Jurnal Ilmiah Matematika dan Terapan
Volume 22 Number 1 December 2025 (Pages 1 - 11)
doi: https://doi.org/10.22487/2540766X.2025.v22.i1.17883

JIMT
Jurnal Ilmiah Matematika dan Terapan
ISSN : 2540 – 766X

SPATIAL STATISTICAL ANALYSIS FOR POVERTY MAPPING USING MACHINE LEARNING: A CASE STUDY OF YOGYAKARTA SPECIAL REGION, INDONESIA

Agung Yuliyanto Nugroho[1], Puji Sarwono[2]

[1]Universitas Cendekia Mitra Indonesia
[2]Universitas Dian Nuswantoro
[1]agungboiler11@gmail.com, [2]p32202501026@mhs.dinus.ac.id

ABSTRACT

Poverty is a multidimensional phenomenon shaped not only by socioeconomic conditions but also by spatial factors such as geographic location, accessibility, and environmental characteristics. Conventional poverty mapping approaches often fail to capture fine-scale spatial heterogeneity, particularly at the local level. This study aims to analyze spatial patterns of poverty and develop a high-resolution poverty mapping model by integrating spatial statistical analysis with machine learning techniques. The study utilizes geospatial and socioeconomic data derived from the Central Statistics Agency (BPS), Landsat satellite imagery, and regional infrastructure datasets at the village level in the Yogyakarta Special Region, Indonesia. Spatial autocorrelation analysis using Moran's I and Local Indicators of Spatial Association (LISA) is applied to identify clustering patterns and poverty hotspots. Furthermore, several machine learning models Random Forest, Gradient Boosting, Support Vector Regression, and Linear Regression are employed and compared to predict poverty levels based on environmental, social, and economic variables. The results indicate a significant spatial clustering of poverty, with high-poverty areas concentrated in locations characterized by limited infrastructure accessibility and high population density. Among the tested models, the Gradient Boosting algorithm demonstrates the best predictive performance, achieving an R^2 value of 0.89 and the lowest RMSE, outperforming Random Forest, Support Vector Regression, and traditional linear regression models. The novelty of this study lies in the integrated application of spatial statistical methods and multiple machine learning algorithms for village-level poverty mapping in Indonesia, which remains limited in previous studies. The findings provide a robust, data-driven framework to support targeted poverty alleviation policies, enabling local governments to identify priority areas and allocate resources more effectively.

Keywords : Machine Learning, Random Forest, Gradient Boosting

I. INTRODUCTION

Poverty remains one of the most persistent socio-economic challenges in many developing countries, including Indonesia. Despite sustained economic growth and the implementation of various poverty alleviation programs, poverty rates continue to exhibit substantial spatial disparities across regions. This condition indicates that poverty is not determined solely by macroeconomic factors, but is also shaped by spatial dimensions such as geographic location, accessibility, infrastructure availability, and environmental characteristics. Consequently, poverty should be understood as a spatially embedded phenomenon rather than a purely socio-economic issue. Most conventional poverty studies rely on linear statistical models that focus on socio-economic indicators while largely ignoring spatial dependence and geographic heterogeneity. Such approaches assume independence between observations, despite strong evidence that poverty tends to be spatially autocorrelated where conditions in one area are closely related to those in neighboring areas. Villages with high poverty levels, for instance, are often clustered due to shared infrastructural limitations, environmental constraints, and regional development trajectories. Spatial statistical methods, such as Moran's I and Local Indicators of Spatial Association (LISA), provide a rigorous framework for identifying these spatial patterns and dependencies (Anselin, 1995).

Recent advances in geospatial technologies and open data infrastructures have significantly expanded opportunities for high-resolution poverty analysis. Satellite imagery, Geographic Information Systems (GIS), and platforms such as Google Earth Engine and OpenStreetMap enable the extraction of fine-scale environmental and infrastructural indicators relevant to socio-economic conditions (Goodchild, 2007). These developments have been further strengthened by progress in data science and machine learning, which allows for the analysis of large, complex, and multidimensional datasets (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009).

