*Research article*

# Regional and spatial dependence of poverty factors in Thailand, and its use in Bayesian hierarchical regression analysis

Statistical Journal of the IAOS
1–16

DOI: 10.1177/18747655261416694
journals.sagepub.com/home/sji

**Irving Gómez-Méndez[1] and Chainarong Amornbunchornvej[2]**

**Abstract**

Poverty in Thailand shows strong spatial dependence that existing administrative boundaries fail to capture, leading to policies that overlook local socioeconomic realities. This study proposes a data-driven regionalization framework to infer geographically coherent "policy regions" that better represent poverty dynamics. Using household-level data from the Thai People Map and Analytics Platform (TPMAP), we analyze spatial autocorrelation across multiple poverty factors through Moran's statistics and principal component analysis, followed by spatially constrained hierarchical clustering to delineate coherent regions. Bayesian hierarchical and geographically weighted regression models are then employed to examine how education influences household income at provincial, regional, and national levels. Our results identified six regions that reflect more accurately poverty structures than official divisions. Northern and Northeastern Thailand emerge as the regions most affected by low education, income, and savings, while Central Thailand shows higher inequality. The inferred regions demonstrate that spatially contiguous provinces often share similar socioeconomic structures, suggesting that policy targeting should align with these patterns rather than provincial borders. Our findings provide a quantitative foundation for evidence-based regional planning, enabling policymakers to design differentiated yet regionally coordinated interventions. The approach illustrates how spatial statistical modeling can bridge the gap between data analysis and effective poverty-alleviation policy.

**Keywords**

Hierarchical models, bayesian regression, income, education, poverty

Received: 2 July 2025; accepted: 6 January 2026

## 1 Introduction

### *1.1 Motivation*

Poverty is a serious issue in many developing countries. In 2030, one of the United Nation (UN)'s plans is to eliminate poverty worldwide.[1] Thus, finding optimal policies are necessary to help the governments cope with the problem.[2–4]

Persistent poverty remains one of Thailand's most complex policy challenges. Despite national progress in reducing deprivation, disparities across regions continue to shape the geography of opportunity. Policies designed at the national or ministerial level often follow administrative boundaries that do not necessarily correspond to socioeconomic structures on the ground.[5] As a result, interventions risk being mismatched to local conditions—some provinces share similar poverty profiles[6–9] but fall under different regional plans, while others within the same administrative region exhibit contrasting needs. *The critical policy question, therefore, is not only how much to spend on poverty alleviation but where and in what form.* For instance, focusing on lacking of health care support, it is possible for some areas to build a new hospital while it might be harder for remote areas in mountain regions to build it. Hence, instead of insisting on having a new hospital for a specific province, providing telehealth[10] might possibly alleviate this issue for several nearby mountainous areas beyond administrative borders.

[1]Artificial Intelligence and Computer Engineering, CMKL University, Bangkok, 10520 Thailand
[2]National Electronics and Computer Technology Center (NECTEC)

**Corresponding author:**
Irving Gómez-Méndez, Artificial Intelligence and Computer Engineering, CMKL University, Bangkok, 10520, Thailand.
Email: gomendez.irving@gmail.com

### 1.2 Policy gap and research question

Most existing studies of Thailand's poverty rely on descriptive regional comparisons or the Multidimensional Poverty Index (MPI),[2,5,11] which summarize deprivation but ignore spatial dependencies between neighboring provinces. Spatial autocorrelation–where nearby provinces tend to share similar income, education, or savings characteristics–implies that policy spillovers extend beyond administrative lines.[5] Recognizing and quantifying this dependence is essential for designing targeted yet scalable policies.

Hence, in this paper, we address the following research question:

*How can data-driven regionalization based on spatial dependence among poverty factors improve the design of geographically targeted poverty-alleviation policies in Thailand?*

### 1.3 Related works

Poverty is widely recognized as a multidimensional phenomenon that extends beyond monetary deprivation to encompass deficits in education, health, and access to essential resources.[1,5,9,12] This multidimensionality complicates both measurement and intervention, as progress along one dimension does not necessarily imply overall improvement in well-being.[1] The Multidimensional Poverty Index (MPI) has become a standard measure for capturing this complexity by integrating indicators across multiple dimensions of deprivation.[12–14] While the MPI provides a comprehensive overview of poverty conditions, it remains limited in guiding spatially differentiated policy responses. Specifically, it does not reveal whether poverty patterns are geographically clustered, or whether similar challenges are shared among neighboring regions.[5,9]

To address the heterogeneity of poverty across spatial scales, the work in[9] proposed a Minimum Description Length (MDL) and Gaussian Mixture Model (GMM) framework[15–17] to infer the optimal resolution of regional partitions for policy formulation. Their approach demonstrated that poverty varies across multiple geographic levels, suggesting that uniform national policies are suboptimal. However, the model did not include a mechanism for spatial dependence, meaning that relationships among adjacent provinces were not modeled and the resulting analysis could not fully capture inter-provincial dynamics.

A subsequent study by[5] advanced this line of research by applying Bayesian hierarchical modeling to capture the nested relationships between larger administrative regions and their subunits. The framework proved valuable for understanding cross-level dependencies and was later utilized in policy evaluation settings.[18,19] Nevertheless, prior works have not yet examined how regional and spatial dependence jointly shape poverty outcomes in Thailand. In particular, the question of whether neighboring provinces share similar deprivation structures–and how such spatial correlations can inform targeted, regionally coordinated policy design–remains open. In this study, we address this gap by combining spatial autocorrelation analysis and hierarchical modeling to infer data-driven, geographically coherent regions that better represent Thailand's poverty dynamics.

### 1.4 Contribution to literature and policy

As noted by,[5] uniform national policies for any administrative region are insufficient to address Thailand's complex poverty landscape, while designing fully customized interventions for every province is impractical due to resource constraints and inter-provincial dependencies. This study introduces a middle-ground approach by identifying data-driven regional clusters that group provinces with similar socioeconomic profiles beyond administrative borders. We integrate spatial statistical analysis, Moran's clustering techniques,[20] and Bayesian hierarchical modeling[21,22] to uncover how poverty-related variables co-vary across space and to estimate the relationship between education and income at provincial, regional, and national levels. The inferred regions provide a more coherent representation of Thailand's poverty dynamics, offering an analytical foundation for regionally coordinated yet locally adapted policy design. Our analysis connects spatial statistics directly to policy design, showing how regional patterns can guide practical poverty-alleviation strategies, transforming spatial statistics into a decision-support framework for poverty alleviation.

A central contribution of this study is to show that Thailand's spatial poverty patterns cannot be adequately understood using only standard administrative regions. When we restrict the analysis to official regional divisions, several important structures remain hidden. In contrast, the data-driven regionalization reveals patterns that more closely follow socioeconomic gradients than ministerial boundaries. In particular, our results highlight the following contrasts between administrative and inferred regions:

- **Northern vs. Northeastern Thailand.** Official classifications often treat the North and Northeast as broad, internally homogeneous regions. Our analysis shows that they in fact contain *distinct* deprivation clusters: Northern provinces concentrate low education, low savings, and high alcohol consumption, whereas Northeastern provinces combine relatively low inequality with very high levels of formal debt.
- **Bangkok–Pattaya as a functional region.** The inferred Bangkok–Pattaya cluster emerges as a cohesive high-education, high-income region that cuts across administrative lines. It is also the only region where average years of education exceed the 12-year

threshold, a pattern that is not visible when using standard ministry-defined regions.
- **Central Thailand as a high-inequality corridor.** Central provinces form a contiguous region with income levels comparable to Bangkok–Pattaya but markedly higher inequality. These provinces are not grouped together in official regional schemes, so their shared profile of "high income but high inequality" would be obscured without a data-driven regionalization.

Taken together, these findings indicate that spatial poverty patterns in Thailand follow continuous socioeconomic gradients rather than administrative borders. The inferred regions group provinces that share common structural challenges, such as education deficits, high household indebtedness, limited savings, or elevated inequality. Therefore, they provide a more appropriate scale for designing coordinated interventions.

