

New approach for estimating socioeconomic indicators in small-areas with sparse samples and high within-region heterogeneity: application in the Peruvian Amazon

Jose Matta-Chuquisapon

jose.matta@unmsm.edu.pe

National University of San Marcos

Julianne Meisner

University of Washington

Sergio E. Recuenco

National University of San Marcos

Article

Keywords: socioeconomic disparities, inequalities, small-area estimation, poverty, education, malaria

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10844989/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

New approach for estimating socioeconomic indicators in small-areas with sparse samples and high within-region heterogeneity: application in the Peruvian Amazon

Jose Matta-Chuquisapon^{a*}, Julianne Meisner^{b,c,d}, Sergio E. Recuenco^e

a. Doctoral Programme in Health Sciences, Faculty of Medicine, Universidad Nacional Mayor de San Marcos, Lima, Perú

b. Department of Global Health, School of Public Health, University of Washington, Seattle, WA, USA

c. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA

d. Center for One Health Research, School of Public Health, University of Washington, Seattle, WA, USA

e. Department of Preventive Medicine and Public Health, Faculty of Medicine, Universidad Nacional Mayor de San Marcos, Lima, Perú

* Correspondence: Jose Matta-Chuquisapon, MSc: jose.matta@unmsm.edu.pe

Abstract

Demographic and Health Surveys (DHS) and national censuses lack the spatial resolution and temporal consistency needed for fine-scale monitoring of socioeconomic inequalities, particularly in settings with difficult geographic access and highly dispersed populations, and substantial within-region heterogeneity.

Using education and poverty indicators derived from the DHS and census-derived auxiliary data, this study developed Bayesian small-area estimation models to generate annual district-level socioeconomic estimates over an 11-year period in highly dispersed regions. This study further illustrates the utility of these estimates by characterizing the spatial relationship between socioeconomic indicators and malaria burden in the Peruvian Amazon, where malaria remains a major public health challenge.

The selected models demonstrated robust predictive performance, producing consistent annual district-level estimates and extending estimates to unsampled areas, with strong agreement in cross-scale validation. The spatial analysis revealed heterogeneous district-level clusters where malaria incidence and socioeconomic indicators co-occurred, highlighting spatial patterns that may be masked by traditional analyses conducted at broader geographic scales. These findings demonstrate that Bayesian small-area estimation models using DHS and census-derived auxiliary data can generate fine-scale, temporally stable socioeconomic estimates in small areas with sparse samples and high within-region heterogeneity, improving the assessment of spatial socioeconomic inequalities in data-limited settings.

Keywords: socioeconomic disparities, inequalities, small-area estimation, poverty, education, malaria

1. Introduction

Education and wealth are key determinants of human wellbeing and central targets of the United Nations' (UN) 2030 Sustainable Development Goals (SDGs), with Goal 4 aiming for universal primary education and Goal 1 for the eradication of extreme poverty (1). Nationally, these indicators are monitored through censuses and Demographic and Health Surveys (DHS), but both approaches are limited by infrequent data collection and insufficient representativeness at subnational levels (2). According to the UN Statistics Division, national censuses should be conducted at least every 10 years, with higher frequency recommended in settings of rapid population change (3).

While ensuring comparability across administrative levels, census data cannot be updated annually (4). Consequently, intercensal projections are commonly used, despite their limitations. By contrast, DHS provide frequent socioeconomic data through complex sampling designs ensuring national and subnational representativeness (2,5). These surveys enable intercensal updates of key indicators, but their estimates are typically representative only at predefined administrative levels, potentially masking heterogeneity at finer spatial scales. In addition to monitoring SDG progress, socioeconomic indicators –particularly education and poverty– are key determinants of health outcomes, including vector-borne diseases such as malaria (6–10). Capturing this relationship at the local level requires socioeconomic data with finer spatial resolution than DHS and finer temporal resolution than censuses, particularly in regions with dispersed populations.

Amazon Basin countries (Peru, Brazil, Venezuela, Colombia, Bolivia, Ecuador, Guyana and Suriname) share highly heterogeneous and spatially dispersed populations, making fine-scale monitoring of socioeconomic indicators especially challenging in this region. Limitations in the periodicity of censuses and DHS in areas with dispersed populations make it difficult to monitor changes in socioeconomic indicators over time at the lowest administrative level. This lack of granular socioeconomic information in turn limits our ability to understand how local social environments influence the dynamics of diseases in these settings, including malaria. To address these limitations, we applied Bayesian small-area methods to DHS and census data to generate annual district-level socioeconomic estimates in the Peruvian Amazon, a setting representative of highly dispersed Amazonian populations. We then illustrate the utility of these estimates by

characterizing the spatial clustering between malaria incidence and the socioeconomic estimates over the study period.

2. Results

Sample characteristics

The analysis of DHS data covered an 11-year period (2010–2020) across the 53 districts of the Loreto in the Peruvian Amazon, corresponding to a total of 1,183 cluster–year combinations and 583 district–year combinations. The spatial–temporal distribution of all cluster–year combinations are shown in Fig 1a. The number of household clusters (primary sampling units, or PSUs) sampled during the 2010–2014 period ranged from 40 to 50 PSUs to represent the entire Loreto department. Starting in 2015, the survey sampling frame changed, and more than 100 PSUs per survey were included to obtain representative estimates. Nevertheless, it is important to note that although the number of PSUs increased, the number of sampled districts did not increase substantially (2010–2014: average of 22 districts; 2015–2020: average of 30 districts). District-level sociodemographic information was obtained from the 2017 National Population Census, which provides complete coverage of all 53 districts in Loreto and was treated as time-invariant over the study period.

Direct estimates

Direct district-year estimates were calculated for districts included in DHS sampling, accounting for the survey sampling design using the survey package in R (11). Although there was a moderate increase in the absolute number of sampled districts, the relative coverage of the total districts in Loreto increased from approximately 35% (19/53 districts) of districts in 2010 to 49% (26/53 districts) in 2016 and 66% (35/53 districts) in 2020.

Regarding education, more than 50% of the population had achieved at least elementary education. However, some districts, such as Iquitos, the capital of Loreto, reach levels close to 90%. In contrast, poverty levels were consistently high across most sampled districts, with Iquitos and a few neighboring districts as exceptions. In several districts, direct poverty estimates reached the upper bound of the parameter space (equal to 1). The small number of sampled households per cluster within districts resulted in unstable direct estimators, with near-zero variance estimates and consequently overly narrow confidence intervals, limiting their reliability for area-level inference.

Small area estimation

Pre-selection of candidate models

Goodness-of-fit metrics for candidate models are summarized in supplementary results 3.1. As a first step in the model selection process, models were ranked primarily according to WAIC (Widely Applicable Information Criterion), where lower values indicate better expected out-of-sample predictive fit, with DIC (Deviance Information Criterion) and $-\log ML$ (negative log marginal likelihood) used as secondary criteria to resolve close competition. Based on these metrics, the initial evaluation identified models Me-4 and Me-3 as the two best-performing models for elementary education, and models Mp-1 and Mp-2 for poverty.

Model performance

In a second step, the predictive performance of the candidates models was evaluated using complementary in-sample accuracy and predictive metrics. Fig 2 compares model-based predictions with direct estimates for elementary education and poverty status across years and summarizes RMSE (root mean squared error), MAE (mean absolute error), and SMAPE (symmetric mean absolute percentage error) values for each model. Across both indicators, error metrics varied within narrow ranges over time, and differences across models were small relative to their interannual variability. However, the differences in the annual MAE and RMSE estimates between the periods 2010–2014 and 2015–2020 are noteworthy for both indicators. Given that both indicators are proportions bounded between 0 and 1, RMSE and MAE values in the range of 0.05–0.07 correspond to average absolute deviations between 5 and 7 percentage points, indicating high predictive accuracy at the district-year level.