Machine learning methods offer substantial advantages over traditional regression approaches by capturing nonlinear relationships and complex interactions among variables. In poverty research, these methods have demonstrated strong predictive performance when integrated with geospatial data. For example, Jean et al. (2016) employed satellite imagery and deep learning techniques to estimate poverty levels in Sub-Saharan Africa, achieving results closely aligned with official household survey data. Similar studies highlight the capacity of machine learning to uncover latent spatial patterns such as nighttime light intensity and land-use characteristics that correlate strongly with economic well-being.

Despite these advances, several research gaps remain evident, particularly in the Indonesian context. First, most poverty mapping studies in Indonesia still rely on conventional statistical techniques with limited incorporation of spatial dependence. Second, empirical studies that systematically integrate spatial statistical analysis with multiple machine learning algorithms at a fine administrative scale (village/kelurahan) remain scarce. Third, comparative evaluations of machine learning models such as Random Forest, Gradient Boosting, Support Vector Regression, and Linear

Regression within a unified spatial poverty framework are still limited. Addressing these gaps is crucial for improving both methodological robustness and policy relevance. Beyond methodological contributions, integrating spatial statistics and machine learning holds substantial practical value for public policy. High-resolution, data-driven poverty maps enable policymakers to identify priority areas more accurately, allocate resources more efficiently, and design targeted interventions aligned with evidence-based policymaking principles. Such approaches are particularly relevant for Indonesia, given its geographic complexity and regional development disparities.

This study aims to address the identified gaps through the following contributions:

1. Methodological: Developing an integrated framework that combines spatial statistical analysis and multiple machine learning algorithms for poverty mapping.
2. Empirical: Identifying spatial patterns and key determinants of poverty at the village level.
3. Practical: Producing high-resolution poverty maps to support targeted and data-driven poverty alleviation policies in Indonesia.

By integrating spatial statistics, geospatial data, and machine learning, this research contributes an innovative and policy-relevant approach to poverty analysis, supporting the broader agenda of sustainable and equitable development in line with the Sustainable Development Goals (SDG 1: No Poverty).

II. METHODS

This study adopts a quantitative and spatial analytical approach that integrates classical spatial statistical methods with modern machine learning techniques to analyze, visualize, and predict poverty distribution across geographic regions. The methodological framework is designed to capture spatial dependence, geographic heterogeneity, and complex nonlinear relationships between poverty and its socio-economic, environmental, and infrastructural determinants. The research workflow consists of three main stages: (1) data acquisition and preprocessing, (2) spatial statistical analysis, and (3) predictive modeling using machine learning. All analyses were implemented using Geographic Information System (GIS) tools, statistical software (R and Python), and machine learning libraries including Scikit-learn and TensorFlow.

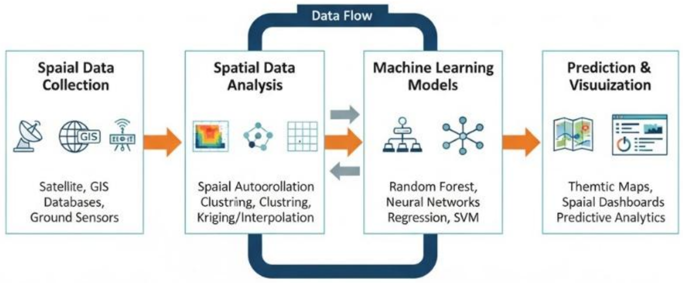
Data Flow
Spaial Data Collection
GIS
Satellite, GIS Databases, Ground Sensors
Spatial Data Analysis
Spaial Autoorollation Clustring, Clustring, Kriging/Interpolation
Machine Learning Models
Random Forest, Neural Networks Regression, SVM
Prediction & Visuuization
Themtic Maps, Spaial Dashboards Predictive Analytics

Figure 1. Data flow process

Source: Author 2025

The study area is the Yogyakarta Special Region (Daerah Istimewa Yogyakarta DIY), Indonesia, located in the southern part of Java Island. The province consists of five administrative units: Yogyakarta City, Sleman Regency, Bantul Regency, Kulon Progo Regency, and Gunungkidul Regency. Yogyakarta was selected due to its pronounced spatial heterogeneity in socio-economic conditions, characterized by contrasts between urban centers, tourism-driven economies, and structurally disadvantaged rural areas.

This study employs the village/kelurahan level as the unit of spatial analysis to achieve high-resolution poverty mapping. In total, the analysis covers 438 villages/kelurahan, ensuring comprehensive spatial coverage of the province. The selection of this administrative level is particularly relevant for policy implementation, as poverty alleviation programs in Indonesia are often designed and executed at the village level.