Building on these insights, this study makes three contributions:

- **Methodological contribution.** We propose a reproducible spatial regionalization framework that combines Moran's statistics, principal component analysis (PCA), and Bayesian hierarchical regression to delineate policy-relevant regions grounded in spatial dependence, rather than in pre-existing administrative divisions.
- **Empirical contribution.** Using household-level TPMAP data, we show that poverty-related indicators are strongly spatially autocorrelated and that six inferred regions better capture these dependencies than Thailand's official regional structure. In particular, the Bangkok–Pattaya, Central, Northern, and Northeastern regions exhibit distinct configurations of education, income, savings, debt, and inequality that would not be apparent under standard administrative groupings.
- **Policy contribution.** By estimating Bayesian hierarchical and geographically weighted regression models, we quantify how returns to education vary across the inferred regions and demonstrate that these variations align with the spatially defined clusters rather than with administrative regions. This provides a quantitative basis for regionally coordinated but locally adapted poverty-alleviation strategies that match the true geography of deprivation.

## 2 Material and methods

We use household-level data from the TPMAP, collected in 2022, which provides province-level indicators of education, income, savings, debt, inequality, and behavioral factors. Our methodological framework integrates spatial dependence diagnostics, multidimensional clustering, and hierarchical regression to examine how poverty-related characteristics co-vary across space and how education relates to income across regions.

### 2.1 Spatial dependence diagnostics

To assess whether poverty indicators exhibit geographic structure, we compute global Moran's I for each variable using a 5-nearest-neighbor spatial weights matrix. This diagnostic quantifies whether provinces with similar socioeconomic profiles are spatially clustered, an essential prerequisite for data-driven regionalization. We next use local Moran's I to detect clusters of high–high or low–low deprivation due to its popularity in spatial analyses.[23–27] These measures help reveal the extent to which poverty is spatially autocorrelated rather than randomly distributed across provinces.

### 2.2 Principal component analysis

Despite their popularity, Moran's clusters present three important drawbacks. First, there is no guarantee that the clusters are not fragmented. Second, since the clusters are created only with provinces whose local Moran's I is significantly different from its expected value under the hypothesis of no spatial correlation, several provinces might not be assigned to any cluster. Third, only one variable can be analyzed at a time, ignoring the multidimensional nature of poverty.

Thus, to consider the multidimensional nature of poverty from a more integrated perspective, we performed a principal component analysis (PCA) over the standardized variables. The first principal component summarizes the dominant socioeconomic gradient across Thailand—contrasting financially secure, high-education provinces with provinces exhibiting low income, low savings, and higher debt. To integrate multiple indicators, we use this component as an integrated spatial signal to examine broad clustering patterns beyond single-variable analysis.

### 2.3 Spatially constrained regionalization

It is important to notice that while the previous approach incorporates the multidimensional nature of poverty, it ignores possible non-linear interactions between variables, and it does not fix the problem of having fragmented clusters.

To infer geographically coherent "policy regions" and account for the possibly non-linear interactions between the variables, we apply agglomerative hierarchical clustering with a contiguity constraint, ensuring all clusters are spatially continuous. The optimal number of regions is selected by balancing geographic coherence (measured by the isoperimetric quotient (*IPQ*)) and feature

coherence (silhouette[28] and Calinski–Harabasz scores[29]). This approach groups provinces that share similar multidimensional poverty profiles while preserving realistic spatial boundaries—addressing the policy challenge that administrative regions do not always reflect socioeconomic structures. Then, as a simple way to count at the same time for the geographical and feature coherence, we sum the *IPQ* with each of the goodness-of-fit metrics.

### 2.4 Hierarchical and spatial regressions

Once we grouped all provinces of Thailand into geographically coherent regions, to quantify how education contributes to income across the inferred regions, we estimate a Bayesian hierarchical model with region-specific intercepts and either a common or region-specific slope for the effect of education as proposed in.[5] This framework accounts for the nested structure of provinces within regions and permits partial pooling, enhancing estimation stability in smaller regions. Posterior distributions provide credible intervals for income levels and returns to education at both regional and national scales.

As a complementary analysis, we estimate a geographically weighted regression model (GWR), which has been extensively used in spatial analyses in social and economic contexts.[30–32] It allows the relationship between education and income to vary smoothly across geographic space. Whereas the hierarchical model quantifies between-region heterogeneity, GWR identifies fine-scale spatial variation within regions. This comparison helps determine whether regional boundaries align with meaningful spatial variation in returns to education.

## 3 Results

### 3.1 Moran clusters

To identify clusters that incorporate the spatial structure of the data, we need to create a network that reflects the connection between the provinces of Thailand. We have seen in practice that considering the five nearest neighbors to build such network maintains the analysis local, while reproducing the spatial structure of Thailand, as it is illustrated in Figure 1, using the monthly average income of households due to its direct relation with poverty. Furthermore, using these spatial weights, we calculate the spatial lag of the variable, and identify its Moran's clusters, which are also shown in Figure 1. We can observe four well-distinguished clusters. Around the area of Bangkok and Eastern Thailand we observe that provinces form a cluster of high income. The other three clusters at Northern, Northeastern and Southern Thailand are clusters where provinces present low income.

To consider the multidimensional nature of poverty, we reproduce this analysis for other eight variables, representing different aspects of the social problem, such as level of education, income, inequality, debt, and living aspects. Four of these variables (percentage of households with no savings, percentage of households with formal debt, alcohol consumption, and smoking) correspond with variables with the largest percentage of households that reported having an issue. Their Moran's clusters are shown in Figure 8. Overall, we can sketch five different clusters, with the following characteristics:

- **Bangkok Metropolitan Area and Eastern Thailand**: Around Bangkok and, to a lesser extent, Eastern Thailand, we detect a cluster that overlaps for several variables, being a region with high level of education, and low percentage of households in debt, as well as low percentage of alcohol consumption and smoking.
- **Northeastern Thailand**: This region forms a cluster with low inequality, but a large percentage of households in debt.
- **Northern Thailand**: This region forms a cluster where households lack savings, and high levels of alcohol consumption.
- **Southern Thailand**: Similar to the previous region, we detect a cluster with a high percentage of households without savings, and high percentage of smoking.
- **Western Thailand**: This cluster is characterized by high levels of inequality.

While the previous analysis can help us to identify clusters based on their poverty dynamics, it has the disadvantage of considering only one variable at a time, which does not capture the possible interactions between them. To consider a more integrated perspective, we perform a principal component analysis of the standardized variables. Then, we apply our previous approach on the first principal component, detecting three clusters, which are shown on the left side of Figure 2. The first cluster corresponds with most of the provinces of Northern Thailand. The second cluster is formed by the most eastern provinces of Northeastern Thailand. Lastly, we detect a third cluster formed by the provinces around Bangkok, Eastern Thailand and some provinces of Western Thailand.

To get an interpretation for the clusters detected through the principal components, in Table 7 we present the loads of the variables for the first two principal components, and show them graphically on the left side of the biplot in Figure 10. The first principal component can be interpreted mostly as a comparison between variables with an intrinsic positive aspect such as monthly income, yearly savings, and years of education; and variables with an intrinsic negative aspect such as alcohol consumption, percentage of formal debt, and smoking. Thus, we can conclude that the first two clusters identified through the first principal component are more correlated with negative aspects. Meanwhile, the third

