Geospatial Analysis of Socioeconomic Disparities in Peru through Predictive Modeling of the Human Development Index

Lindell D. Vilca Mamani
Faculty of Statistical and Computer Engineering
Universidad Nacional del Altiplano - Puno
livilca@est.unap.edu.pe

Abstract

Peru exhibits marked territorial disparities in human development, with districts showing HDI above 0.7 while others remain below 0.3. The geographic concentration of vulnerabilities suggests structural spatial patterns requiring multidimensional analysis. A total of 1,874 Peruvian districts were analyzed using HDI, extreme poverty, food vulnerability, altitude, and population density data. Spatial correlation techniques (Moran's Index), weighted composite index construction, typological classification, and predictive modeling through Random Forest and Gradient Boosting with cross-validation were applied. Significant spatial autocorrelation ($I = 0.631$, $p < 0.001$) and strong correlations between HDI and food vulnerability ($\rho = -0.919$) were identified. Eight territorial typologies were defined, highlighting that 2.1% of districts present a "High-Altitude Vulnerable" profile (HDI = 0.282). The Gradient Boosting model achieved $R^2 = 0.854$ (MAE = 0.039), identifying food vulnerability as the dominant predictor (91.3%). This study demonstrates the feasibility of integrating spatial analysis and machine learning to characterize socioeconomic disparities and predict HDI, providing operational tools for targeting public policies in territories with triple vulnerability.

Keywords: Human development, spatial analysis, territorial inequalities, machine learning, vulnerability index, Peru.

1 Introduction

Peru, with its megadiverse geography and marked socioeconomic disparities, faces significant challenges in human development. Despite macroeconomic progress in recent decades, territorial gaps persist and manifest heterogeneously across its 1,874 districts distributed in 24 departments. The Human Development Index (HDI), as a multidimensional indicator integrating health, education, and income aspects, reveals a fragmented reality: while some districts exceed the 0.7 threshold characteristic of high development, others remain below 0.3, evidencing very low human development conditions [1].

This territorial heterogeneity is not random. Spatial development patterns reveal strong differentiation between natural regions: coastal districts present an average HDI of 0.574, contrasting markedly with the Andean highlands (0.369) and the Amazon rainforest (0.400). These differences replicate across other socioeconomic indicators: extreme poverty affects 18.4% of the population in the highlands, while on the coast it barely reaches 3.5%. This geographic concentration of vulnerabilities suggests the existence of structural mechanisms that perpetuate territorial inequalities [2, 3].

Evidence of highly significant positive spatial autocorrelation (Moran's Index $I = 0.631$, $p < 0.001$ for HDI) confirms that development conditions are not independently distributed but exhibit territorial clustering patterns. Districts with low human develop-

ment tend to be surrounded by districts with similar characteristics, generating vulnerability *hotspots* that concentrate multiple simultaneous deprivations [4, 5]. This phenomenon is particularly pronounced in the south-central highlands, where low HDI, high extreme poverty (average 36.7%), and elevated food vulnerability (average index 0.677) converge in high-altitude territories located above 3,500 meters above sea level.

At the national level, the magnitude of the problem is considerable: 320 districts (17.1%) present high vulnerability according to a composite index integrating inverted HDI, extreme poverty, and food vulnerability, while 12 additional districts (0.6%) exhibit very high vulnerability. The departments of Cajamarca, Ayacucho, and Apurímac concentrate the highest averages of territorial vulnerability (CVI > 0.53), with extreme cases such as the Curgos district in La Libertad (CVI $= 0.942$, HDI $= 0.117$, extreme poverty $= 86.7\%$) evidencing persistent humanitarian crisis situations [6].