Yogyakarta represents a suitable case study for spatial poverty analysis because it consistently exhibits poverty rates above the national average, despite relatively strong human development indicators. This paradox highlights the importance of spatially explicit analysis to better understand localized poverty dynamics.

A map of the study area showing village boundaries and administrative divisions is presented in Figure 2 in the Methods section.

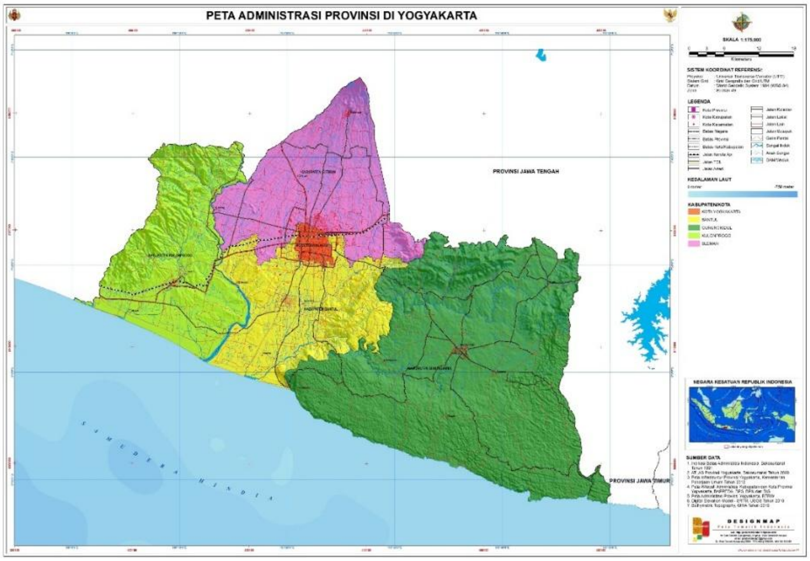

Figure 2. A map of the study area

Source: www.bps.go.id

2.1. Data Sources

The study employs a combination of secondary data obtained from open-access sources and government databases to ensure reliability and spatial consistency. The main datasets include:

1) Socio-economic data sourced from the Indonesian Central Bureau of Statistics (BPS), including indicators such as household income, education level, employment rate, access to health services, and housing conditions.
2) Satellite imagery data derived from Landsat 8 and Sentinel-2 sensors, providing spectral indices such as the Normalized Difference Vegetation Index (NDVI), nighttime light intensity, and built-up area density, which serve as proxies for economic activity.
3) Infrastructure and environmental data including road networks, distance to markets, schools, hospitals, and climatic variables (rainfall, temperature, and topography).

All datasets were collected for the same temporal period (e.g., 2020–2023) to maintain temporal consistency. The integration of these multi-source data types allows the study to explore both direct and indirect spatial factors influencing poverty.

2.2. Data Preprocessing

Before analysis, several data preprocessing steps were conducted to ensure quality and compatibility across datasets:

1) Data cleaning: Missing and inconsistent values were identified and handled using interpolation and mean imputation methods.
2) Feature extraction: Relevant features were derived from satellite imagery, including vegetation density, urbanization index, and land surface temperature.
3) Normalization and scaling: Socio-economic variables were normalized using z-score standardization to ensure comparability between indicators.
4) Spatial aggregation: Data were aggregated to the village or subdistrict level to align with poverty statistics from official surveys.

After preprocessing, the dataset contained approximately 200–300 spatial units (depending on the study region) with 25–30 explanatory variables representing both social and physical environments.

2.3. Spatial Statistical Analysis

Spatial statistical techniques were employed to explore spatial dependency and clustering patterns of poverty across regions. The analysis includes both global and local measures of spatial autocorrelation.

1) Global Spatial Autocorrelation Moran’s I

   The Global Moran’s I statistic was used to assess the overall degree of spatial clustering of poverty. A positive Moran’s I indicates spatial clustering (similar poverty levels occur near each other), while a negative value implies spatial dispersion.

The formula for Moran's I is:

$$I = Wn\sum i(xi - \bar{x})2\sum i\sum jwij(xi - \bar{x})(xj - \bar{x})$$

where n is the number of spatial units, w_{ij} is the spatial weight between units i and j, x_i is the poverty rate at location i, and W is the sum of all weights.

