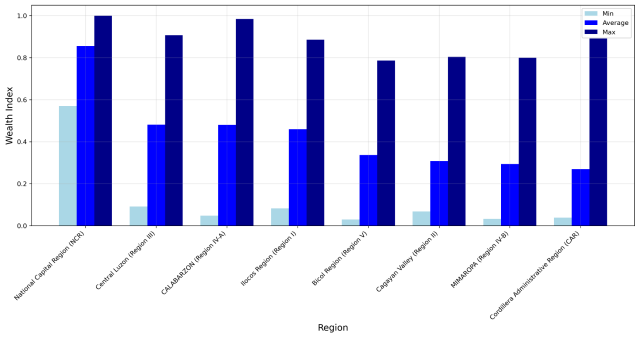


a. Random Forest Mapping Processed Data | *b. Random Forest Mapping PCA Data* | *c. Random Forest Mapping Clustered PC Data*

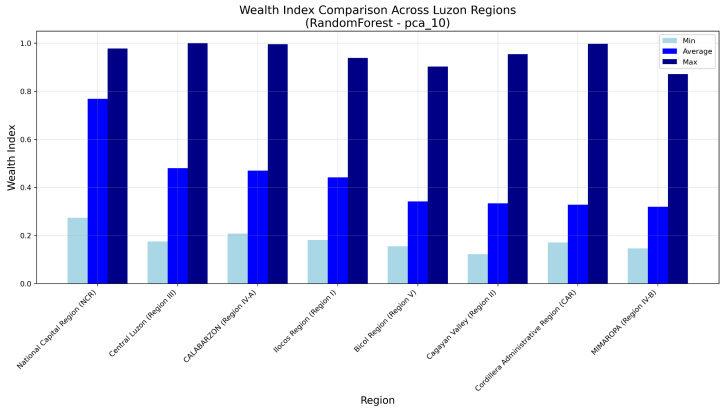
Figure 3

Random Forest grouped bar charts by region: (a) Processed, (b) PCA, (c) Clustered PCA.

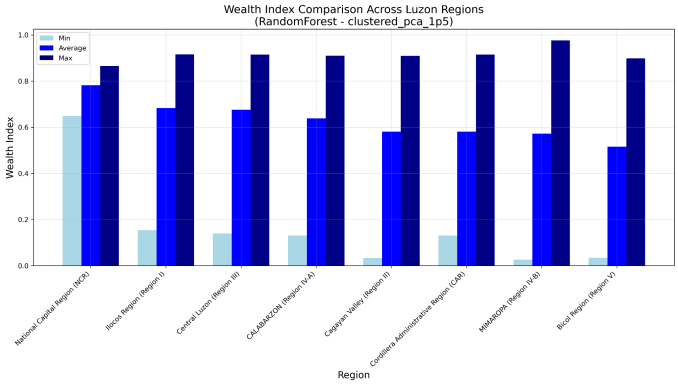
a. Random Forest Grouped Bar Chart Processed Data



b. Random Forest Grouped Bar Chart PCA Data



c. Random Forest Grouped Bar Chart PCA Clustered Data



Note: Figures 2. a and 3.b illustrate the wealth index distribution across Luzon using Random Forest predictions on processed data. While most regions show similar averages, notable gaps exist between the richest and poorest areas. NCR leads with the highest average wealth index, reflecting its developed economy. MIMAROPA, Cagayan, Bicol, and particularly CAR, recorded lower values, highlighting regional poverty. The map visually confirms this trend, with darker urban centers (e.g., Metro Manila) contrasting against lighter rural zones.