To further characterize prediction error, Fig 3 presents the MSE (mean squared error) and its decomposition into bias squared and variance. Although MSE is expressed on the squared scale of the outcome, its decomposition is informative for distinguishing systematic error from random variability. In the case of education, all models exhibited similar bias and variance values, with Me-3 and Me-4 showing slightly lower variance and bias. In the case of the poverty indicator, models Mp-1 and Mp-2 exhibited notably higher variance components, while their bias remained similar to that of the other models. This translates into substantially wider confidence intervals, despite point estimates being broadly comparable across models. Mp-3 and Mp-4 showed lower variance; however, Mp-4, which had the lowest variance component, produced the

narrowest confidence intervals. Overall, these patterns indicate meaningful differences in uncertainty across models, primarily driven by differences in the variance component rather than systematic bias.

Predictive ability was additionally assessed using conditional predictive ordinate (CPO) values, which evaluate how well each observation is predicted when it is left out of the model fitting process. Fig 3c summarizes the $-2 \log(\text{CPO})$, where lower values indicate better predictive performance. Although the differences between models were small, Me-4 (5819.82) showed the best predictive performance for primary education, followed by Me-3 (5820.08) and Me-2 (5831.19), while Mp-1 (4999.726) showed the best predictive performance for poverty, followed by Mp-3 (5003.455) and Mp-4 (5003.955).

Taken together, MAE, RMSE, SMAPE, MSE, and $-2 \log(\text{CPO})$ indicate that all proposed models yield accurate and stable predictions, with no evidence of substantial overfitting or systematic bias, thereby providing a robust basis for final model selection.

Model validation

Because DHS does not provide district-level direct estimates for many district-year combinations due to its sampling design, validation was conducted indirectly by aggregating district-level model-based predictions to the departmental level, and comparing these aggregates with corresponding departmental estimates obtained directly from the DHS survey (Fig 4). Aggregate sociodemographic indicators were computed through resampling from the posterior distribution of district-level predictions, thereby propagating uncertainty into the departmental-level estimates.

In the case of elementary education, a particular pattern is observed. For all models, during the 2010–2014 period, the aggregated values were nearly identical to the direct DHS estimates. However, from 2015 onwards, all models showed values that were, on average, below the DHS estimates by less than 5% (<0.05 on a 0–1 scale), while preserving the same overall trend: a decline in 2016, a progressive increase up to 2019, and a plateau in 2020. In the case of poverty, unlike elementary education, the aggregated predictions from all models remained nearly identical to the DHS estimates throughout the entire study period.

As shown in Fig 4, despite the discrepancy observed for elementary education during the 2015–2020 period, these differences are small (0.04–0.05) compared to the observed temporal variation and to the differences between model-based and direct departmental

estimates. In the context of highly heterogeneous settings with limited data availability, where direct district-level estimates are not available, discrepancies of this magnitude indicate strong cross-scale consistency with DHS departmental benchmarks and good predictive performance in unsampled districts.

Final model selection

Based on model fit, predictive performance, and cross-scale consistency, Me-3 and Mp-3 were selected as the final models for elementary education and poverty status, respectively. Model selection was guided by WAIC, $-2 \log \text{CPO}$, and agreement between district-level and departmental estimates. Although these models did not consistently achieve the best performance across all metrics, differences relative to the best-performing alternatives were negligible ($\text{SMAPE} < 0.1$), indicating similar predictive accuracy across specifications. For elementary education, where Me-3 and Me-4 showed nearly identical performance, the simpler model was preferred to enhance parsimony and interpretability. In the case of poverty, Mp-3 was not the best-performing model across all metrics, but it was the most consistent across all the evaluated metrics. In addition, the confidence intervals of the district-level predictions from its competitors, Mp-1 and Mp-4, were either too wide or, conversely, too narrow. These final models were subsequently used to generate district-level estimates and to support the spatial analysis of the relationship between socioeconomic indicators and malaria incidence. Fig 5 shows the direct estimates and model-based predictions for each district in 2010, 2012, 2014, 2016, 2018 and 2020, including unsampled areas, while predictions for all years are provided in Supplementary Fig 3.2.

Spatial analysis of the relationship between socioeconomic indicators and malaria incidence

Fig 1b shows the division of Loreto into its provinces (administrative level-2) and presents both provincial- and district-level patterns of malaria incidence from 2010 to 2020. Malaria incidence peaked between 2013 and 2018, with the highest burden concentrated in the border provinces of Datem del Marañon and Loreto, which over the entire study period (2010–2020) recorded cumulative incidences of 172.15 and 120.81 cases per 1,000 inhabitants, respectively. At the district level (Fig 1c), this pattern is reflected within these same high-burden provinces: districts in the northern zone—particularly those located along the borders with Ecuador and Colombia—show a clear upward trend, with a pronounced intensification between 2013 and 2018. In contrast,

districts in the southern part of the region consistently reported lower incidence levels throughout the study period.

Figs 6 and 7 show bivariate maps describing the distribution at the district level by year of elementary education and poverty status and malaria incidence, respectively. For elementary education and malaria incidence, two large and well-differentiated areas are consistently observed over time. In the north of the department, districts have a medium-to-high malaria incidence and a low proportion of individuals with at least primary education, which evolves to an intermediate level from 2018 onwards. In contrast, the south shows districts with low malaria incidence and a high proportion of individuals with at least primary education, a pattern that has expanded toward the center of the department since 2017. However, in the eastern part of the department, there are districts with a high incidence of malaria and a high proportion of at least primary education, mainly between 2014 and 2017 (Fig 6). With respect to the poverty status and malaria incidence, the northern part of the department forms a consistent cluster of districts with high malaria incidence and medium-to-high levels of poverty. In contrast, the rest of the department shows a more heterogeneous pattern, with some districts experiencing medium-to-high malaria incidence and low poverty, while others exhibit high poverty but low malaria incidence (Fig 7).

To further assess these spatial patterns, we performed a bivariate LISA analysis between each socioeconomic indicator and malaria incidence, using a conditional permutation test (999 simulations), and pseudo p-value <0.05 for significance. For elementary education, Low-Low clusters were identified from 2010-2014, which progressively shifted to High-Low clusters (high malaria incidence and low primary-education coverage) in the northern area bordering Ecuador. For poverty, a Low-Low cluster (low malaria incidence and low poverty) was identified in central Loreto, where the regional capital, Iquitos, is located. Notably, an adjacent district (Alto Nanay district) displayed a High-Low pattern, with high malaria incidence despite low poverty levels.

3. Discussion

Despite the recognized need to monitor socioeconomic inequalities, the lack of high-resolution, temporally consistent socioeconomic data, especially in hard-to-reach areas where disease is endemic, such as the Amazon, has limited systematic monitoring. In this study, for the first time in the Americas, we bridge this gap by producing annual, subnational socioeconomic estimates for highly dispersed and heterogeneous

populations using Bayesian small-area estimation models that combine national census-derived socioeconomic indices with household survey data. These estimates demonstrate consistently strong temporal predictive performance, enabling reliable longitudinal monitoring across the region. Building on this granularity, we applied these estimates to reveal the fine-scale spatial patterning of malaria burden and its systematic co-occurrence with socioeconomic disadvantage across an 11-year period in the Peruvian Amazon.

The first major contribution of this study is that the spatiotemporal Bayesian models generated district-level socioeconomic estimates for all years, including in districts that were never sampled during the study period. In contrast to traditional approaches that often rely on auxiliary data sources such as satellite or geospatial information to predict in non-sampled areas (26–28), our models produced stable estimates even in years with less than 50% sampling coverage. This framework is readily extensible to a wide range of sociodemographic indicators collected in DHS surveys, including morbidity, under-five mortality, and health service quality (29–32).