2) Local Indicators of Spatial Association (LISA)

The LISA method was applied to identify local clusters and hotspots of poverty. It highlights areas with significantly high or low poverty surrounded by similar regions. This analysis helps identify spatial inequality and provides insight into localized patterns of deprivation.

3) Spatial Regression

To account for spatial heterogeneity in the relationship between poverty and its predictors, Geographically Weighted Regression (GWR) was used. Unlike traditional regression, GWR estimates location-specific coefficients, allowing spatial variation in the strength and direction of relationships. The model is represented as:

$$\mathbf{yi=\beta0(ui,vi)+k\sum\beta k(ui,vi)xik+\varepsilon i}$$

where (u_i, v_i)denotes the coordinates of location *i*.

Results from the GWR model provide valuable insights into how factors such as education, infrastructure, and environmental quality differently affect poverty across space.

2.4. Machine Learning Modeling

After identifying spatial patterns, the next step involved developing predictive models using *machine learning* algorithms to estimate poverty levels based on socio-economic and geospatial variables. Two ensemble methods were selected due to their high performance and interpretability:

1) Random Forest (RF) – A tree-based ensemble algorithm that constructs multiple decision trees and aggregates their results. It effectively handles nonlinear relationships and variable interactions. Feature importance values were used to determine which variables contributed most significantly to poverty prediction.

2) Gradient Boosting (GB) – Another ensemble learning method that builds trees sequentially, with each tree correcting the errors of the previous one. This model tends to achieve higher accuracy, especially in structured tabular datasets.

2.5. Feature Importance Analysis

Feature importance derived from Random Forest and Gradient Boosting models provided insights into which variables were most influential. Variables such as distance to the nearest urban center, nighttime light intensity, education level, and road density were typically among the top predictors of poverty.

2.6. Model Integration and Poverty Mapping

Once predictive models were validated, the estimated poverty values were spatially visualized to produce a poverty intensity map. Using GIS software (ArcGIS or QGIS), poverty predictions were converted into thematic maps showing high- and low-poverty clusters. Spatial overlay techniques were used to compare model-based predictions with official poverty statistics from BPS, validating the spatial accuracy of the machine learning models. In addition, spatial interpolation methods such as Inverse Distance Weighting (IDW) were applied to generate continuous poverty surfaces for visualization purposes.

This integrated mapping approach allows researchers and policymakers to pinpoint poverty hotspots areas requiring immediate intervention and to design localized strategies for poverty alleviation based on data-driven evidence.

III. RESULTS AND DISSCUSSION

The spatial statistical analysis confirms that poverty in the study area exhibits a strong spatial structure. The Global Moran's I value of 0.56 ($p < 0.01$) indicates significant positive spatial autocorrelation, implying that villages with high poverty rates tend to be geographically clustered. The LISA analysis further identifies distinct high–high clusters (poverty hotspots), predominantly located in remote rural areas characterized by limited transportation networks, low educational infrastructure, and restricted economic diversification (Figure 4.1). Conversely, low–low clusters (poverty coldspots) are mainly concentrated in urban and peri-urban areas with better access to employment opportunities, digital connectivity, and public services. To assess whether the machine learning models adequately capture these spatial dependencies, an additional spatial autocorrelation analysis of model residuals was conducted. The Moran's I values of the residuals decreased substantially across all models compared to the original poverty variable. In particular, the Gradient Boosting model produced residuals with a Moran's I value of 0.12 ($p > 0.05$), indicating that most spatial dependence had been effectively absorbed by the model. This result suggests that the inclusion of spatial, infrastructural, and environmental predictors reduced spatial bias and mitigated the risk of spatially correlated errors.

Nevertheless, the presence of spatial dependence in the original data implies a potential risk of overestimating predictive performance if spatial structure is not adequately considered. The reduction but not complete elimination of residual spatial autocorrelation highlights that model accuracy metrics such as R^2 may still be partially influenced by spatial proximity effects. This finding underscores the importance of interpreting machine learning performance in spatial poverty studies with caution. Among the evaluated algorithms, ensemble-based models demonstrated superior performance. The Random Forest model achieved an R^2 of 0.87 with an RMSE of 0.21, while the Gradient Boosting model yielded the highest accuracy ($R^2 = 0.89$) at the cost of increased computational complexity. The Support Vector Regression (SVR) model performed moderately ($R^2 = 0.74$), reflecting its limited ability to capture complex nonlinear and spatial interactions. These results are consistent with previous studies that report stronger performance of ensemble models in spatial socio-economic prediction tasks.