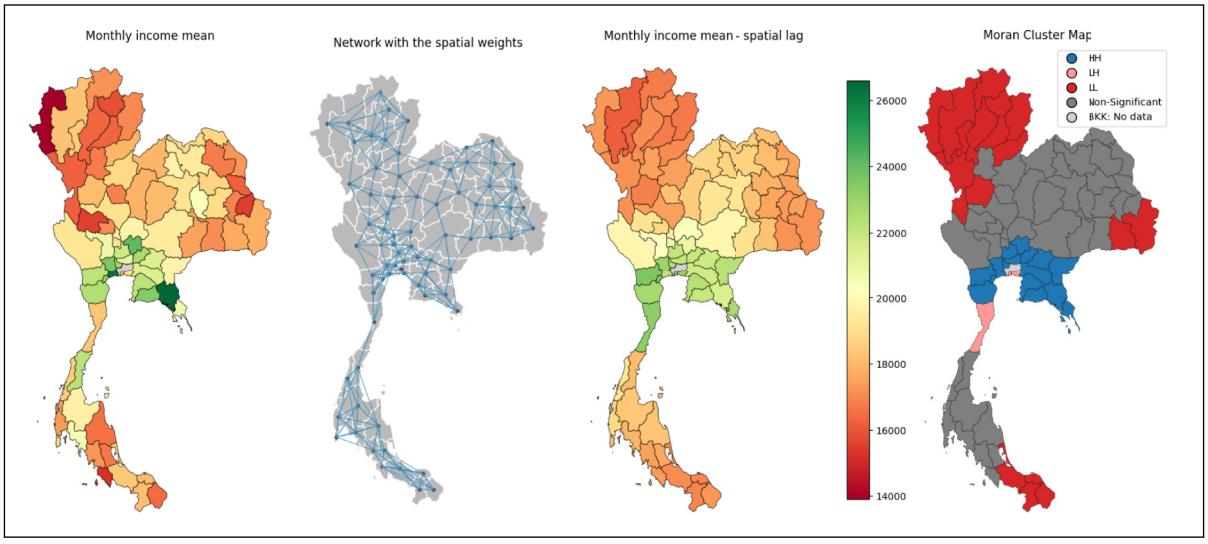


**Figure 1.** From left to right, the maps depict the following information. (i) Monthly average household income (in Thai baht) for each province in Thailand. (ii) The spatial network topology connecting each province to its five nearest neighbors. (iii) The spatial lag of monthly average income, computed as the mean income of each province's five nearest neighbors. (iv) Statistically significant Moran's clusters of income. Four clear clusters emerge: a high-income region centered on Bangkok and three low-income clusters in Northern, Northeastern, and Southern Thailand. Together, these panels illustrate how spatial dependence structures income patterns beyond administrative boundaries, motivating data-driven regionalization.

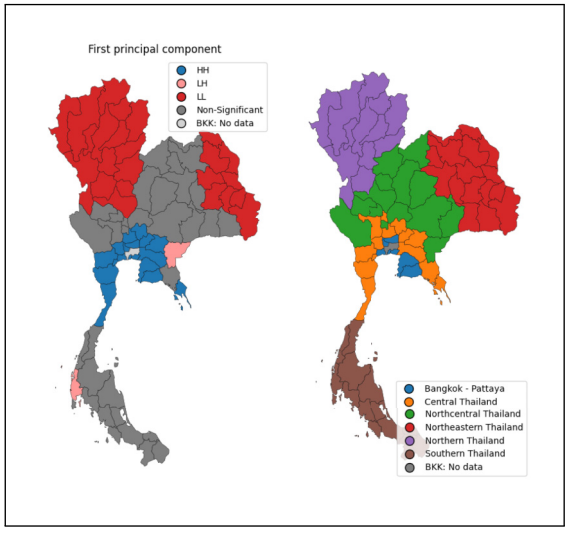


**Figure 2.** Left: Moran's clusters using the first principal component. The first principal component can be interpreted as a comparison between positive aspects such as high levels fo education and income, and negative aspects such as alcohol consumption, smoking, and debt. We have detected three clusters; two clusters at Northern Thailand and Northeastern Thailand correlated with negative aspects, and a third cluster showing a wealthy area around Bangkok and Eastern Thailand. Right: Proposed regions using agglomerative hierarchical clustering, adding a spatial constraint. This technique guarantees the detection of coherent regions, while considering nonlinear interactions between variables. Comparing both panels shows that while Moran's clusters capture broad gradients, the spatially constrained clustering yields contiguous regions suitable for policy design.

cluster, corresponding with the provinces around Bangkok, corresponds to a wealthy region. This helps to guide policy makers to identify regions that need more attention and to establish the high-priority aspects to mitigate.

## 3.2 *Regionalization*

As discussed in Section 2, Moran's clusters present the disadvantage that several provinces do not end into any cluster, as well as ignoring non-linear relations in the variables. To take into consideration these aspects, we fit an agglomerative hierarchical clustering. Identifying 6 different regions in Thailand, which we have named as: Bangkok-Pattaya, Central Thailand, Northcentral Thailand, Northeastern Thailand, Northern Thailand, and Southern Thailand, which are shown on the right side of Figure 2. It is interesting to notice that Northern Thailand and Northeastern Thailand correspond approximately with two of the Moran's clusters for the first principal component. While Bangkok-Pattaya and Central Thailand form, approximately, the third cluster detected with the first principal component.

We observe that the Bangkok-Pattaya region, corresponds with the wealthiest region of Thailand, principally associated with high levels of education, income and savings. And, at the same time, keeping the lowest levels for the rest of the variables, which posse an inherent negative aspect.

On the other hand, while Central Thailand presents similar values to Bangkok-Pattaya for the monthly income and the amount of savings, important differences appear for the

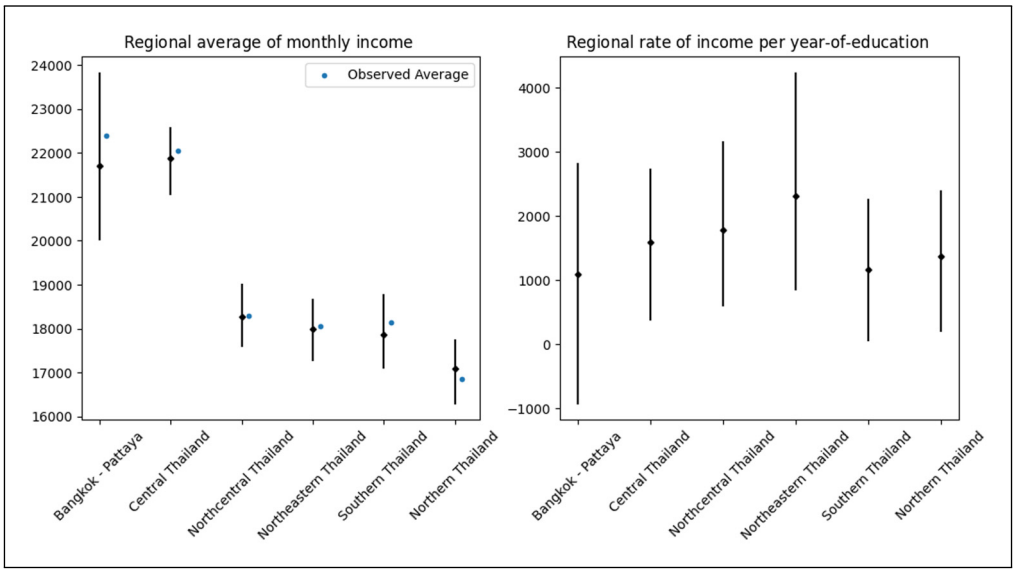


**Figure 3.** Left: Credible intervals of 0.95 posterior probability for the regional average monthly income. We observe that Bangkok-Pattaya and Central Thailand show a similar income, being far higher than the income for the rest of the regions. While Northern Thailand presents the lowest income between all the regions. Right: Credible intervals of 0.95 posterior probability for the regional rate of income per year-of-education (i.e., marginal income associated with one additional year of schooling). The large overlapping for most of the regions might indicate that all the regions share a common rate.

years of education, with a dramatic decrease in the levels of education. We also find Central Thailand to be the region with the largest values of inequality. Thus we can identify this region as a buffer zone of transition from the wealthy region of Bangkok-Pattaya and the rest of the country.

Finally, Northern Thailand, Northeastern Thailand, and to a lesser extent Northcentral Thailand correspond with the regions associated with low levels of education, households without savings, high levels of debt, smoking and households in debt.

### 3.3 Spatial, and Bayesian hierarchical regression

Following the approach proposed in,[5] we fit a hierarchical regression model to estimate the impact of years of education into the income for the proposed regions.

In Figure 3 we present the credible intervals of 0.95 posterior probability for the regional average monthly income and the regional rate of income per year-of-education. Bangkok-Pattaya and Central Thailand show a similar income, being far higher than the income for the rest of the regions. While Northern Thailand presents the lowest income between all the regions. For the rate of income per year-of-education, we observe a large overlapping for most of the regions, which might indicate that all the regions share a common rate. Thus, we might better opt for a hierarchical model that considers one common rate of income per year-of-education. For this purpose, we also follow the model presented by[5] with a common rate of income for all the regions.

In Tables 1 and 2 we present the estimated values for the monthly income and the rate income per year-of-education at a regional and a national levels for the hierarchical models considering different slopes, and a common slope, respectively. Together with the posterior mean of the parameters, we present an interval of 0.95 posterior probability. While we obtain mostly the same estimates for the monthly income, and the rate income per year-of-education at a national level, it is important to note that the posterior interval is slightly smaller when we consider a common slope. On the other hand, when we consider different slopes for the regions, the posterior interval for some of these slopes include negative values, which is very unlikely. Thus, a common slope for all regions seems to be a model that can explain better our phenomenon.