Second, we demonstrate that strong predictive performance in non-sampled areas can be achieved using only national census data as auxiliary information. While most existing approaches rely on richer auxiliary sources such as administrative records or satellite-derived covariates (26,31,33,34), we show that robust and temporally consistent spatial predictions are feasible even when relying on a single, widely available data source. This is particularly relevant for settings where alternative high-resolution data are scarce or inconsistent over time. Importantly, this approach is not specific to Peru but is highly transferable to other Amazonian countries, which share similar patterns of population dispersion and sociodemographic heterogeneity (35,36).

The estimates generated by this study directly support monitoring of Sustainable Development Goals 1 (No Poverty) and 4 (Quality Education) at the district level, on an annual basis in a region where official disaggregated progress assessments are rarely available (19,37). In Loreto specifically, these models revealed district-level trajectories of education and poverty between 2010 and 2020, where previously only department-level trends from DHS data were available (16). In our analysis, socioeconomic indicators evolved heterogeneously across districts, with some showing sustained improvement while others experienced minimal gains. This fine-scale evidence can

267 directly inform the strategic allocation of resources and the design of targeted
268 interventions in the Peruvian Amazon.

269 Using these district-level estimates, our LISA analysis identified clusters where high
270 malaria incidence co-occurs with low levels of education, as well as clusters of low
271 malaria incidence that co-occur with low levels of poverty, patterns consistent with
272 previous research (7,9,38–40). This demonstrates the ability of these models to detect
273 heterogeneity in the distribution of health inequalities on an annual basis in areas where
274 analysis is traditionally conducted at the national level, leading to the masking of such
275 variations.

276 This study has some limitations. First, in 2015, a change in the DHS sampling design,
277 including the number of PSUs, the weighting scheme, and the districts covered,
278 introduced systematic noise in district-level predictions from 2015 onward, likely resulting
279 in underestimates for affected districts. However, cross-scale validation at the
280 departmental level showed that differences did not exceed 5%, suggesting that the
281 overall patterns remained interpretable. Second, the poverty status variable was derived
282 from the DHS wealth index. While some components are directly observable, others rely
283 on self-reported asset ownership, which may be subject to underreporting. This limitation
284 is common in household surveys, and wealth index-based estimates have been shown
285 to be consistent with census data, allowing results to remain comparable. Third, this
286 study relied exclusively on the 2017 Peruvian national census, as it was the only census
287 available within the study period. Under this constraint, the burden of estimation falls
288 primarily on the random effects, which were still able to capture the temporal variability
289 observed in the DHS direct estimates. Fourth, estimates for 2020 should be interpreted
290 with caution. The COVID-19 pandemic disrupted both the malaria surveillance system,
291 through the temporary closure of primary healthcare facilities, and DHS data collection,
292 which shifted to telephone interviews. These conditions may have affected the
293 completeness of case reporting and the representativeness of the survey data.

294 In conclusion, this study applied a Bayesian small-area estimation framework to
295 generate annual, subnational socioeconomic indicators in small areas with sparse
296 samples and high within-region heterogeneity using exclusively census-derived auxiliary
297 data. Our models provide finer spatial resolution than DHS surveys and finer temporal
298 resolution than census data. By enabling fine-scale and temporally stable inference

beyond survey coverage, this approach improves the assessment of spatial socioeconomic inequalities in data-limited settings.

4. Methods

4.1. Study area and design

Peru is a country located in South America and is divided into 24 departments (administrative level-1) (12). Loreto, located in the northeastern region of the Peruvian Amazon, is the largest department in Peru by area (~ 374 000 km²), and contains 53 districts (administrative level-3) (13). Loreto exhibits considerable geographical and socio-demographic heterogeneity, with a highly dispersed human population, and is also endemic for several vector-borne diseases such as malaria and dengue (14). Further details on the characteristics of Loreto can be found in Supplementary information 1.1. This study included all 53 districts in Loreto, for the period of 2010-2020.

We conducted a retrospective ecological study at the district level to generate annual district-level estimates of poverty status and elementary education, which were subsequently applied to characterize their spatial clustering with malaria incidence. Bayesian hierarchical small area estimation models were fitted using the Integrated Nested Laplace Approximation (INLA). Primary sampling units (PSUs), defined as clusters of households in the DHS, were aggregated to obtain cluster-level number of cases (y) and total number of households (n) by districts and years, which were used as inputs (15,16). District-level random effects were included to produce model-based predictions for each district-year combination with covariates from the 2017 Peruvian census incorporated to improve predictions. Finally, malaria case data from the Peruvian public health surveillance system were integrated with small-area model predictions to generate the spatial outputs used in the LISA analysis (17).

4.2. Ethics declarations

This study is a product of a doctoral dissertation approved by the Comité de Ética en Investigación, part of the Instituto de Ética en Salud of the Facultad de Medicina of the Universidad Nacional Mayor de San Marcos, under study number N° 0073-2025, which also waived the requirement for informed consent because this study used secondary data that had been anonymized. All methods were performed in accordance with the relevant guidelines and regulations. This study has satisfactorily complied with the recommendations regarding scientific, technical, and ethical aspects for research involving human subjects.

4.3. Data sources and definition of variables

4.3.1. Poverty and elementary education

Poverty status and education were obtained from the DHS surveys conducted in Peru between 2010 and 2020. The DHS is a nationally representative survey conducted annually to measure demographic dynamics, maternal and child health, socioeconomic conditions, and factors associated with communicable and non-communicable diseases (16). It uses a complex two-stage, probabilistic, stratified, balanced, and independent sampling design that allows precise estimations at national, rural/urban, and departmental levels (16).

Wealth and education variables from the DHS were recategorized into dichotomous variables. Poverty status was defined as households in the first two quintiles (quintile 1-2) and elementary education was defined as households where at least one member has completed primary schooling, aligning with SDG targets and WHO wealth metrics (18,19), and UNESCO and World Bank indicators (20,21). Both binary variables were aggregated at the cluster level for the 2010–2020 period, resulting in the total number of households per PSU (n) and the number of events (y) for each PSU-year.

4.3.2. Predictors

Predictors were obtained from the Peru's National Population and Housing Census, conducted approximately every 10 years (most recently in 2017), providing statistical, demographic, social, and economic information at the district level (22). Variables from the housing, household, and population modules hypothesized to be associated with education and poverty were included, capturing domains such as basic housing services, construction materials, native language, educational attainment, employment status, and health insurance coverage. After transforming multi-category variables into dummy variables, a final set of 94 variables was obtained from an initial selection of 31. The complete list of national census variables included in the study is shown in Supplementary information 1.3.

Principal component analysis (PCA) was applied to district-level proportions to reduce dimensionality and address multicollinearity. The first 10 components explaining more than 80% of the total variance were retained and evaluated for inclusion as fixed-effects (pc_1 – pc_{10}) in the Bayesian SAE models. More information about PCA applied in Supplementary methods 2.1.

4.3.3. Malaria data

Malaria cases from 2010 to 2020 were obtained from the Peruvian public health surveillance system of the National Center for Epidemiology and Disease Control of the Peruvian Ministry of Health (MINSA) (17). This system consolidates all reports of malaria cases at the individual level from different health centers throughout the country. We aggregated individual malaria cases reported in the national surveillance system between 2010 and 2020 to the district level. Population data from WorldPop (23) were used to calculate incidence rates for each district throughout the study period.

4.4. Statistical Analysis

4.4.1. Descriptive assessment of sampled household clusters in the DHS

The georeferenced locations of household clusters were obtained from the DHS survey for the period 2010–2020 and plotted on a base map of Loreto. This mapping of the location and frequency of sampled clusters was conducted to descriptively assess coverage, temporal availability, and changes in the number of sampled households by district.