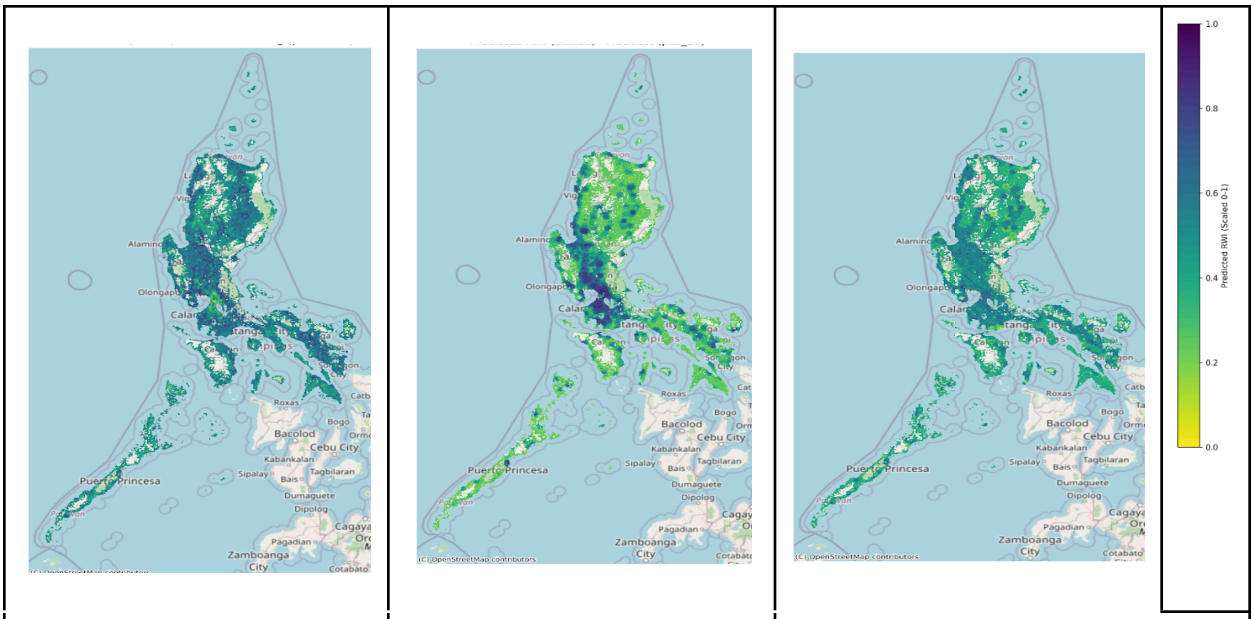
Figures 2a and 3b, using PCA-transformed data, show NCR maintaining the highest average index, with Central Luzon, CALABARZON, and CAR reaching the highest peak values. Ilocos also ranks high in average scores, indicating more equitable wealth distribution in these areas.

Figure 2c reveals strong urban-rural contrasts, with high RWI values clustered around Metro Manila and Southern Luzon. Rural provinces like Bicol, Cordillera, and Cagayan Valley show lower wealth levels. Figure 3c supports this by showing MIMAROPA's high peaks, while NCR, Ilocos, and Central Luzon retain the highest average wealth indices. In contrast, Bicol and Cagayan reflect consistently low predictions.

Figures 4 and 5 present comparable outputs from XGBoost. The model also identifies NCR as the most affluent region, followed by CALABARZON and Central Luzon. In contrast, Bicol, CAR, and Cagayan Valley consistently rank among the least wealthy.

Figure 4

XGBoost predicted wealth maps: (a) Processed, (b) PCA, (c) Clustered PCA.



a. XGBoost Mapping Processed Data

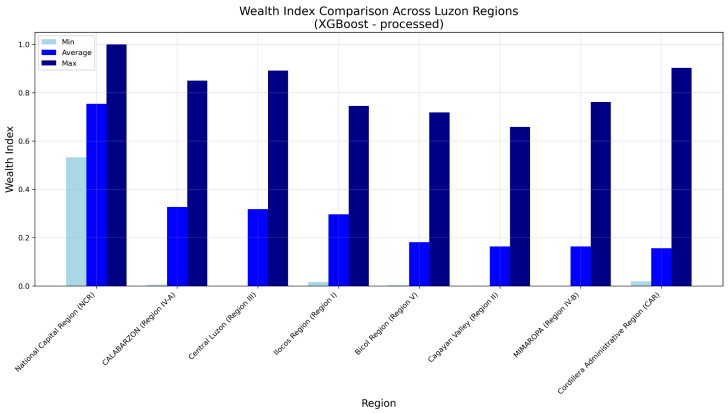
b. XGBoost Mapping PCA Data

c. XGBoost Mapping Clustered PC Data

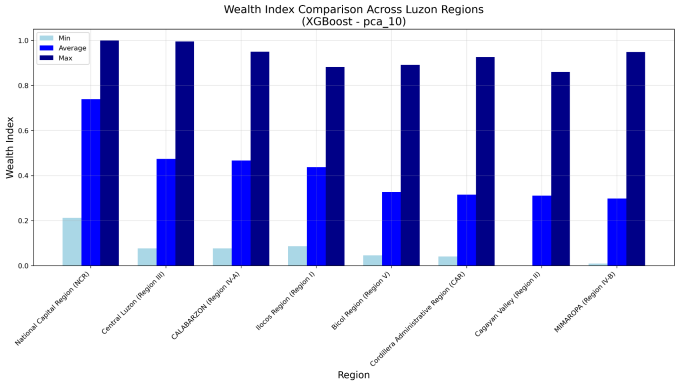
Figure 5

XGBoost grouped bar charts by region: (a) Processed, (b) PCA, (c) Clustered PCA.