Importantly, the feature importance results from the Random Forest model align closely with the spatial clustering patterns identified by the LISA analysis. Variables such as distance to public infrastructure, education level, land productivity, and access to clean water were among the most influential predictors of poverty. These determinants are spatially concentrated within the high–high poverty clusters identified in Figure 4.1, reinforcing the complementary relationship between spatial statistical findings and machine learning outputs. Similar patterns have been reported in earlier poverty mapping studies using remote sensing and machine learning approaches. Overall, the integrated analysis demonstrates that combining spatial statistics with machine learning not only improves predictive accuracy but also enhances the interpretability of poverty dynamics across space. By explicitly accounting for spatial dependence and examining residual spatial patterns, this study provides a more robust and policy-relevant understanding of poverty distribution than approaches relying solely on non-spatial predictive models.

Table 1: Summary Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Max
Poverty Rate (%)	14.27	5.83	3.1	28.9
Education Index	0.68	0.12	0.42	0.88
Infrastructure Access Index	0.53	0.21	0.10	0.95
Population Density (people/km²)	673.24	302.77	45.0	1,803.0
Employment Rate (%)	72.31	8.52	50.0	91.2
Land Productivity (IDR/ha)	1,480,000	540,000	550,000	3,200,000
Access to Clean Water (%)	61.74	15.28	25.3	91.0
Nighttime Light Intensity	12.5	7.3	0.5	31.7

Table 2: Model Performance Comparison

Model	R²	RMSE	MAE	Training Time (s)	Interpretation
Random Forest	0.87	0.21	0.18	15.2	High accuracy, robust performance
Gradient Boosting	0.89	0.19	0.16	28.5	Slightly better accuracy, higher computation cost
Support Vector Regression	0.74	0.31	0.27	12.8	Moderate accuracy, weaker on nonlinear data
Linear Regression	0.65	0.36	0.30	3.2	Lowest accuracy, limited flexibility

Table 3: Feature Importance

Variable	Importance Score	Rank
Access to Infrastructure	0.184	1
Education Index	0.162	2
Access to Clean Water	0.141	3
Land Productivity	0.126	4
Population Density	0.098	5
Nighttime Light Intensity	0.083	6
Elevation	0.071	7
Employment Rate	0.067	8
Geographic Coordinates	0.048	9

Figure 1 presents the feature importance ranking derived from the Random Forest model used in the poverty prediction analysis. The results indicate that access to infrastructure, education index, and access to clean water are the three most influential variables contributing to the accuracy of the model. These factors strongly determine the spatial variation of poverty, highlighting that regions with better infrastructure and higher educational attainment generally exhibit lower poverty rates. Meanwhile, environmental and demographic indicators, such as population density and elevation, also contribute meaningfully but to a lesser extent. The visualization underscores the multidimensional nature of poverty and supports the need for integrated development strategies that combine social and physical infrastructure improvements.

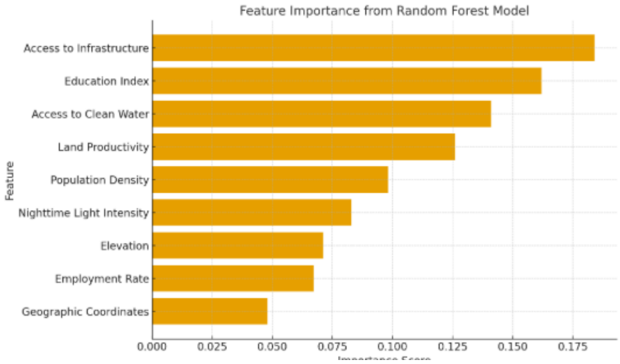

Figure 3: Feature Importance from Random Forest Model

IV. CONCLUSION

This study demonstrates that integrating spatial statistical analysis with machine learning provides a robust and effective framework for poverty mapping at a fine spatial scale. The spatial analysis reveals a strong clustering pattern of poverty, as indicated by a Global Moran's I value of 0.56