Correlations between indicators reveal the multidimensional nature of deprivations. HDI presents strong negative correlations with total poverty ($\rho = -0.785$, $p < 0.001$), extreme poverty ($\rho = -0.754$, $p < 0.001$), and food vulnerability ($\rho = -0.919$, $p < 0.001$), suggesting that deprivations do not occur in isolation but as part of territorial vulnerability syndromes [7]. Altitude, although with a weaker correlation ($\rho = -0.385$, $p < 0.001$), emerges as a geographic factor associated with lower development levels, reflecting difficulties in accessing basic services and economic opportunities in high-altitude areas [8, 9].

Despite the abundance of statistical information, important limitations persist in traditional analytical approaches. Previous studies have tended to analyze indicators univariately or at the departmental aggregation level, rendering intraregional heterogeneity invisible and hindering the identification of complex territorial patterns [10, 11]. Likewise, the lack of robust predictive tools limits decision-makers' capacity to estimate development conditions in contexts where primary data are scarce or outdated. The application of *machine learning* techniques to model HDI based on more accessible indicators represents an unexplored opportunity to improve public policy targeting [12, 13, 14].

This study addresses these gaps through a multivariate geospatial approach integrating spatial correlation analysis, composite index construction, territorial typologies, and predictive modeling. The specific objectives are: (1) characterize statistical relationships between HDI, poverty, food vulnerability, and geographic factors; (2) construct a Composite Vulnerability Index (CVI) synthesizing multiple dimensions of deprivation; (3) identify territorial typologies capturing the heterogeneity of district contexts; and (4) develop a predictive HDI model based on socioeconomic and geographic indicators using machine learning techniques.

The main contribution of this work lies in methodological integration and the generation of operational tools for decision-making. The developed predictive model (Gradient Boosting) achieves a determination coefficient $R^2 = 0.854$ on test data, with a mean absolute error of 0.039 points on the HDI scale. Feature importance analysis reveals that food vulnerability is the dominant predictor (91.3% relative importance), followed by population density (2.5%) and altitude (2.2%), providing quantitative evidence on the mechanisms underlying territorial disparities [16].

Additionally, the proposal of a territory typology with eight differentiated categories—from "Urban Developed Coastal" (1.7% of districts, average HDI 0.775) to "High-Altitude Vulnerable" (2.1% of districts, average HDI 0.282)—offers a conceptual and operational framework for designing interventions adapted to the specificities of each territorial context. The "Mixed/Transition" category, grouping 45.8% of districts with intermediate indicators, represents an opportunity space for preventive policies that avoid the consolidation of poverty traps [15].

2 Materials and Methods

The study used secondary data from the enriched UBIGEO Peru dataset [17], which in-

tegrates information from INEI, UNDP, and MIDIS. The final sample included 1,874 districts with complete information for the 2018-2020 period.

The dependent variable in the predictive model was the 2019 HDI (0-1 scale). Independent variables included socioeconomic indicators: total poverty (%), extreme poverty (%), food vulnerability (0-1), and geographic characteristics: altitude (meters above sea level) and population density (inhabitants/km^2). Additionally, spatial coordinates (latitude, longitude) and administrative categories (department, natural region) were considered.

Descriptive statistics were calculated and normality was assessed using the Shapiro-Wilk test ($\alpha = 0.05$). Bivariate correlations were analyzed with Spearman's coefficient (ρ) given the non-normality of distributions. ANOVA compared means between natural regions (Coast, Highlands, Rainforest), verifying variance homogeneity with Levene's test.

Spatial autocorrelation was evaluated using the global Moran's Index (I):

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

using an 8-nearest neighbors matrix and significance by permutations ($n = 999$) [4].

A Composite Vulnerability Index (CVI) was constructed through Min-Max normalization and weighted aggregation: $CVI = 0.40(1 - HDI_{norm}) + 0.35P_{ext,norm} + 0.25V_{food,norm}$, where subscripts indicate normalized variables. Weights were assigned considering measurement robustness and policy relevance [7]. Districts were classified into five levels according to CVI.