4.4.2. Cluster-level small area estimation models for socioeconomic indicators

Model structure

Bayesian hierarchical generalized linear models with spatio-temporal structure were fitted for each socioeconomic indicator using the Integrated Nested Laplace Approximation (INLA) framework.

Elementary education

A Bayesian hierarchical model with a Binomial likelihood was specified for the education outcome. Let y_{it} denote the number of households with at least one member having completed elementary education in primary sampling unit (PSU) i at time t , and n_{it} the corresponding total number of surveyed households within each PSU-year, where i indexes PSUs and t indexes year (2010–2020). The model is specified as:

$$y_{it}|p_{it} \sim \text{Binomial}(n_{it}, p_{it})$$

where $\mu_{it} = E(p_{it} | \xi, v, w)$ is the expected prevalence of elementary education completion given all random effects. The mean is linked to a linear predictor through a logit transformation:

$$\text{logit}(\mu_{it}) = \eta_{it}$$

The linear predictor included spatial and temporal random effects and was progressively extended to evaluate alternative model specifications. Model development followed a sequential strategy. All model variants shared a common baseline structure consisting of a spatial effect $\xi(s_i)$ modeled as a Gaussian random field (GRF) with Matérn covariance structure (SPDE approach) and a temporal effect v_t modeled as a second-order random walk (RW2). Alternative specifications were obtained by (i) including or excluding socioeconomic covariates through principal components $pc_{k,it}$, and (ii) including or excluding a province-level independent and identically distributed (iid) random effect $w_{p(i)}$.

The linear predictor is specified as:

$$\eta_{it} = \beta_0 + \sum_{k=1}^K \beta_k pc_{k,it} + \xi(s_i) + w_{p(i)} + v_t$$

Where:

- β_0 is the intercept
- $pc_{k,it}$ is the value of the k -th principal component for PSU i at time t .
- $\xi(s_i)$ is a spatial random effect modeled as a Gaussian random field (GRF) with Matérn covariance structure, approximated via the SPDE approach fitted on a triangular mesh over the Loreto department
- v_t is a temporal random effect modeled as a second-order random walk (RW2), allowing for smooth nonlinear temporal variation over the study period
- $w_{p(i)}$ is a province-level iid random effect capturing residual heterogeneity at an intermediate spatial scale (included only in selected model specifications)

Poverty status

A Bayesian hierarchical Beta-Binomial model was specified to account for overdispersion in clustered binary outcomes. Let y_{it} denote the number of households classified as poor in primary sampling unit (PSU) i at time t , and n_{it} the corresponding total number of surveyed households within each PSU-year, where i indexes PSUs and t indexes year (2010–2020). The model is specified hierarchically as:

$$p_{it} \sim \text{Beta}(\alpha_{it}, \beta_{it}),$$

$$y_{it} | p_{it} \sim \text{Binomial}(n_{it}, p_{it})$$

with $\alpha_{it} = \mu_{it}\phi$ and $\beta_{it} = (1 - \mu_{it})\phi$, where $\mu_{it} = E(p_{it} | u, v, w)$ is the expected poverty prevalence given all random effects and $\phi > 0$ is a precision parameter controlling overdispersion. The mean is linked to a linear predictor through a logit transformation:

$$\text{logit}(\mu_{it}) = \eta_{it},$$

The linear predictor included spatial and temporal random effects and was progressively extended to evaluate alternative model specifications. Model development followed a sequential strategy. All model variants shared a common baseline structure consisting of a district-level spatial effect $u_{s(i)}$ modeled using the BYM2 model and a temporal effect v_t modeled as a second-order random walk (RW2). Alternative specifications were obtained by (i) including or excluding socioeconomic covariates through principal components $pc_{k,it}$, derived from census-based variables, and (ii) including or excluding a province-level independent and identically distributed (iid) random effect $w_{p(i)}$. The linear predictor is specified as:

$$\eta_{it} = \beta_0 + \sum_{k=1}^K \beta_k pc_{k,it} + u_{s(i)} + w_{p(i)} + v_t$$

Where:

- β_0 is the intercept
- $pc_{k,it}$ is the value of the k -th principal component for PSU i at time t .
- $u_{s(i)}$ is a district-level spatial random effect defined over districts in the Loreto department, modeled using the BYM2 formulation, combining a structured component defined through an intrinsic conditional autoregressive (ICAR) model and an unstructured independent and identically distributed (iid) component capturing area-level heterogeneity
- v_t is a temporal random effect modeled as a second-order random walk (RW2), allowing for smooth nonlinear temporal variation over the study period
- $w_{p(i)}$ is a province-level iid random effect capturing residual heterogeneity at an intermediate spatial scale (included only in selected model specifications)

This specification allows for overdispersion relative to the standard binomial model while producing smoothed estimates at the district level through the hierarchical spatial structure. We considered four model variants based on alternative combinations of the socioeconomic covariates and the province-level random effect.

Although all models are fitted to PSU-level data, predictions are reported at the district level. In SPDE-based models, predictions are derived by evaluating the latent spatial field at district centroids, whereas in the BYM2 specification spatial effects are defined directly at the district level. Full details of prior distributions and random effects specifications are provided in Supplementary Methods 2.2.

Model selection

A total of 4 models were fitted for each socioeconomic indicator: models Me-1 – Me-4 for elementary education and models Mp-1 – Mp-4 for poverty status. These models differed in the number of principal components included and the specification of the spatio-temporal random effects. The selection of the best models was based on the Watanabe–Akaike information criterion (WAIC), the deviance information criterion (DIC) and the negative log Marginal Likelihood (-logML). In-sample accuracy was evaluated using metrics including mean absolute error (MAE), root mean squared error (RMSE) and symmetric mean absolute percentage error (SMAPE).

Model validation

To assess predictive performance, the conditional predictive ordinate (CPO) was used as an efficient approximation to leave-one-out cross-validation, computed directly from INLA without refitting the model. In addition, the mean squared error (MSE) was decomposed into squared bias and variance components to evaluate the sources of prediction error and further characterize the predictive accuracy of the models.

Finally, we conducted a cross-scale validation by aggregating district-level predictions to the departmental level for poverty and elementary education. These were obtained by resampling district-level posterior predictions from the proposed models and contrasting them with the direct departmental estimates from DHS described above. The MAE was then calculated to quantify the differences between the two sets of estimates. All data cleaning, manipulation, and analysis were performed using R version 4.4.2 (24).

4.4.3. Descriptive analyses and local spatial autocorrelation (LISA) of socioeconomic indicators and malaria incidence

We mapped the distribution of malaria cases as incidence rates per 1,000 inhabitants for each district and year. We then produced district-level bivariate maps combining the socioeconomic estimates obtained from the best-performing model selected in the previous step with malaria incidence, using the *biscale* package in R (25), in order to

486 explore their joint spatial distribution across the 2010–2020 period. Finally, we assessed
487 spatial dependence between each socioeconomic indicator and malaria incidence by
488 computing a local bivariate Moran's I statistics for each year of the study period. Local
489 indicators of spatial association were estimated for each district and year using spatial
490 weights based on first-order queen contiguity, allowing the identification of High–High,
491 Low–Low, High–Low, and Low–High spatial clusters.

Data availability

National census data are available free of charge upon request from the Peruvian National Institute of Statistics and Informatics (<https://censo2017.inei.gob.pe/resultados-definitivos-de-los-censos-nacionales-2017/>). The Peruvian DHS microdata (Demographic and Family Health Survey—ENDES) are publicly available through the open data platform of the National Institute of Statistics and Informatics (<https://proyectos.inei.gob.pe/microdatos/>), and the code to download them directly is provided in the repository of this paper. Malaria epidemiological surveillance data are available from Peru's Centro Nacional de Epidemiología, Prevención y Control de Enfermedades through Peru's National Open Data Platform (<https://www.datosabiertos.gob.pe/dataset/vigilancia-epidemiologica-de-malaria>); access may be restricted outside Peru.