Finally, in Figure 4 we present the regression fitted for each one of the regions. For reference, we have added a vertical dashed line at 12 years of education, which corresponds (approx.) with a complete senior high school. To complement this analysis, in Table 3 we present the average years of education for each region. We observe that Bangkok-Pattaya is the only one whose average years of education is above the 12 years threshold, with most of the provinces in the region above it, with the only exceptions of Chonburi and Rayong which are a little below with 11.98 and 11.93 years, respectively. On the other hand, except for Phuket with 12.79 years of education on average, all the provinces that do not belong to the Bangkok-Pattaya region have an average years of education below the 12 years threshold, while the national average is 10.87 years.

**Table 1.** Posterior mean and interval of 0.95 posterior probability for the hierarchical model that consider different intercepts and slopes. Most of the intervals for the slope present a large overlap, which might indicate a common slope for all the regions. Note that the posterior interval for some of these slopes include negative values (bangkok-pattaya), which is very unlikely. Income is measured in Thai baht per month.

| Region | Monthly income mean | Rate income per year-of-education |
|---|---|---|
| National level | 19004; (18553, 19449) | 1673; (1138, 2267) |
| Bangkok - Pattaya | 21683; (19927, 23550) | 1042; (-1062, 2676) |
| Central Thailand | 21854; (21090, 22520) | 1587; (399, 2803) |
| Northcentral Thailand | 18261; (17522, 18950) | 1777; (559, 3164) |
| Northeastern Thailand | 18057; (17380, 18700) | 2339; (947, 4300) |
| Northern Thailand | 17859; (17166, 18682) | 1181; (47, 2262) |
| Southern Thailand | 17078; (16316; 17727) | 1408; (156, 2445) |

**Table 2.** Posterior mean and interval of 0.95 posterior probability for the hierarchical model that consider different intercepts and a common slope for all the regions. A common slope for all regions seems to be a model that can explain better our data. Income is measured in Thai baht per month.

| Region | Monthly income mean | Rate income per year-of-education |
|---|---|---|
| National level | 19015; (18554, 19485) | 1545; (982, 2059) |
| Bangkok - Pattaya | 21779; (19828, 23567) | |
| Central Thailand | 21852; (20804, 22892) | |
| Northcentral Thailand | 18310; (17638, 18982) | |
| Northeastern Thailand | 18084; (17441, 18754) | |
| Northern Thailand | 18194; (17242, 19167) | |
| Southern Thailand | 16943; (16157; 17722) | |

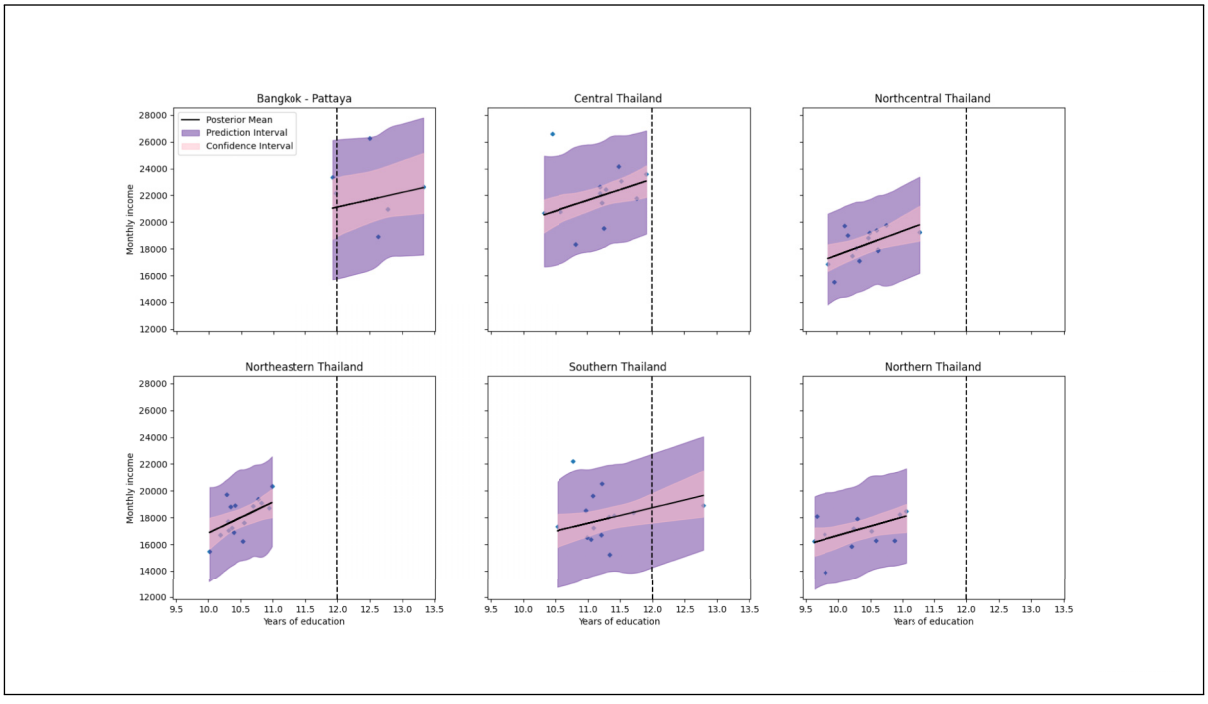


**Figure 4.** Regression fitted for each region. Explaining the average monthly income as function of the years of education. For reference, we have added a vertical dashed line at 12 years of education, which corresponds (approx.) with a complete senior high school.

**Table 3.** Average years of education for each region. Bangkok-pattaya is the only one whose average years of education is above the 12 years threshold, while the national average is of 10.87 years.

| Region | Years of education |
|---|---|
| Bangkok-Pattaya | 12.52 |
| Southern Thailand | 11.26 |
| Central Thailand | 11.16 |
| Northeastern Thailand | 10.49 |
| Northcentral Thailand | 10.42 |
| Northern Thailand | 10.32 |
| Thailand | 10.87 |

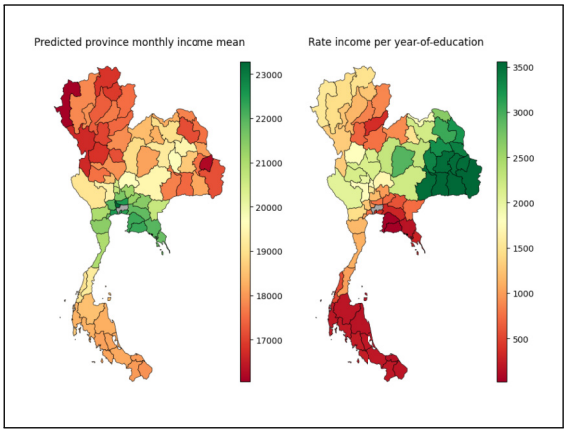


**Figure 5.** Results of the geographically weighted regression (GWR) model. Left: we present the estimated monthly average income for each province, getting similar values to the observed ones. Right: we present the estimated rate income per year-of-education for each province. This figure illustrates within-region spatial heterogeneity that complements the between-region variation captured by the hierarchical model.

Taking into account the spatial weights, we can also fit a geographically weighted regression (GWR) model, explaining the province income in terms of the average years of education in the province. On the left side of Figure 5 we present the estimated monthly average income for each province; while on the right side we present the estimated rate income per year-of-education for each province, it worth notice that there are no negative values for these estimations, even though we did not impose this constraint explicitly in the model.

To evaluate our spatial regression model, we analyze its residuals, without observing any geographical pattern, being an indicative that the model can recover and estimate correctly the geographical pattern of the data. Furthermore, we got a $p$-value of 0.371 for the hypothesis of no spatial correlation in the residuals, which is sufficiently large to consider that the residuals do not present any spatial structure, being a desirable property of a well-fitted spatial regression model.

## 4 Discussion

Thailand's persistent poverty disparities cannot be effectively addressed through a uniform national strategy. While fully customized provincial interventions are impractical due to resource and administrative constraints, policies based on data-driven regional clusters offer a balanced alternative, which allowing interventions to reflect regional realities while maintaining feasible implementation scales.