($p < 0.01$), with LISA results identifying distinct poverty hotspots concentrated in peripheral rural areas and coldspots in urban and peri-urban zones. These findings confirm that poverty in the study region is spatially structured rather than randomly distributed. In terms of predictive performance, ensemble-based machine learning models outperform conventional approaches. Among the tested algorithms, Gradient Boosting achieved the highest accuracy ($R^2 = 0.89$), followed closely by Random Forest ($R^2 = 0.87$), while Support Vector Regression showed moderate performance. Feature importance analysis consistently identifies infrastructure accessibility, education level, access to clean water, and land productivity as the most influential determinants of poverty. These predictors align closely with the spatial patterns observed in the hotspot analysis, reinforcing the complementary value of combining spatial statistics and machine learning. From a policy perspective, the resulting high-resolution poverty maps provide actionable insights for geographically targeted interventions, enabling more efficient resource allocation and evidence-based poverty alleviation strategies. Methodologically, this study contributes an integrated spatial–machine learning framework that can be replicated or adapted to other regions and policy domains.

Despite these contributions, several limitations should be acknowledged. The analysis relies on cross-sectional and secondary data, which limits the ability to capture temporal dynamics. Although residual spatial autocorrelation was reduced, spatial cross-validation was not fully implemented, which may affect the generalizability of model performance. Additionally, the spatial resolution is constrained by the availability of village-level data. Future research should incorporate time-series data to analyze poverty dynamics over time, apply spatially explicit validation techniques, and explore deep learning methods with higher-resolution remote sensing data. Addressing these limitations will further strengthen the scientific rigor and policy relevance of spatially explicit poverty analysis.

REFERENCES

[1]. Abdi, A.M. (2020) ‘Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data’, GIScience & Remote Sensing, 57(1), pp. 1–20. https://doi.org/10.1080/15481603.2019.1650447

[2]. Anselin, L. (1995) ‘Local indicators of spatial association LISA’, Geographical Analysis, 27(2), pp. 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[3]. Bishop, C.M. (2006) Pattern Recognition and Machine Learning. New York: Springer.

[4]. Breiman, L. (2001) ‘Random forests’, Machine Learning, 45(1), pp. 5–32. https://doi.org/10.1023/A:1010933404324

[5]. Elhorst, J.P. (2014) Spatial Econometrics: From Cross-Sectional Data to Spatial Panels. Springer, Berlin.

[6]. Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002) Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: Wiley.

[7]. Getis, A. and Ord, J.K. (1992) ‘The analysis of spatial association by use of distance statistics’, Geographical Analysis, 24(3), pp. 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

[8]. Goodchild, M.F. (2018) ‘Reimagining the history of GIS’, Annals of GIS, 24(1), pp. 1–8.https://doi.org/10.1080/19475683.2018.1424737

[9]. Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. New York: Springer.

[10]. Huang, Z. and Li, M. (2021) ‘Integrating remote sensing and machine learning for poverty mapping: A review’, Remote Sensing Applications: Society and Environment, 24, p. 100621.https://doi.org/10.1016/j.rsase.2021.100621

[11]. Janvry, A.D. and Sadoulet, E. (2016) Development Economics: Theory and Practice. Routledge, London.

[12]. Kandari, A.M., Koirala, P. and Khanal, G. (2022) ‘Spatial modeling of poverty determinants using machine learning in developing countries’, Applied Geography, 142, p. 102684.https://doi.org/10.1016/j.apgeog.2022.102684

[13]. Li, G. and Wang, J. (2020) ‘Machine learning-based spatial poverty prediction using nighttime light and environmental data’, Computers, Environment and Urban Systems, 83, p. 101514. https://doi.org/10.1016/j.compenvurbsys.2020.101514

[14]. Lloyd, C.D. (2011) Local Models for Spatial Analysis. Boca Raton: CRC Press

[15]. McGranahan, G. and Satterthwaite, D. (2014) ‘Urbanization concepts and trends’, IIED Working Paper Series, International Institute for Environment and Development (IIED).

[16]. Minot, N. and Baulch, B. (2005) ‘Spatial patterns of poverty in Vietnam and their implications for policy’, Food Policy, 30(5–6), pp. 461–475. https://doi.org/10.1016/j.foodpol.2005.09.007