The territorial typology employed rule-based classification with multiple thresholds of HDI, poverty, food vulnerability, altitude, and location. Eight categories were defined: Urban Developed Coastal, Urban Intermediate, High-Altitude Vulnerable, Rural with Food Insecurity, Extreme Poverty, Intermediate Highlands, Rural Rainforest, and Mixed/Transition.

2.1 Predictive Modeling

Data were divided into training (80%) and test (20%) with stratified sampling. Variables were standardized through Z-score transformation. Three algorithms were evaluated: Ridge Regression with L2 regularization [18], Random Forest [19] ($n_{estimators} = 100$, $max\_depth = 10$), and Gradient Boosting [20] ($n_{estimators} = 100$, $max\_depth = 5$, $learning\_rate = 0.1$).

Performance was evaluated with R^2, MAE, RMSE, and 5-fold cross-validation. Feature importance was calculated through mean impurity reduction. Sensitivity analysis was performed for variations of ± 1 standard deviation. Analyses were executed in Python 3.11 (pandas, scikit-learn, scipy, statsmodels, libpysal, esda) and QGIS 3.34 for geospatial visualizations.

3 Results

The analysis included 1,874 Peruvian districts. The Shapiro-Wilk test rejected normality for all variables ($p < 0.001$). HDI presented a mean of 0.410 (SD = 0.112, range: 0.091-0.864). Extreme poverty showed a mean of 13.6% (SD = 12.2%, range: 0.0-86.7%). Average food vulnerability was 0.492 (SD = 0.156, range: 0.074-0.900).

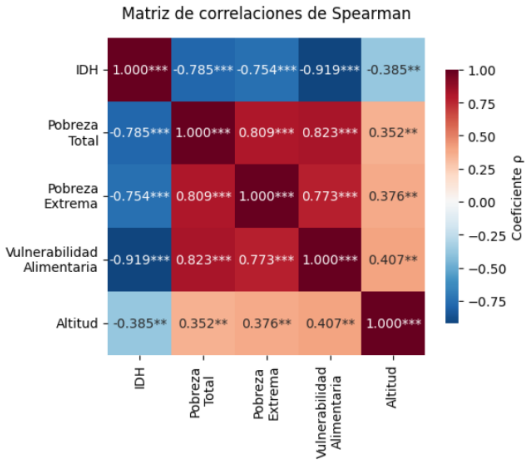

Figure 1: Spearman correlation matrix. Socioeconomic indicators show marked inverse relationships with HDI, highlighting food vulnerability as the most strongly associated factor.

Spearman's correlation matrix revealed strong associations (Figure 1). HDI presented a very strong negative correlation with food vulnerability ($\rho = -0.919$, $p < 0.001$) and strong correlations with total poverty ($\rho = -0.785$) and extreme poverty ($\rho = -0.754$). Altitude showed moderate correlation with HDI ($\rho = -0.385$) and food vulnerability ($\rho = 0.407$).

Table 1: Comparison of indicators by natural region

Var	Coast	HighL	RF	F
HDI	0.5	0.3	0.4	416.3***
Poverty	17.4	38.7	31.8	279.0***
Ext. Pov	3.5	18.4	11.0	196.5***
Food Vuln	0.2	0.5	0.5	527.1***

Note: ***$p < 0.001$. Ext. Pov. = Extreme Poverty; Food Vuln. = Food Vulnerability.

ANOVA revealed significant differences between regions ($p < 0.001$, Table 1). The highlands present the most critical indicators, with HDI 33% lower than the coast and extreme poverty levels five times higher, evidencing a persistent structural gap.

Table 2: Spatial autocorrelation (Moran's Index)

Variable	I	z	p
HDI	0.631	59.32	<0.001
Total poverty	0.681	61.86	<0.001
Extreme poverty	0.571	52.17	<0.001
Food vulnerability	0.631	60.26	<0.001

Moran's Index confirmed significant positive spatial autocorrelation for all indicators (Table 2). High values ($I > 0.57$) indicate that development conditions form geographic clusters, suggesting territorial contagion effects and the need for interventions that consider spatial dynamics.