Data are also accessible through the Peruvian Standard Transparency Portal (https://www.transparencia.gob.pe/enlaces/pte_transparencia_enlaces.aspx?id_entidad=143). All code used to download, clean, and analyze the data, as well as to generate the figures and tables presented in this study, is available at https://github.com/JoseMatta95/SAE_INLA_Loreto.

5. References

1. Measuring and monitoring progress towards the SDGs | UNECE [Internet]. [cited 2025 Sep 15]. Available from: <https://unece.org/measuring-and-monitoring-progress-towards-sdgs>
2. The DHS Program - Quality information to plan, monitor and improve population, health, and nutrition programs [Internet]. [cited 2025 Sep 6]. Available from: <https://dhsprogram.com/>
3. United Nations. Handbook on Registers-based Population and Housing Censuses [Internet]. 2022. Available from: <https://unstats.un.org/unsd/demographic-social/publication/handbook-registers-phc.pdf>
4. Skinner C. Issues and Challenges in Census Taking. *Annu Rev Stat Its Appl*. 2018 Mar 7;5(Volume 5, 2018):49–63. doi:10.1146/annurev-statistics-041715-033713
5. United Nations. Designing Household Survey Samples: Practical Guidelines [Internet]. 2005. Available from: <https://unstats.un.org/unsd/demographic/sources/surveys/handbook23june05.pdf>
6. World Health Organization. World malaria report 2023 [Internet]. 2023 [cited 2024 Feb 2]. Available from: <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023>
7. Carrasco-Escobar G, Fornace K, Benmarhnia T. Mapping socioeconomic inequalities in malaria in Sub-Sahara African countries. *Sci Rep*. 2021 Jul 23;11(1):15121. doi:10.1038/s41598-021-94601-x PubMed PMID: 34302015; PubMed Central PMCID: PMC8302762.
8. Were V, Buff AM, Desai M, Kariuki S, Samuels A, ter Kuile FO, et al. Socioeconomic health inequality in malaria indicators in rural western Kenya: evidence from a household malaria survey on burden and care-seeking behaviour. *Malar J*. 2018 Apr 16;17(1):166. doi:10.1186/s12936-018-2319-0
9. Apeagyei AE, Patel NK, Cogswell I, O'Rourke K, Tsakalos G, Dieleman J. Examining geographical inequalities for malaria outcomes and spending on malaria in 40 malaria-endemic countries, 2010-2020. *Malar J*. 2024 Jul 10;23(1):206. doi:10.1186/s12936-024-05028-4 PubMed PMID: 38982498; PubMed Central PMCID: PMC11234708.
10. Verma AK, Shrivastava S, Saxena N. Malaria and malnutrition in tribal areas of India: implications for paediatric health. *Front Gastroenterol*. 2025;4:1526806. doi:10.3389/fgstr.2025.1526806 PubMed PMID: 41822271; PubMed Central PMCID: PMC12952398.
11. Lumley T, Gao P, Schneider B. survey: Analysis of Complex Survey Samples [Internet]. 2025 [cited 2025 Aug 29]. Available from: <https://cran.r-project.org/web/packages/survey/index.html>
12. Perú: información territorial departamental [Internet]. 2023 [cited 2025 Jun 19]. Available from: <https://www.gob.pe/institucion/pcm/campa%C3%B1as/4107-peru-informacion-territorial-departamental>

- 550 13. Loreto: información territorial [Internet]. 2021 [cited 2025 Jun 19]. Available from:
551 [https://www.gob.pe/institucion/pcm/campa%C3%B1as/4292-loreto-informacion-](https://www.gob.pe/institucion/pcm/campa%C3%B1as/4292-loreto-informacion-territorial)
552 territorial
- 553 14. PAHO and the Regional Health Management of Loreto committed to the elimination of
554 trachoma - PAHO/WHO | Pan American Health Organization [Internet]. 2024 [cited
555 2025 Aug 20]. Available from: [https://www.paho.org/en/news/23-1-2024-paho-and-](https://www.paho.org/en/news/23-1-2024-paho-and-regional-health-management-loreto-committed-elimination-trachoma)
556 regional-health-management-loreto-committed-elimination-trachoma
- 557 15. R-INLA Project [Internet]. [cited 2025 Jun 21]. Available from: [https://www.r-](https://www.r-inla.org/home)
558 inla.org/home
- 559 16. Encuesta Demográfica y de Salud Familiar - ENDES 2025 [Internet]. 2025 [cited 2025
560 Jun 21]. Available from: [https://www.gob.pe/institucion/inei/informes-](https://www.gob.pe/institucion/inei/informes-publicaciones/3364616-encuesta-demografica-y-de-salud-familiar-endes-2025)
561 publicaciones/3364616-encuesta-demografica-y-de-salud-familiar-endes-2025
- 562 17. Vigilancia epidemiológica de Malaria | Plataforma Nacional de Datos Abiertos
563 [Internet]. [cited 2025 Jul 4]. Available from:
564 <https://datosabiertos.gob.pe/dataset/vigilancia-epidemiol%C3%B3gica-de-malaria>
- 565 18. WHO. Health Inequality Monitor dataset [Internet]. [cited 2025 Jul 4]. Available from:
566 <https://www.who.int/data/sets/health-inequality-monitor-dataset>
- 567 19. UNDP [Internet]. [cited 2025 Jul 4]. Objetivo 1: Fin de la pobreza | Programa De Las
568 Naciones Unidas Para El Desarrollo. Available from:
569 <https://www.undp.org/es/sustainable-development-goals/fin-pobreza>
- 570 20. Completion rate (primary education, lower secondary education, upper secondary
571 education) [Internet]. 2020 [cited 2025 Jul 4]. Available from:
572 [https://uis.unesco.org/en/glossary-term/completion-rate-primary-education-lower-](https://uis.unesco.org/en/glossary-term/completion-rate-primary-education-lower-secondary-education-upper-secondary-education)
573 secondary-education-upper-secondary-education
- 574 21. Glossary | DataBank [Internet]. [cited 2025 Jul 4]. Available from:
575 <https://databank.worldbank.org/metadataglossary/jobs/series/SE.PRM.CMPT.ZS>
- 576 22. Censos Nacionales 2017 – XII de Población, VII de Vivienda y III de Comunidades
577 Indígenas [Internet]. 2017 [cited 2025 Jun 19]. Available from:
578 <https://censo2017.inei.gob.pe/>
- 579 23. Google for Developers [Internet]. [cited 2026 May 5]. WorldPop Global Project
580 Population Data: Estimated Residential Population per 100x100m Grid Square | Earth
581 Engine Data Catalog. Available from: [https://developers.google.com/earth-](https://developers.google.com/earth-engine/datasets/catalog/WorldPop_GP_100m_pop?hl=es-419)
582 engine/datasets/catalog/WorldPop_GP_100m_pop?hl=es-419
- 583 24. Posit team. RStudio: Integrated Development Environment for R. Posit Software, PBC,
584 Boston, MA. Posit [Internet]. 2024 [cited 2024 Jul 3]. Available from:
585 <http://www.posit.co/>
- 586 25. Prener C, Grossenbacher T, Zehr A, Stevens J. biscale: Tools and Palettes for
587 Bivariate Thematic Mapping [Internet]. 2022 [cited 2025 Jul 5]. Available from:
588 <https://cran.r-project.org/web/packages/biscale/index.html>