Our analysis demonstrates that poverty and its associated dimensions (e.g. education, income, inequality, debt, and savings) are spatially dependent and geographically clustered. The consistent positive values of Moran's I across variables confirm that provinces with similar poverty profiles tend to be adjacent, implying that spatial spillovers are a defining feature of Thailand's socioeconomic landscape. Recognizing and incorporating these inter-provincial dependencies is therefore crucial for effective policy design.

By integrating spatial clustering and hierarchical regionalization, we identified six coherent regions that better represent the country's underlying poverty dynamics than administrative boundaries. These inferred regions are both statistically consistent and geographically continuous, enabling targeted analysis of the factors that shape regional deprivation. For instance, Northern and Northeastern Thailand exhibit the strongest spatial clustering of low income and education levels, while Central Thailand shows higher inequality despite higher overall income. In contrast, Bangkok–Pattaya stands out as the only region with average education levels exceeding twelve years, corresponding to complete senior high school.

These patterns reveal that policy effectiveness depends not only on the magnitude of poverty but also on its spatial configuration. Neighboring provinces often face shared structural barriers, such as low educational attainment, limited savings, or high debt exposure, which cannot be resolved through isolated provincial programs. Consequently, policy coordination within each inferred region is essential. For example:

- **Northern and Northeastern Thailand** would benefit most from initiatives emphasizing education expansion and debt management.

- **Central Thailand** requires inequality mitigation through skill upgrading and wage redistribution measures.
- **Southern Thailand** faces behavioral and savings-related constraints, calling for financial literacy and health behavior programs.

Incorporating these spatial insights into policymaking can enhance both efficiency and equity. The data-driven regions identified here can serve as operational policy zones, guiding ministries and local agencies in designing regionally differentiated yet nationally coherent interventions. Moreover, the application of Bayesian hierarchical and geographically weighted regression demonstrates how education contributes to income generation differently across these regions, providing quantitative evidence to support such targeted planning.

Beyond Thailand, this framework highlights the broader potential of spatial statistics as a decision-support tool. It moves beyond traditional descriptive poverty mapping by integrating spatial dependence directly into policy analysis. The approach enables governments to identify functional socioeconomic regions, align resource allocation with actual deprivation patterns, and monitor how spatial inequalities evolve over time.

## 5 Conclusion

This study illustrates that Thailand's poverty exhibits strong spatial dependence—neighboring provinces share similar deprivation structures—making uniform national policies or policies that depend on only administrative regions inefficient. Using spatial autocorrelation, principal component analysis, and hierarchical regionalization, we identify six coherent regions that better represent socioeconomic realities than administrative boundaries. These regions reveal distinct policy priorities: Northern and Northeastern Thailand face low education and income, Central Thailand shows higher inequality, while Southern Thailand struggles with savings and behavioral issues. By applying Bayesian hierarchical and geographically weighted regression, we quantify how education influences income across these spatially defined regions, providing evidence for differentiated interventions. The framework demonstrates how spatial statistics can translate into actionable policy design, supporting regionally coordinated yet locally adapted poverty-alleviation strategies. More broadly, it offers a reproducible, data-driven approach for aligning resource allocation with the true geography of inequality.

### Acknowledgments

Portions of this manuscript were prepared with assistance from a large language model (ChatGPT). All analyses, interpretations, and final revisions were conducted and validated by the authors.

### ORCID iDs

Irving Gómez-Méndez https://orcid.org/0009-0006-5892-8428

Chainarong Amornbunchornvej https://orcid.org/0000-0003-3131-0370

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Code availability

Codes to reproduce our results are available in https://github.com/IrvingGomez/SpatialPovertyFactors

### References

1. Zhou Y and Liu Y. The geography of poverty: Review and research prospects. *J Rural Stud* 2022; 93: 408–416.
2. Amornbunchornvej C, Surasvadi N, Plangprasopchok A, et al. Framework for inferring empirical causal graphs from binary data to support multidimensional poverty analysis. *Heliyon* 2023; 9: e15947.
3. Zhang Y, Yang C, Yan S, et al. Alleviating relative poverty in rural China through a diffusion schema of returning farmer entrepreneurship. *Sustainability* 2023; 15: 1380.
4. Okpala E, Manning L and Baines R. Socio-economic drivers of poverty and food insecurity: Nigeria a case study. *Food Rev Inter* 2023; 39: 3444–3454.
5. Gómez-Méndez I and Amornbunchornvej C. Income, education, and other poverty-related variables: A journey through bayesian hierarchical models. *Heliyon* 2024; 10: e27968.
6. Berdegue J, Escobar G, et al. *Rural diversity, agricultural innovation policies and poverty reduction*. Agricultural Research and Extension Network, 2002.
7. Commins P. Poverty and social exclusion in rural areas: characteristics, processes and research issues. *Sociol Ruralis* 2004; 44: 60–75.
8. Pringle D, Cook S, Poole M, et al. *Cross-border deprivation analysis: a summary guide*. Oak Tree Press, 2000.
9. Amornbunchornvej C, Surasvadi N, Plangprasopchok A, et al. Identifying linear models in multi-resolution population data using minimum description length principle to predict household income. *ACM Trans Knowl Discov Data* 2021; 15: 1–30.
10. Blandford A, Wesson J, Amalberti R, et al. Opportunities and challenges for telehealth within, and beyond, a pandemic. *The Lancet Global Health* 2020; 8: e1364–e1365.
11. Pienkhumtod A, Amornbunchornvei C and Nantharath P. Quantitative analysis of poverty indicators: The case of khon

kaen province, thailand. *J Asian Finance, Econ Business* 2020; 7: 131–141.
12. Alkire S, Kanagaratnam U, Nogales R, et al. Revising the global multidimensional poverty index: Empirical insights and robustness. *Rev Income Wealth* 2022; 68: S347–S384.
13. Alkire S, Kanagaratnam U and Suppa N. The global multidimensional poverty index (MPI) 2021. *OPHI MPI Methodol Note* 2021; 51: 1–39.
14. Alkire S, Roche JM, Ballon P, et al. *Multidimensional poverty measurement and analysis*. Oxford University Press, USA, 2015.
15. Grün B, Leisch F, et al. Applications of finite mixtures of regression models. https://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf 2007.
16. Grün B and Leisch F. Fitting finite mixtures of linear regression models with varying & fixed effects in R. In: *Proceedings in computational statistics*, 2006, pp.853–860. Physica Verlag - Springer.
17. Leisch F. Flexmix: A general framework for finite mixture models and latent class regression in R. *J Stat Software, Articles* 2004; 11: 1–18.
18. Fienberg SE. Bayesian models and methods in public policy and government settings. *Stat Sci* 2011; 26: 212–226.
19. Finucane MM, Martinez I and Cody S. What works for whom? a bayesian approach to channeling big data streams for public program evaluation. *Am J Evalu* 2018; 39: 109–122.
20. Li H, Calder CA and Cressie N. Beyond moran's i: Testing for spatial dependence based on the spatial autoregressive model. *Geograph Anal* 2007; 39: 357–375.
21. Gelman A and Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.
22. Congdon PD. *Bayesian hierarchical models: with applications using R*. CRC Press, 2019.
23. Rey S, Arribas-Bel D and Wolf LJ. *Geographic data science with python*. CRC Press, 2023.
24. Zongfan B, Ling H, Xuhai J, et al. Spatiotemporal evolution of desertification based on integrated remote sensing indices in duolun county, inner mongolia. *Ecol Informat* 2022; 70: 101750.
25. Hou C, Xie Y and Zhang Z. An improved convolutional neural network based indoor localization by using jenks natural breaks algorithm. *China Commun* 2022; 19: 291–301.
26. Khamis N, Sin TC and Hock GC. Segmentation of residential customer load profile in peninsular malaysia using jenks natural breaks. In: *2018 IEEE 7th international conference on power and energy (PECon)*, 2018, pp.128–131. IEEE.
27. Chen J, Yang S, Li H, et al. Research on geographical environment unit division based on the method of natural breaks (jenks). *Inter Archives Photogram Remote Sens Spat Inform Sci* 2013; 40: 47–50.
28. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987; 20: 53–65.
29. Caliński T and Harabasz J. A dendrite method for cluster analysis. *Commun Stat-theory Methods* 1974; 3: 1–27.
30. Fotheringham AS, Brunsdon C and Charlton M. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.
31. Li S, Zhou C, Wang S, et al. Spatial heterogeneity in the determinants of urban form: an analysis of chinese cities with a gwr approach. *Sustainability* 2019; 11: 479.
32. Farahmand S, Sameti M and Salahaldin Sasan S. Spatial variations of $\beta$-convergence coefficient in asia (the gwr approach). *Iran Econ Rev* 2014; 18: 81–101.
33. Moran PA. The interpretation of statistical maps. *J R Stat Soc Ser B (Methodol)* 1948; 10: 243–251.
34. Mendez C and Santos-Marquez F. Regional convergence and spatial dependence across subnational regions of asean: Evidence from satellite nighttime light data. *Regional Sci Policy Pract* 2021; 13: 1750–1777.
35. Mendez C. *Convergence Clubs in Labor Productivity*. Springer, 2020.