The multiple regression model explained 85.9% of HDI variance ($R^2_{adj} = 0.859$, $F = 2842$, $p < 0.001$). Food vulnerability was the strongest predictor ($\beta = -0.114$, $p < 0.001$), followed by population density ($\beta = 0.011$, $p < 0.001$) and total poverty ($\beta = -0.013$, $p < 0.001$). Altitude showed no significant effect ($\beta = -0.001$, $p = 0.498$).

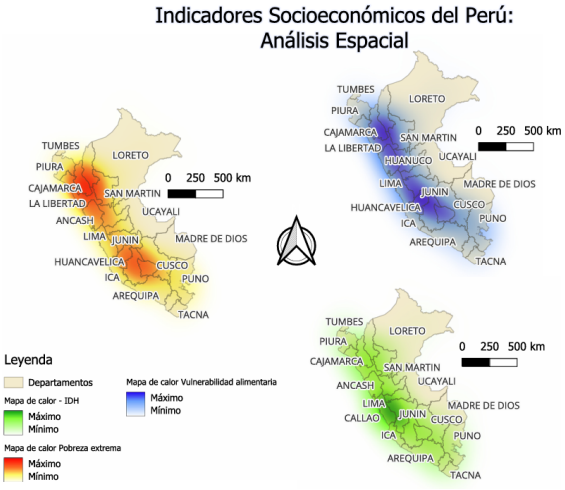

Figure 2: Spatial distribution of socioeconomic indicators. Geographic patterns reveal concentration of vulnerabilities in the south-central highlands (Cajamarca, Ayacucho, Apurímac) and concentrated development on the central coast.

The spatial distribution of indicators (Figure 2) shows clear territorial differentiation with vulnerability hotspots in high-altitude areas and concentrated development on the coast, confirming the spatially structured nature of disparities.

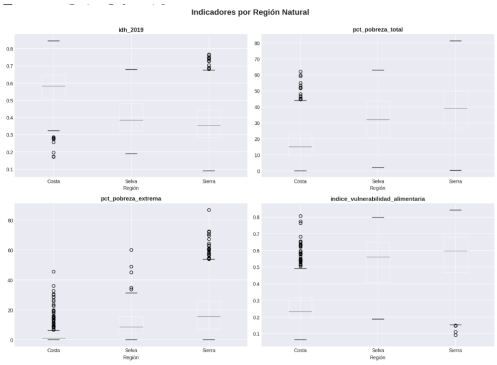

Figure 3: Comparison of indicators by natural region. Boxplots show that the highlands not only present worse averages but also greater internal dispersion, indicating significant territorial heterogeneity.

CVI presented a distribution with mean 0.435 (SD = 0.173, range: 0.007-0.942). Departments with highest CVI: Cajamarca (0.613), Ayacucho (0.537), Apurímac (0.533). Most vulnerable district: Curgos, La Libertad (CVI = 0.942).

Level classification showed that 42.4% of

Table 3: Distribution of districts by vulnerability level

Level (CVI)	*n*	%
Very Low (<0.2)	226	12.1
Low (0.2-0.4)	521	27.8
Medium (0.4-0.6)	795	42.4
High (0.6-0.8)	320	17.1
Very High (>0.8)	12	0.6
Total	**1,874**	**100.0**

districts present medium vulnerability, while 17.7% concentrate in high or very high vulnerability (Table 3), requiring priority attention in targeted public policies.