- 589 26. Permatasari N, Ubaidillah A. Small Area Estimation of poverty using remote sensing
590 data. *Stat J IAOS*. 2025 Mar 1;41(1):180–90. doi:10.1177/18747655241308390
- 591 27. Sikov A, Cerda-Hernandez J. Prediction in non-sampled areas under spatial small area
592 models. *Stat Methods Appl*. 2024 Sep;33(4):1079–116. doi:10.1007/s10260-024-
593 00754-0
- 594 28. Masaki T, Newhouse D, Silwal AR, Bedada A, Engstrom R. Small Area Estimation of
595 Non-Monetary Poverty with Geospatial Data. Vol. 38. 2022;38(3). doi:10.3233/SJI-
596 210902
- 597 29. Simkin J, Dummer TJB, Erickson AC, Otterstatter MC, Woods RR, Ogilvie G. Small
598 area disease mapping of cancer incidence in British Columbia using Bayesian spatial
599 models and the smallareamapp R Package. *Front Oncol*. 2022;12:833265.
600 doi:10.3389/fonc.2022.833265 PubMed PMID: 36338766; PubMed Central PMCID:
601 PMC9627310.
- 602 30. Wang H, Smits L, Putrik P. Spatial variation in tobacco smoking among pregnant
603 women in South Limburg, the Netherlands, 2016-2018: Small area estimations using a
604 Bayesian approach. *Spat Spatio-Temporal Epidemiol*. 2022 Aug;42:100525.
605 doi:10.1016/j.sste.2022.100525 PubMed PMID: 35934326.
- 606 31. Gascoigne C, Smith T, Paige J, Wakefield J. Estimating subnational under-five
607 mortality rates using a spatio-temporal Age-Period-Cohort model. *Spat Spatio-
608 Temporal Epidemiol*. 2025 Feb 1;52:100708. doi:10.1016/j.sste.2024.100708
- 609 32. Allorant A, Fullman N, Leslie HH, Sarr M, Gueye D, Eliakimu E, et al. A small area
610 model to assess temporal trends and sub-national disparities in healthcare quality. *Nat
611 Commun*. 2023 Jul 28;14(1):4555. doi:10.1038/s41467-023-40234-9
- 612 33. United Nations Statistic Division. Small Area Estimation with Geospatial Data: A Primer
613 [Internet]. 2025. Available from:
614 [https://unstats.un.org/UNSDWebsite/statcom/session_56/documents/BG-3p-
615 Geospatial_SAE_Primer-E.pdf](https://unstats.un.org/UNSDWebsite/statcom/session_56/documents/BG-3p-Geospatial_SAE_Primer-E.pdf)
- 616 34. Burgard JP, Morales D, Wölwer AL. Small area estimation of socioeconomic indicators
617 for sampled and unsampled domains. *ASTA Adv Stat Anal*. 2022 Jun 1;106(2):287–
618 314. doi:10.1007/s10182-021-00426-4
- 619 35. Pereira ASA de P, dos Santos VJ, Alves S do C, Amaral e Silva A, da Silva CG, Calijuri
620 ML. Contribution of rural settlements to the deforestation dynamics in the Legal
621 Amazon. *Land Use Policy*. 2022 Apr 1;115:106039.
622 doi:10.1016/j.landusepol.2022.106039
- 623 36. Assunção J, Rocha R. Rural Settlements And Deforestation in The Amazon [working
624 paper]. *Clim Policy Initiat*. 2016.
- 625 37. UNDP [Internet]. [cited 2025 Sep 6]. Objetivo 4: Educación de calidad | Objetivos de
626 Desarrollo Sostenible | Programa De Las Naciones Unidas Para El Desarrollo.
627 Available from: [https://www.undp.org/es/sustainable-development-goals/educacion-
628 calidad](https://www.undp.org/es/sustainable-development-goals/educacion-calidad)

- 629 38. Okova D, Lukwa AT, Edusei M, Bodzo P, Atta-Obeng C, Chiwire P, et al. Between-and
630 within-socioeconomic groups temporal inequality in the uptake of malaria prevention
631 strategies among pregnant women and under-five children in Ghana (2003-2022).
632 Malar J. 2025 Aug 22;24(1):273. doi:10.1186/s12936-025-05512-5 PubMed PMID:
633 40847346; PubMed Central PMCID: PMC12372236.
- 634 39. Gunderson AK, Kumar RE, Recalde-Coronel C, Vasco LE, Valle-Campos A, Mena CF,
635 et al. Malaria Transmission and Spillover across the Peru–Ecuador Border: A
636 Spatiotemporal Analysis. Int J Environ Res Public Health. 2020 Jan;17(20):7434.
637 doi:10.3390/ijerph17207434
- 638 40. Krisher LK, Krisher J, Ambuludi M, Arichabala A, Beltrán-Ayala E, Navarrete P, et al.
639 Successful malaria elimination in the Ecuador–Peru border region: epidemiology and
640 lessons learned. Malar J. 2016 Nov 28;15:573. doi:10.1186/s12936-016-1630-x
641 PubMed PMID: 27894320; PubMed Central PMCID: PMC5126842.

642

643

Funding

This work was funded by the Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) and the Programa Nacional de Investigación Científica y Estudios Avanzados (PROCIENCIA) within the framework of the E077-2023-01-BM “Becas en Programas de Doctorado en Alianzas Interinstitucionales” contest, grant number (PE501090537-2024) and the E033-2023-01-BM “Alianzas Interinstitucionales para Programas de Doctorado” contest, grant number (PE501084306-2023).

AI use statement

No AI-assisted technologies were used in the preparation of this manuscript.

Author contributions

Conceptualization: J.M.-C, S.R. | Data curation: J.M.-C | Formal analysis: J.M.-C., S.R., J.M. | Methodology: J.M.-C., S.R. J.M. | Writing – original draft: J.M.-C., S.R. J.M. | Writing – review & editing: J.M.-C., S.R. J.M.

Competing interest

The authors declare that they have no competing interests

659 **Figures**

660 **Figure 1.** Direct estimates of education and poverty and malaria incidence in Loreto. a. Map of the department of Loreto showing the location of the
661 PSUs across the territory. The stratified distribution is displayed, distinguishing whether a PSU was sampled in 2010–2014 or 2015–2020. b. On the
662 left, a map of Loreto showing the districts (administrative level 3), and their grouping into the region's eight provinces (administrative level 2). On the
663 right, the cumulative malaria incidence over the years (2010–2020), stratified by each province of Loreto. c. Distribution of malaria incidence per
664 thousand inhabitants in the department of Loreto between 2010 and 2020. The incidence is shown using a log10 transformation to smooth values
665 close to zero. Districts that did not report any cases of malaria are shown in grey. For panel c, values are displayed at two-year intervals for clarity;
666 the full annual series is provided in the supplementary figure 3.4.
667