# Appendix A—Statistical methods and complementary maps

## A.1 Global Moran's I and Moran's clusters

To identify clusters that incorporate the spatial structure of data, we need to construct a network that resembles the topological connections. A typical approach to building such network is to connect each province $i$ with its $k$-nearest neighbors, knn($i$),[30] creating weights of the form

$$w_{ij} = \begin{cases} 1/k & \text{if province } j \in \text{knn}(i), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Using these spatial weights, we can calculate the spatial lag of a variable for province $i$, defined as

$$\text{lag-}y_i = \sum_{j=1}^{n} w_{ij} y_j, \quad (2)$$

where $y_i$ is the observed value of the attribute for the province $i$.

To determine if there is statistical evidence of spatial structure for the attribute, we can estimate its spatial autocorrelation. When the variable is geographically distributed such that high values are nearby other high values and low values are nearby other low values, we say that the variable has a positive spatial autocorrelation. On the other hand, when the attribute is distributed such that high and low values are close, we say that the variable presents a negative spatial autocorrelation. Probably, the most commonly used

statistic to estimate the spatial autocorrelation of a variable is Moran's I[33] (see for example[34,35]), which is given by

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j / S_0}{\sum_{\ell=1}^{n} z_\ell^2 / n}, \tag{3}$$

where $z_i = y_i - \bar{y}$, and $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. It can be proved that, under absence of spatial autocorrelation, the expected value of Moran's I is

$$\mathbb{E}(I) = -\frac{1}{n-1}. \tag{4}$$

On the other hand, since $\sum_{j=1}^{n} w_{ij} = 1$, then $S_0 = n$. And, Moran's I statistic can be rewritten as

$$\begin{aligned} I &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{\ell=1}^{n} z_\ell^2} \\ &= \frac{\sum_{i=1}^{n} z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{\ell=1}^{n} z_\ell^2} \\ &= \frac{\sum_{i=1}^{n} z_i \text{lag-}z_i}{\sum_{\ell=1}^{n} z_\ell^2}, \end{aligned} \tag{5}$$

which corresponds with the least squares estimator of the slope in the linear regression

$$\text{lag-}z = \beta_0 + \beta_1 z + \varepsilon. \tag{6}$$

Also note that, because $z_i = y_i - \bar{y}$, then $\hat{\beta}_0 = 0$.

Therefore, we can create a scatter plot in which the variable of interest is displayed against its spatial lag, whose estimated slope by least squares corresponds with its Moran's I. This graph is known as Moran's plot or Moran's scatter plot.

Since, by definition, the spatial lag is a weighted average of the $k$-nearest neighbors of each observation, then the top-right quadrant in Moran's plot presents observations with an attribute above the average whose neighbors' attribute is also above the average, and corresponding with a positive spatial autocorrelation. Thus, we name this quadrant as high-high (HH). Similarly, we can name the other quandrants as low-high (LH) for the top-left, low-low (LL) for the bottom-left, and high-low (HL) for the bottom right.

Consider the expression of Moran's I statistic given by equation 5, and note that it can be expressed as

$$\begin{aligned} I &= \frac{1}{n} \frac{\sum_{i=1}^{n} z_i \text{lag-}z_i}{\sum_{\ell=1}^{n} z_\ell^2 / n} \\ &= \frac{1}{n} \sum_{i=1}^{n} I_i, \end{aligned} \tag{7}$$

where

$$I_i = \frac{z_i \text{lag-}z_i}{\sum_{\ell=1}^{n} z_\ell^2 / n} \tag{8}$$

is known as the local Moran's I for the observation $i$.

UsingMoran's plot and local Moran's I, we can create four clusters– namely: HH, HL, LL, and LH –form by observations whose local Moran's I is significantly different from its expected value under the hypothesis of no spatial correlation, depending on the quadrant where the observation is located in the Moran's plot. These clusters are usually named Moran's clusters.

To estimate the statistical significance of Moran's I, we permute the variable of interest 1000 times, calculating the statistic for each permutation, which gives us a reference distribution in the absence of spatial structure. On the left side of Figure 6 we present such reference distribution, we also show its expected value in absence of spatial structure, and the observed value for the Moran's I of the monthly income mean. Since the I statistic is positive, we can conclude that provinces with similar income tend to be

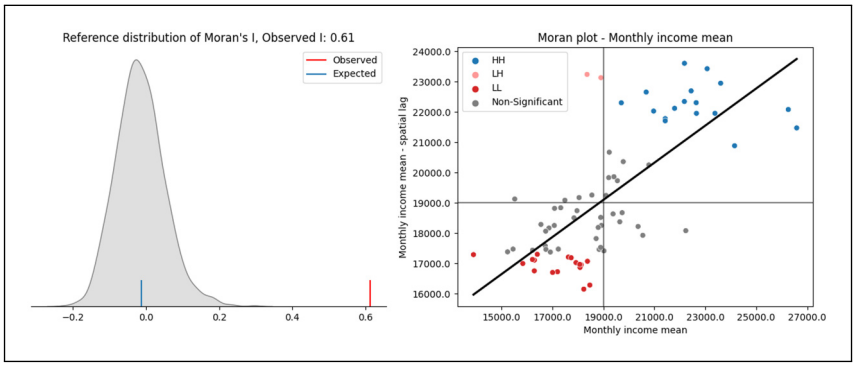


**Figure 6.** Left: Reference distribution for Moran's I in the absence of spatial autocorrelation for the monthly average income. It shows that the observed Moran's I is far from what would be expected under no spatial correlation. Thus, we can conclude that average monthly income presents a significant positive spatial correlation. Right: Moran's plot, the observations have been colored if their local Moran's I is significantly different from its expected value in the absence of spatial correlation.

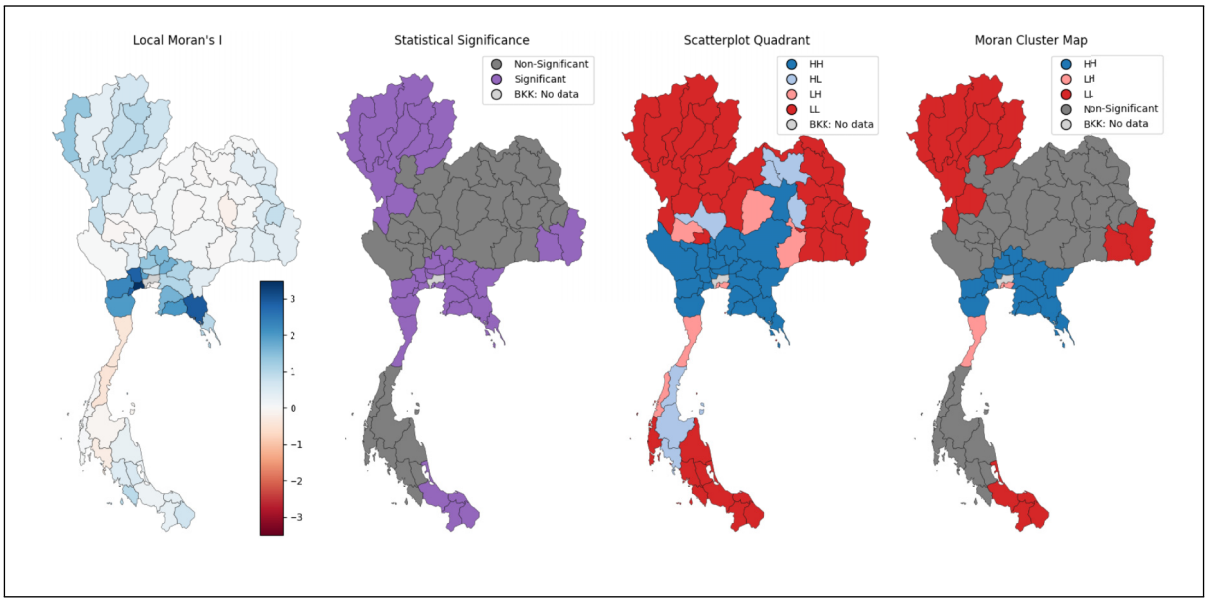


**Figure 7.** Methodology to determine Moran's clusters for the average monthly income. From left to right, the maps depicts the following information. First: We present the local Moran's I statistic for each province. Second: We present the provinces for which we reject the hypothesis of non-spatial autocorrelation. Third: We colored the provinces accordingly to the quadrant where they belong in Moran's plot. Fourth: We keep those provinces whose local Moran's I is significantly different from its expected value under no-spatial autocorrelation.

close, creating possible clusters of provinces with similar income.