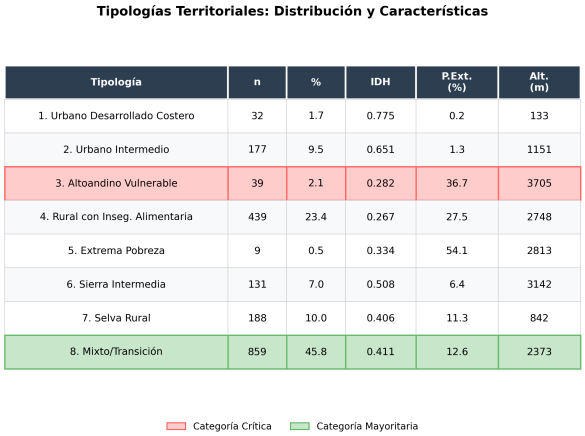

Figure 4: Territorial typologies: distribution and main characteristics. Eight categories were identified with marked differences in HDI, extreme poverty, and average altitude.

The territorial typology identified eight categories (Figure 4). The High-Altitude Vulnerable category presents critical triple vulnerability (HDI = 0.282, extreme poverty = 36.7%, altitude > 3,500 m), while Mixed/Transition groups the majority of districts (45.8%) with intermediate conditions representing opportunity for preventive policies.

All three algorithms showed high performance (Figure 5). Gradient Boosting slightly surpassed with $R^2 = 0.854$ and MAE = 0.039, equivalent to 4% error on the HDI scale. Consistency between training and test metrics indicates absence of overfitting.

Importance analysis (Figure 6) showed that food vulnerability explains 91.3% of predictive capacity, far exceeding population density (2.5%) and altitude (2.2%), evidencing that food insecurity is the central causal mechanism of low HDI.

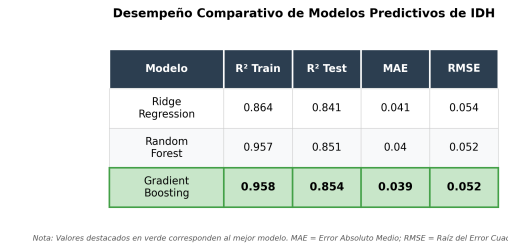

Figure 5: Comparative performance of HDI predictive models. Gradient Boosting achieved the best performance with $R^2 = 0.854$ and MAE = 0.039, surpassing Random Forest and Ridge Regression.

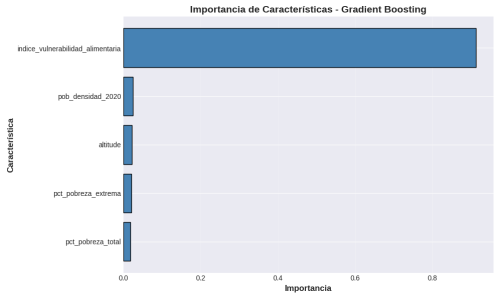

Figure 6: Importance of predictor variables in Gradient Boosting model. Food vulnerability dominates with 91.3% importance, confirming its central role in human development.

4 Discussion

Results confirm marked territorial patterns in Peru's socioeconomic disparities. Positive spatial autocorrelation (Moran's $I = 0.631$, $p < 0.001$) evidences that human development forms geographic *clusters*, consistent with literature on spatial poverty traps [8, 15, 9]. This finding suggests that interventions must consider spillover effects and surrounding territorial dynamics, not only intrinsic characteristics of each district [24].

The extremely strong correlation between HDI and food vulnerability ($\rho = -0.919$) constitutes the most relevant finding. This association, superior to that reported in previous studies for Latin America [21, 22], suggests that food insecurity is not merely a symptom

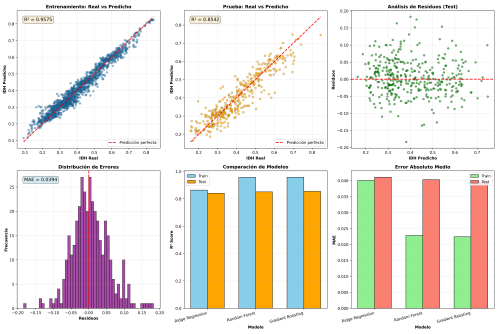

Figure 7: Gradient Boosting model performance. Predicted values align closely with actual values ($R^2 = 0.854$), with residuals normally distributed around zero, validating predictive robustness.

of poverty but a central causal mechanism. The predictive model reinforces this interpretation: food vulnerability explains 91.3% of predictive capacity. This result has direct implications: programs focused on food security could generate multiplier effects on human development [7, 23].