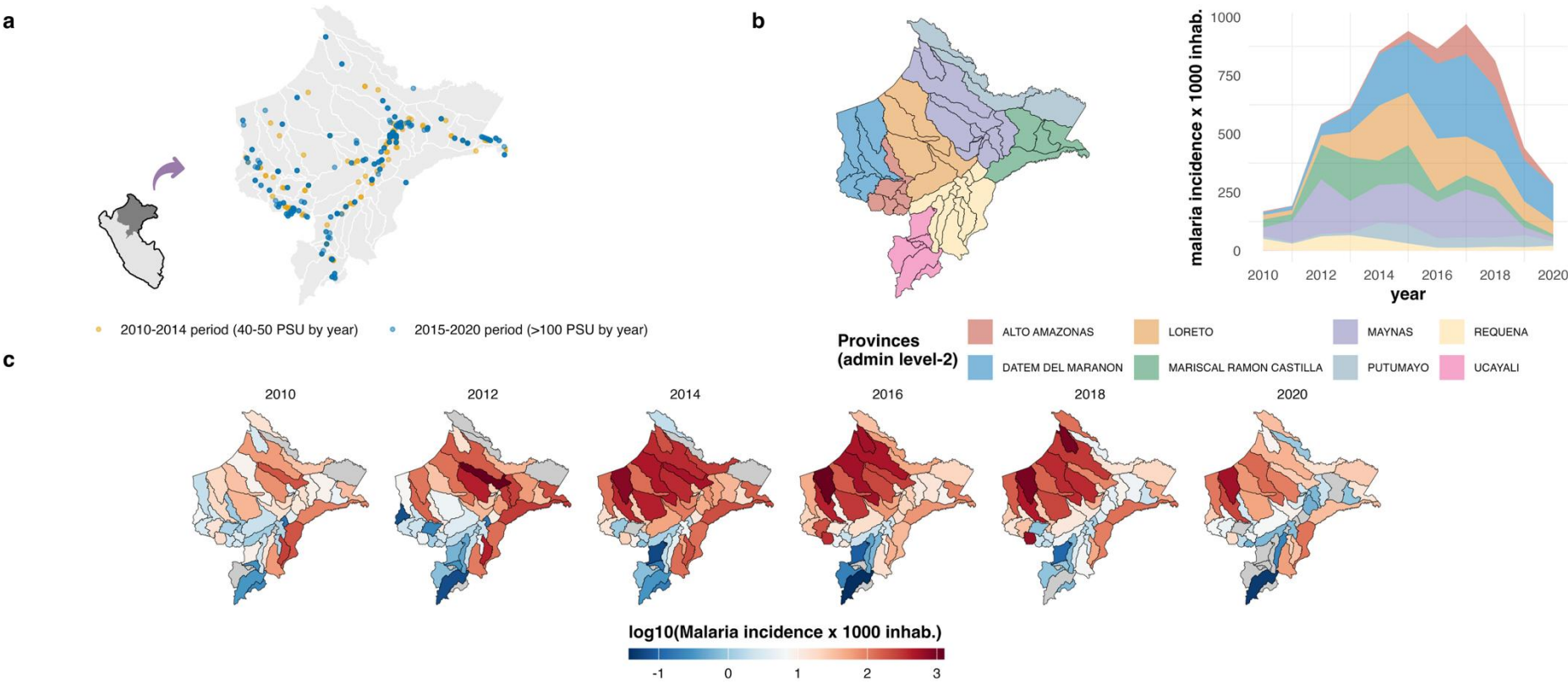
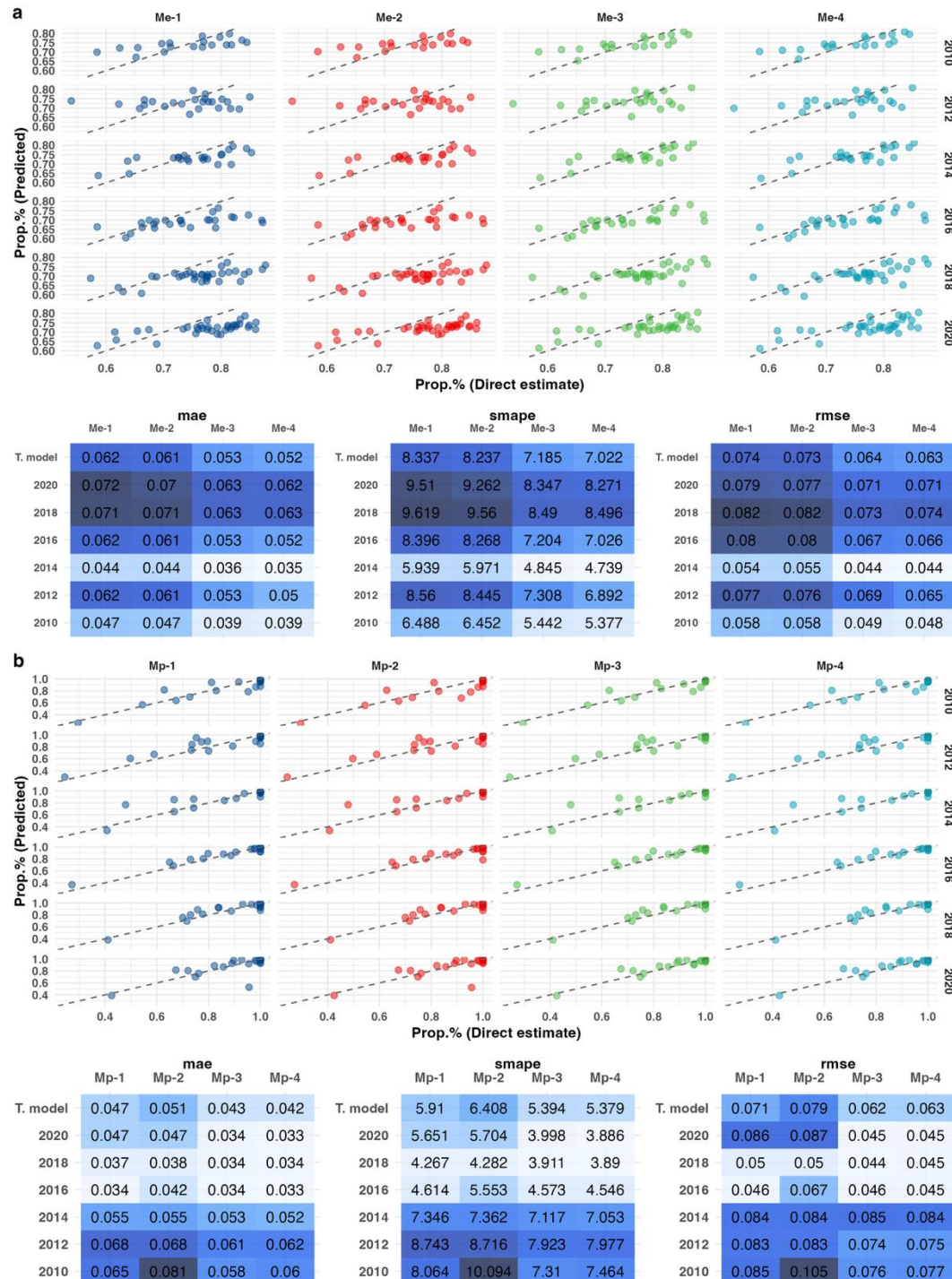