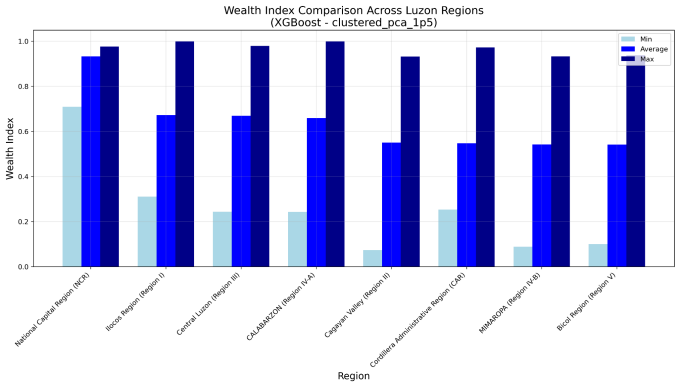
a. XGBoost Grouped Bar Chart Processed Data



b. XGBoost Grouped Bar Chart PCA Data



c. XGBoost Grouped Bar Chart PCA Clustered Data



Note: Figure 4a shows NCR with the highest average wealth index, followed by Central Luzon and CALABARZON. In contrast, Cagayan Valley, Bicol, MIMAROPA, and CAR rank lower. Peak values mirror this trend, with NCR and Central Luzon leading, while Cagayan Valley consistently scores the lowest. Figure 5a confirms this disparity, highlighting concentrated wealth in urban regions and lower levels in rural areas. Compared to Random Forest, XGBoost produces less extreme contrasts.

Figures 4b and 5b, using PCA data, maintain NCR and Central Luzon at the top in both average and peak scores, followed by MIMAROPA and CALABARZON. The map reflects these regional variations, with darker shades in urban centers and lighter tones in less developed areas, suggesting PCA effectively preserved spatial wealth patterns.

Figures 4c and 5c, based on clustered PCA, show Ilocos and CALABARZON at peak wealth values, with NCR and CAR closely trailing. NCR remains the leader in average wealth. Ilocos, Central Luzon, and CALABARZON appear moderately wealthy, while Cagayan Valley, CAR, MIMAROPA, and Bicol remain low. The map affirms this spatial gradient with clear urban-rural contrasts.

While both models agree on regional rankings, Random Forest captures sharper wealth divides, especially between urban and rural areas. XGBoost produces smoother gradients, suggesting less sensitivity to extreme variation.

Similarities and Differences between the two Models

Both Random Forest and XGBoost consistently identify Metro Manila (NCR) as the wealthiest region, with CALABARZON and Central Luzon as moderately wealthy. Conversely, Bicol, Cagayan Valley, and MIMAROPA rank among the least wealthy. Differences emerge in how each model presents disparities. Random Forest, especially in clustered PCA outputs, highlights more pronounced wealth gaps, sharply distinguishing urban regions like Metro Manila, Ilocos, and Central Luzon from rural areas such as Cagayan Valley, CAR, and Bicol. The sharp contrasts between dark (wealthy) and light (poor) areas on the maps and the clear ranking gaps in the bar chart suggest that Random Forest captures a more segmented, unequal landscape that also aligns with the model's sensitivity to categorical variables and its emphasis on standalone asset indicators.

XGBoost, on the other hand, smooths the extremes. While still identifying NCR as the top region, it elevates regions such as Ilocos, Central Luzon, and CALABARZON in clustered PCA outputs. This results in reduced visual and statistical gaps between the richest and poorest regions. The model's gradient-boosting architecture contributes to a more balanced depiction of regional wealth, with transitions appearing more gradual and less fragmented compared to Random Forest. Both models agree on the ranking hierarchy but differ in how they portray the severity of regional inequalities.

How do wealth disparities vary across Luzon's regions?