On the right side of Figure 6, we present Moran's scatter plot which presents the average monthly income for each province against its corresponding spatial lag. We have colored those observations where we rejected the absence of spatial autocorrelation considering their local Moran's I statistic and a significance level of 0.05.

In Figure 7 we present the clusters found for the average monthly income using the local I statistics. From left to right, the maps present: the value of each local Moran's I statistic, those provinces whose local I is statistically different from its expected value under no spatial correlation assumption, the quadrant where each province belongs in the Moran's scatter plot, and Moran's clusters.

We found a positive value for the Moran's I statistic for all the variables analyzed in this study, which are reported in Table 4, and reject the hypothesis of no spatial structure . To estimate the years of education, we followed the approach proposed in,[5] where the number of years of education is assigned accordingly to the rule presented in Table 5. We present the clusters identified for each variable in Figure 8.

**Table 4.** Poverty factors, and their respective moran's I statistic. All the factors present a positive spatial correlation, which is significantly different from the expected value under the no spatial correlation assumption. Thus, we can conclude that close provinces share similar values and poverty dynamics.

| Variable | Aspect | Moran's I |
|---|---|---|
| Years of education | Education | 0.64 |
| Monthly income | Income | 0.61 |
| Yearly savings | | 0.25 |
| Percentage of households without savings | | 0.56 |
| Monthly income ratio 20:20 | Inequality | 0.47 |
| Gini index | | 0.31 |
| Percentage of households with formal debt | Debt | 0.72 |
| Alcohol consumption | Living aspect | 0.53 |
| Smoking | | 0.68 |

**Table 5.** Estimated years of education for each educational grade, as proposed by.[5]

| Education | Years of education |
|---|---|
| Uneducated | 0 |
| Kindergarten | 0 |
| Pre-elementary school | 3 |
| Elementary school | 6 |
| Junior high school | 9 |
| Senior high school | 12 |
| Vocational degree | 14 |
| Bachelor degree | 16 |
| Post-graduate | 19 |

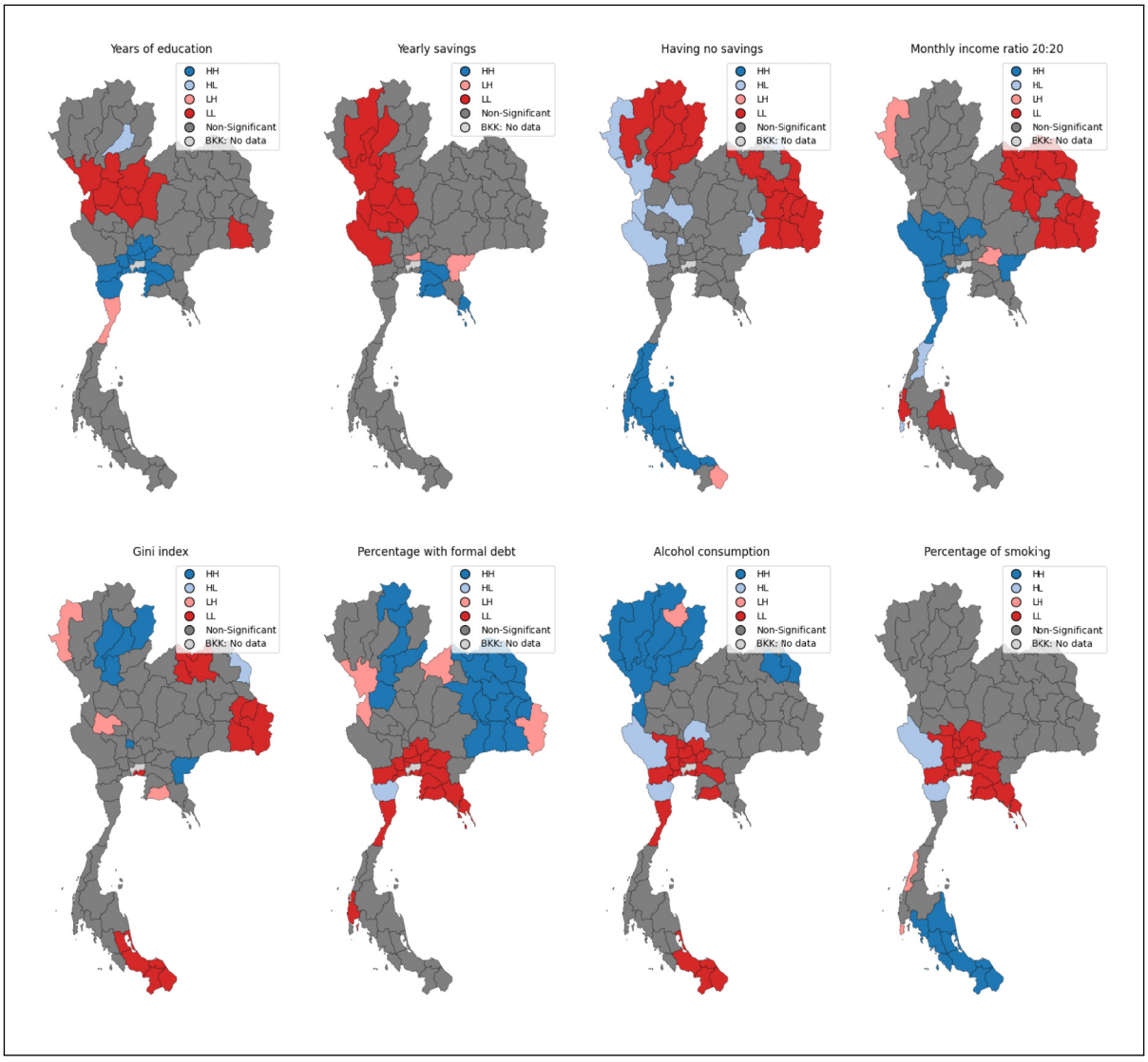


**Figure 8.** Moran's clusters for the considered poverty factors. Note that this methodology can only be applied to one variable at a time, which makes challenging the detection of clusters that consider the multidimensional nature of proverty.

## A.2 Regionalization

Regionalizationmethods are clustering techniques that impose a spatial constraint on clusters. In other words, the result of a regionalization algorithm contains clusters with areas that are geographically coherent, in addition to having coherent data profiles.

For the geographic coherence we consider the isoperimetric quotient of the regions, $IPQ = 4\pi A/L^2$, where $A$ is its area and $L$ its perimeter. According to the isoperimetric inequality, $IPQ \leq 1$, and $IPQ = 1$ if and only if we consider a circumference, while non-convex shapes ("wormy" shapes) would have a low $IPQ$. Thus, we try to achieve a high $IPQ$ for each region in order to achieve geographic coherence.

On the other hand, to measure the feature coherence of the regions we consider two metrics: the silhouette score[28] and the Calinski-Harabasz score.[29] Both of them tried to maintain the feature coherence comparing the variance within the regions against the variance between them.

The average value over the regions for the isoperimetric quotient ($IPQ$), the silhouette score, and the Calinski-Harabasz score are presented on the left side of Figure 9, against the number of regions, which vary from 2 to 9 regions, normalizing the metrics, so the maximum takes the value of 1 while the minimum takes the value of 0. We observe that while the $IPQ$ tends to increase with the number of regions, the silhouette score and the Calinski-Harabasz score tend to decrease. Thus, it is necessary to make a trade-off between the geographic and the feature coherence. As a simple way to count at the same time for the geographical and feature coherence, we sum the $IPQ$ with each of the goodness-of-fit metrics. According to these new metrics, the optimum number of regions is around 6 or 7, as depicted on the right side of Figure 9.