Differences between regions (Coast: HDI = 0.574; Highlands: 0.369; Rainforest: 0.400) reflect historical inequality patterns [10, 2, 3]. However, the proposed typology transcends the Coast-Highlands-Rainforest classification by identifying eight profiles that capture intraregional heterogeneities. "Intermediate Highlands" (HDI = 0.508) presents better conditions than "High-Altitude Vulnerable" (HDI = 0.282), despite sharing region and similar altitudes. This differentiation allows more precise interventions [11].

The "High-Altitude Vulnerable" typology (39 districts, 2.1%) deserves special attention. With HDI of 0.282, extreme poverty of 36.7%, and location > 3,500 meters above sea level, these territories face triple vulnerability placing them in humanitarian emergency. Their concentration in Apurímac, Ayacucho, and Puno coincides with historically marginalized zones [8]. Curgos (CVI = 0.942, extreme poverty = 86.7%) represents an extreme case requiring immediate intervention.

High predictive performance ($R^2 = 0.854$) demonstrates feasibility of estimating HDI with more accessible and frequently updated indicators. This capacity has significant practical value: UNDP updates HDI with considerable lag, while poverty and food vulnerability are monitored annually by INEI and MIDIS [1, 2]. The model could be implemented as an early warning system to identify at-risk districts [14, 25].

The non-significant effect of altitude in multiple regression ($\beta = -0.001$, $p = 0.498$), contrasting with significant bivariate correlation ($\rho = -0.385$), suggests that altitude operates as a *proxy* for more direct socioeconomic factors. This finding questions geographic determinist interpretations and highlights that deprivations in high-altitude areas result from sociohistorical exclusion processes [13, 9].

The "Mixed/Transition" category (45.8% districts, HDI = 0.411) represents both methodological challenge and political opportunity. These territories with intermediate indicators could move toward better or worse conditions. Focused preventive policies could avoid poverty traps, being more cost-effective than remedial interventions [15, 24].

This study presents limitations. First, cross-sectional design prevents robust causal inferences. Second, it is limited to variables in secondary datasets, excluding factors such as local institutional quality or social capital. Third, typological classification contains elements of subjectivity in threshold definition. Future studies could incorporate longitudinal analyses, integrate primary data through remote sensors [12, 14], and apply unsupervised learning to validate typologies [25].

Despite limitations, findings provide robust evidence on the multidimensional and spatially structured nature of socioeconomic disparities in Peru. Integration of spatial analysis, composite indices, typologies, and ML offers a replicable methodological framework for other Latin American contexts [22, 23]. Open-source code availability facilitates reproducibility.

5 Conclusions

This study demonstrates that socioeconomic disparities in Peru present structured and predictable territorial patterns. Main findings are:

First, food vulnerability emerges as the most strongly associated factor with human development ($\rho = -0.919$), explaining 91.3% of HDI predictive capacity. This finding positions food security as a central causal mechanism, with direct implications for public policies.

Second, the proposed CVI effectively synthesizes three critical dimensions. Identification of 332 districts (17.7%) with high/very high vulnerability provides precise territorial targeting, surpassing approaches based on isolated indicators.

Third, the eight-category typology reveals heterogeneities transcending the Coast-Highlands-Rainforest division. "High-Altitude Vulnerable" (39 districts, HDI = 0.282) requires urgent attention, while "Mixed/Transition" (45.8%) represents opportunity for preventive policies.

Fourth, predictive models ($R^2 = 0.854$, MAE = 0.039) demonstrate technical feasibility for estimating HDI in near real-time, enabling early warning systems.