The results show that wealth disparities across Luzon are largely shaped by a strong urban-rural divide, as also observed in Clausen's study (2006). Metro Manila emerges as the consistent wealth leader due to its infrastructure and economic concentration. Surrounding regions like CALABARZON, CAR, and Central Luzon follow, benefiting from industrial growth and proximity to NCR. In contrast, Bicol, MIMAROPA, and Cagayan Valley rank lower, with their limited infrastructure, economic activity, and employment contributing to weaker wealth indices. Both models reinforce this gradient: Random Forest emphasizes the contrast more distinctly, while XGBoost provides a smoother distribution that still supports the broader narrative. Ilocos, for example, gains prominence in XGBoost outputs, suggesting that some rural regions demonstrate stronger-than-expected wealth signals under models that account for nonlinear feature interactions. These consistent patterns across models underscore the need for region-targeted policies aimed at addressing infrastructure gaps and improving asset access in underdeveloped areas.

5. CONCLUSION

Summary of Findings

This study investigated regional wealth disparities in Luzon using a multidimensional framework integrating socioeconomic, infrastructural, and geospatial indicators. Among the four regression models evaluated, Random Forest demonstrated the highest predictive performance

($R^2 = 0.61$), followed closely by XGBoost ($R^2 = 0.59$). Both models remained consistent even after applying dimensionality reduction using Principal Component Analysis (PCA).

Predicted wealth distributions revealed a distinct urban-rural divide. Metro Manila (NCR) consistently registered the highest scores, followed by CALABARZON and Central Luzon. In contrast, regions such as Bicol, CAR, MIMAROPA, and Cagayan Valley scored lower. SHAP analysis identified household assets—particularly refrigerators and televisions—as strong wealth predictors, alongside infrastructure proxies like nighttime light intensity and road density. PCA results further highlighted the significance of infrastructure, with the first component accounting for 55% of the variance.

These findings align with observed patterns in the 2017 Philippine Demographic and Health Survey (DHS), reinforcing that the country's economic center remains concentrated in urbanized regions while more remote and less developed areas face persistent disadvantages. Notably, the model outputs are consistent with official poverty statistics released by the Philippine Statistics Authority (2023), which identified the same regions—NCR, Central Luzon, and CALABARZON—as having lower poverty incidence, while regions like MIMAROPA and Bicol exhibited higher rates.

Implications

The analysis provides evidence-based insights to guide interventions addressing regional wealth disparities. The identification of household asset ownership and infrastructure quality as key predictors addresses earlier limitations in fragmented, static indices, as noted by Tingzon et al. (2019), and supports Howe's (2009) emphasis on multidimensional wealth measurement. Infrastructure's strong influence on the wealth index, as quantified in PCA, also extends the spatial inequality discussions raised by Rondinelli (1980).

The results reaffirm Clausen's (2006) findings on the urban-rural divide and support Navarro's (2023) recommendation for subnationally disaggregated data. SHAP analysis emphasized digital infrastructure, particularly 4G density, as a wealth driver, strengthening Szabó et al. 's (2024) framework on the digital divide in the Philippines.

By integrating geospatial data sources like DHS, VIIRS, OpenStreetMap, and Ookla, the study operationalizes a scalable, multidimensional model that can inform targeted development initiatives and support local decision-makers in addressing wealth inequality. These insights complement ongoing discussions around SDG 10 (Reduced Inequalities) by promoting regional equity through data-driven resource allocation.

Limitations

Despite strategies conducted within the study, several limitations were identified throughout. One of the main constraints is the models' reliance on data captured in 2017, which hinders further dynamic analysis. Confined geographical analysis exclusively on Luzon impeded the generalizability of findings to other areas in the Philippines with differing socioeconomic contexts. Although the study used diverse data sources, limited detail in some subregional metrics may obscure local disparities, affecting data quality. Complex interactions in ensemble

models hinder full interpretability, which is common in machine learning applied in socioeconomic contexts. Additionally, the reliance on available datasets may exclude key cultural and social dimensions of wealth, which highlights persistent difficulties in capturing the full scope of multidimensional inequality.

Future Research

Future work should incorporate the latest DHS data (2022–2025) to enable longitudinal analysis and assess post-pandemic trends. Incorporating climate indicators (e.g., precipitation, temperature), and finer spatial grids can improve granularity and capture environmental effects on regional wealth.

Advancing modeling techniques—such as convolutional neural networks (CNNs) or causal inference approaches like Double Machine Learning—can better capture nonlinear interactions and the impact of external shocks. Real-time data from mobile networks or social media platforms can further improve responsiveness to evolving socioeconomic conditions.