## A.3 Principal component analysis

In Table 6 we present the cummulative percentage of the variance explained by the principal components. To get an interpretation for the clusters detected through principal components, in Table 7 we present the loads of the variables for the first two principal components. These loads are presented graphically on the left side of the biplot in Figure 10. While on the right side we present the average value of the two principal components for the detected regions.

To complementthe profile of each region, in Figure 11 we present the radar chart for each one of the identified regions, normalizing the variables. Providing a whole profile which can help us to identify the principal issues faced for each region.

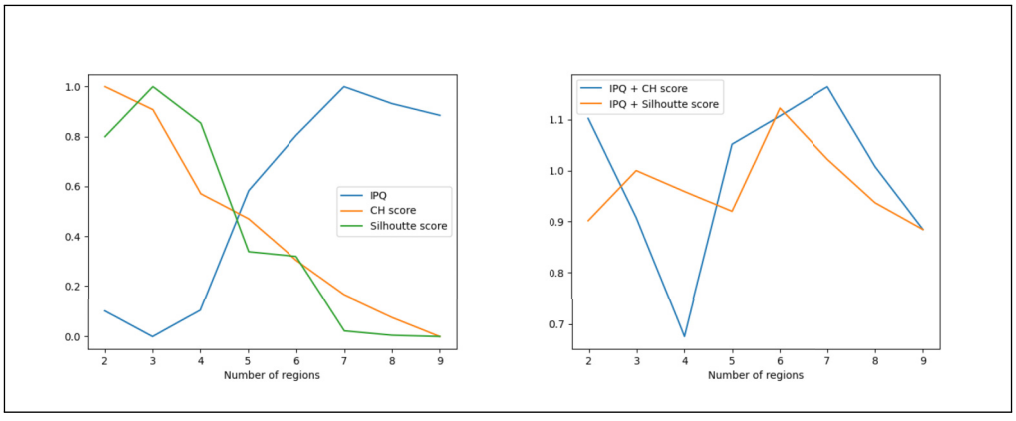


**Figure 9.** Left: We present the *IPQ* score, the silhouette score and the Calinski-Harabasz score as functions of the number of regions considered, we have normalized these scores to be between zero and one. Right: We consider the *IPQ* plus each one of the feature coherence score to account for both, the spatial coherence and the goodness-of-fit. According to this analysis, the optimum number of regions to preserve coherence is around six or seven.

**Table 6.** Cummulative percentage of variance explained by the principal components.

| $PC_1$ | $PC_2$ | $PC_3$ | $PC_4$ | $PC_5$ | $PC_6$ | $PC_7$ | $PC_8$ | $PC_9$ |
|---|---|---|---|---|---|---|---|---|
| 38% | 61% | 74% | 86% | 90% | 94% | 97% | 98% | 100% |

**Table 7.** Loads of the poverty factors for the first two principal components. Note that the first principal component is mainly a comparison between inherently positive aspects such as years of education and monthly income, and inherently negative aspects such as alcohol consumption, smoking, and debt. The second principal component is more challenging to be interpreted, but it gives the largest loads to variables related with inequality and income.

| Variable | $PC_1$ | $PC_2$ |
|---|---|---|
| Years of education | 0.51 | −0.10 |
| Monthly income | 0.40 | 0.33 |
| Yearly savings | 0.24 | 0.13 |
| Percentage of households without savings | 0.34 | −0.42 |
| Monthly income ratio 20:20 | 0.09 | 0.43 |
| Gini index | −0.02 | 0.61 |
| Percentage of households with formal debt | −0.39 | 0.04 |
| Alcohol consumption | −0.45 | 0.11 |
| Smoking | −0.21 | −0.34 |

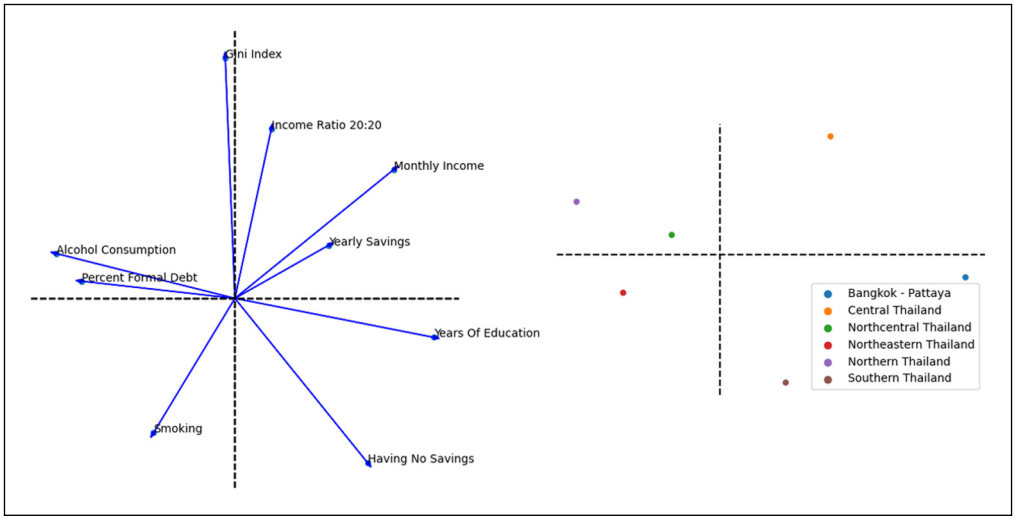


**Figure 10.** Biplot considering the first two principal components. Left: The loads of the variables are depicted as vectors in the cartesian coordinate form by the two first principal components. Right: We present the average value of the first two principal components for each one of the proposed regions.

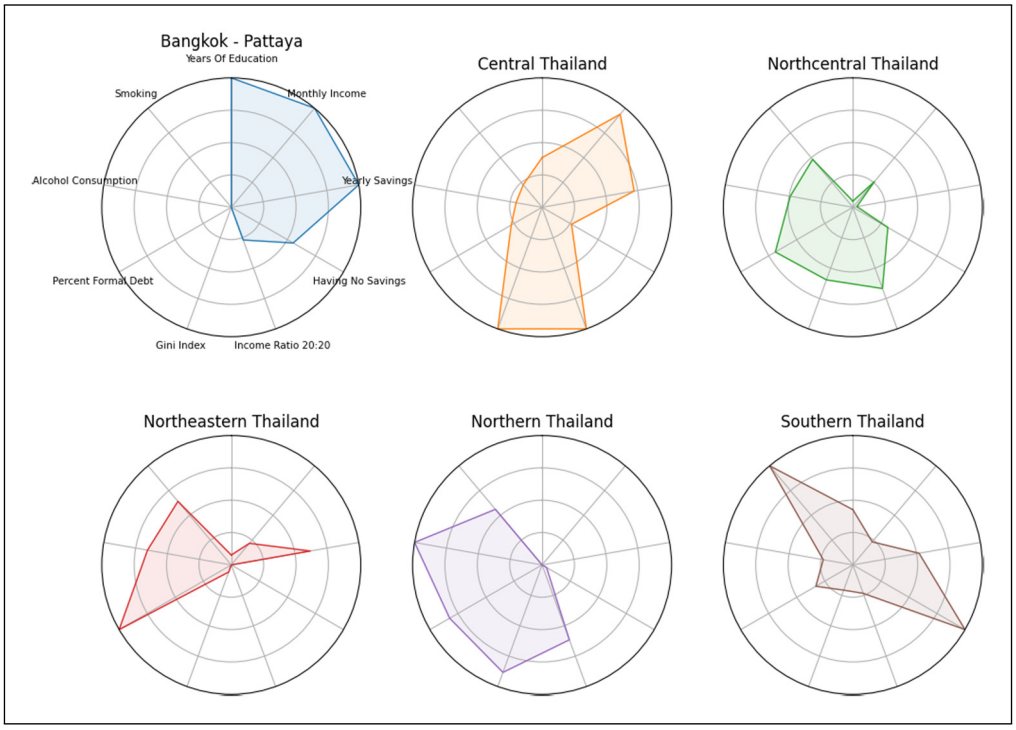


**Figure 11.** Radar chart for each one of the identified regions, normalizing the variables. Providing a whole profile which can help us to identify the principal issues faced for each region (all the variables have been normalized).