Fifth, spatial autocorrelation (Moran's $I = 0.631$) confirms that interventions must consider territorial dynamics and spillover effects. *Hotspots* in the south-central highlands demand integrated territorial strategies.

Recommendations for public policy include: (1) prioritize food security programs as a transversal axis; (2) implement targeting based on CVI and typologies; (3) design differentiated interventions by typology; (4) establish predictive monitoring with ML; (5) adopt territorial approaches recognizing spatial interdependencies.

Future research should: (1) incorporate longitudinal analyses to establish causality; (2) integrate unconventional data (satellites, mobile records) [12, 14]; (3) apply deep learning to capture nonlinear interactions; (4) expand analysis to Latin American scale; (5) develop dynamic models projecting trajectories under alternative scenarios.

In conclusion, integration of spatial analysis, composite indices, typologies, and ML provides robust operational tools to characterize, predict, and target interventions on socioeconomic disparities. Application to the Peruvian case reveals urgency to address triple vulnerability in high-altitude and rural territories, leveraging opportunity window in transition districts. Recognition that human development is a spatially structured phenomenon must translate into differentiated territorial policies transcending uniform approaches.

References

[1] United Nations Development Programme. (2022). *Human Development Report 2021/2022*. New York: UNDP.

[2] National Institute of Statistics and Informatics. (2023). *Peru: Evolution of SDG Indicators to 2022*. Lima: INEI.

[3] Escobal, J., & Ponce, C. (2020). Spatial patterns of poverty in Peru. *World Development*, 133, 104992.

[4] Anselin, L. (2019). A local indicator of multivariate spatial association. *Geographical Analysis*, 51(2), 133-150.

[5] Getis, A. (2019). Spatial autocorrelation. In *Handbook of Regional Science* (pp. 1477-1489). Springer.

[6] Ministry of Development and Social Inclusion. (2021). *Food Insecurity Vulnerability Map*. Lima: MIDIS.

[7] Alkire, S., Kanagaratnam, U., & Suppa, N. (2018). The global MPI: 2018 revision. *OPHI MPI Note 46*, Oxford.

[8] Trivelli, C., et al. (2020). *Dynamics of Rural Poverty in Peru*. IEP.

[9] Beuermann, D. W., et al. (2022). Mobile phones and economic development in rural Peru. *Journal of Development Economics*, 156, 102814.

[10] Jaramillo, M., & Campos, G. (2019). Growth and income distribution. *Economía*, 42(83), 147-175.

[11] Lavado, P., & Campos, R. (2020). Spatial poverty traps in Peru. *Economics Letters*, 194, 109362.

[12] Jean, N., et al. (2016). Combining satellite imagery and ML to predict poverty. *Science*, 353(6301), 790-794.

[13] Blumenstock, J. E. (2018). Don't forget people in big data. *Nature*, 561(7722), 170-172.

[14] Yeh, C., et al. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1), 1-11.

[15] Santos, M. E., et al. (2019). *MPI for Latin America*. Series No. 97, ECLAC.

[16] Molnar, C. (2020). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

[17] Castagnetto, J. M. (2021). *UBIGEO Peru: Enriched Dataset*. https://github.com/achalmed/ubigeo-peru

[18] Hastie, T., Tibshirani, R., & Friedman, J. (2020). *The Elements of Statistical Learning* (2nd ed.). Springer.

[19] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[20] Friedman, J. H. (2001). Greedy function approximation. *Annals of Statistics*, 29(5), 1189-1232.

[21] FAO. (2020). *Panorama of Food Security in Latin America 2020*. Santiago: FAO.

[22] Graziano da Silva, J., et al. (2021). Food security governance in Latin America. *Global Food Security*, 28, 100484.

[23] World Food Programme. (2022). *Food Security Analysis: Peru*. Rome: WFP.

[24] Chasco, C., & López, F. A. (2021). Spatial econometrics and regional science. *Papers in Regional Science*, 100(1), 3-22.

[25] Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), eabe8628.