Developing an open-source web-based wealth index prediction tool, validated through partnerships with local government units, would support actionable planning. Ensuring ethical practices, including GDPR-compliant privacy measures and inclusive data collection, remains essential to address biases and ensure equity. Preliminary provincial-level results show the model scales effectively, enabling more granular policy design and planning within Luzon and beyond.

REFERENCES

Allen, F., & Gale, D. (2007). Understanding financial crises. Oxford University Press.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Clausen, A. (2006). Disparities of poverty and wealth in the Philippines. An analysis of policy effect(iveness). https://kups.ub.uni-koeln.de/2002/

Fujita, M., Krugman, P., & Venables, A. J. (1999). The spatial economy: Cities, regions, and international trade. MIT Press.

Howe, L. D. (2009). The wealth index as a measure of socio-economic position. https://doi.org/10.17037/pubs.00768490

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Krugman, P. (1991). Increasing returns and economic geography. Journal of Political Economy, 99(3), 483–499. https://doi.org/10.1086/261763

Mojica, M. B., & Mapa, C. D. (2016). An Index of Financial Inclusion in the Philippines: Construction and analysis. The Philippine Statistician, 66(No. 1 (2017)), 59–74. https://api.semanticscholar.org/CorpusID:210175717

Navarro, A. (2023). Subnational infrastructure development and internal migration in the Philippines. https://doi.org/10.62986/dp2023.20

North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge University Press.

Pagaduan, J. A. (2023). Spatial Income Inequality, Convergence, and Regional Development in a Lower Middle‐Income Country: Satellite Evidence from the Philippines. The Developing Economies, 61(2), 117–154. https://doi.org/10.1111/deve.12354

Philippine Statistics Authority. (2023). *Poverty Incidence Among Population, by Region*. Retrieved from https://seis.pids.gov.ph/index.php?id=250&r=databank%2Ffrontend%2Fdataset%2Fview

Rondinelli, D. A. (1980). Regional Disparities and investment allocation policies in the Philippines: Spatial dimensions of poverty in a developing country. Canadian Journal of Development Studies/Revue Canadienne D Études Du Développement, 1(2), 262–287. https://doi.org/10.1080/02255189.1980.9669816

Salvador, E. L. (2024). Use of boosting Algorithms in Household-Level Poverty Measurement: A machine learning approach to predict and classify household wealth quintiles in the Philippines. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.13061

Sawada, Y., Addawe, E., Martinez, A., Albert, J., Bulan, N., Durante, R., Fenz, K., Hoffer, M., Martillan, M., Mitterling, T., Claire, T., Mapa, D., Bautista, R., Esquivias, M., Astrologo, C., Guillen, W., De Costo, S., Balamban, B., Buenaventura, P. S., . . . Zaini, B. (2021). Mapping the spatial distribution of poverty using satellite imagery in the Philippines. https://doi.org/10.22617/tcs210076-2

Smith, J. P. (n.d.). The Impact of Socioeconomic Status on Health over the Life-Course. The Journal of Human Resources, XLII(4), 739–764. https://doi.org/10.3368/jhr.xlii.4.739

Szabó, Roland. (2024). Overcoming the digital divide: A conceptual framework. Journal of Infrastructure, Policy and Development. 8. 10082. 10.24294/jipd10082.

Tang, G., Tian, R., & Wu, B. (2022). An overview of clustering methods in the financial world. Advances in Economics, Business and Management Research/Advances in Economics, Business and Management Research. https://doi.org/10.2991/aebmr.k.220307.084

Tingzon, I., Orden, A., Go, K. T., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M., & Kim, D. (2019). Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W19, 425–431. https://doi.org/10.5194/isprs-archives-xlii-4-w19-425-2019

World Commission on Environment and Development. (1987). Our common future. Oxford University Press.

Xu, J., Song, J., Li, B., Liu, D., & Cao, X. (2021). Combining night time lights in prediction of poverty incidence at the county level. Applied Geography, 135, 102552. https://doi.org/10.1016/j.apgeog.2021.102552

Zhao, M., Zhou, Y., Li, X., Cao, W., He, C., Yu, B., Li, X., Elvidge, C. D., Cheng, W., & Zhou, C. (2019). Applications of satellite Remote sensing of nighttime Light observations: advances, challenges, and perspectives. *Remote Sensing*, *11*(17), 1971. https://doi.org/10.3390/rs11171971