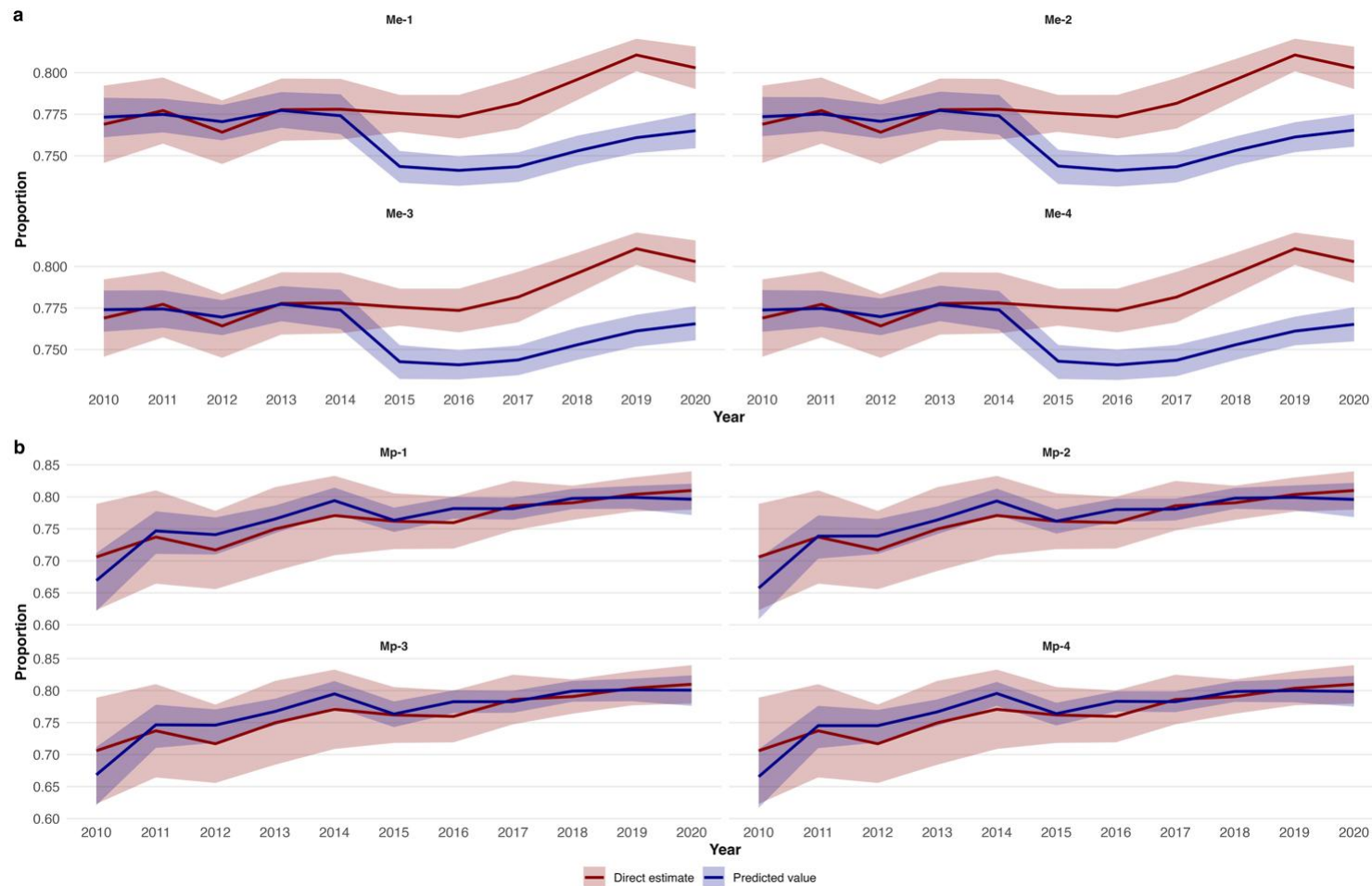
Figure 2. Predicted values and direct estimates of education (elementary education, a) and poverty (poverty status, b) and accuracy metrics in-sample by year. The dotted line has a slope of 1, showing the relationship between the direct estimate and the values predicted by each model for each year. The MAE, SMAPE, and RMSE values are shown for each model for each year of the study period included. The heatmap shows the years in which the models have a higher accuracy metric value. MAE: Mean Absolute Error; SMAPE: Symmetric Mean Absolute Percentage Error; RMSE: Root Mean Squared Error. Mp: Models for poverty indicator, Me: Models for elementary education indicator. Panels a and b, values are displayed at two-year intervals for clarity; the full annual series is provided in the supplementary figure 3.5 and 3.6.



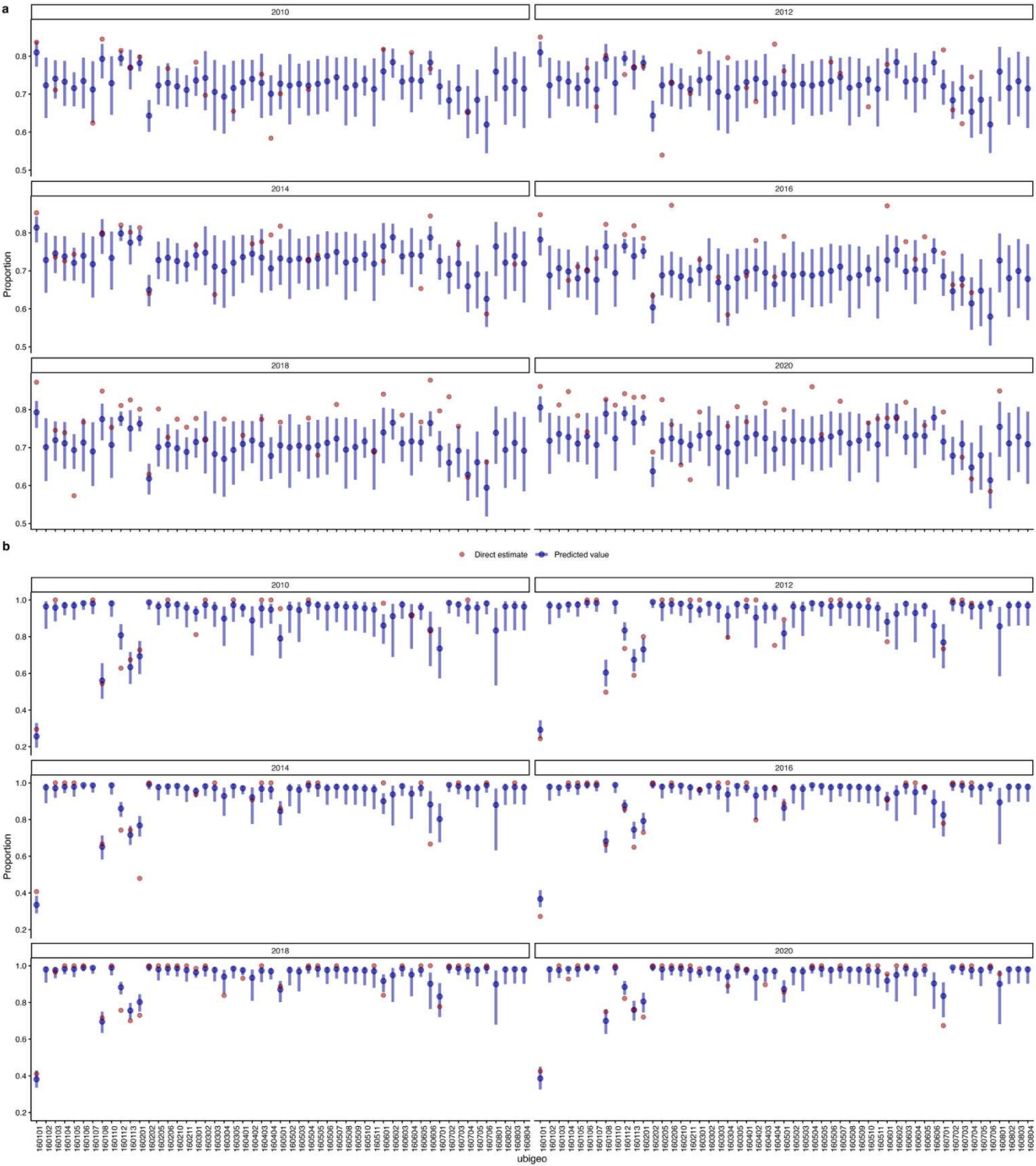
677 **Figure 3.** Predictive performance of models for poverty and education. Decomposition of the Mean
 678 Squared Error (MSE) into squared bias and variance for primary education (Elementary education,
 679 a) and poverty (poverty status, b). Values by model and area are displayed, while the larger, darker
 680 figures represent the annual average values for each model. Panel c shows the $-2\log(\text{CPO})$ values
 681 of each pre-selected model for each socioeconomic indicator. MSE: Mean Squared Error. Mp: Models
 682 for poverty indicator, Me: Models for elementary education indicator.
 683



684 **Figure 4.** Estimates of aggregate predictions at the departmental level for each model by year for primary education (Elementary education, a) and
685 poverty (poverty status, b). Department-level estimates were obtained by aggregating district-level model predictions, with associated 95% credible
686 intervals. For comparison, the corresponding estimates from the DHS survey, with their 95% confidence intervals, are also shown. The difference
687 between the two sources is summarised in terms of the Mean Absolute Error (MAE). Mp: Models for poverty indicator, Me: Models for elementary
688 education indicator.

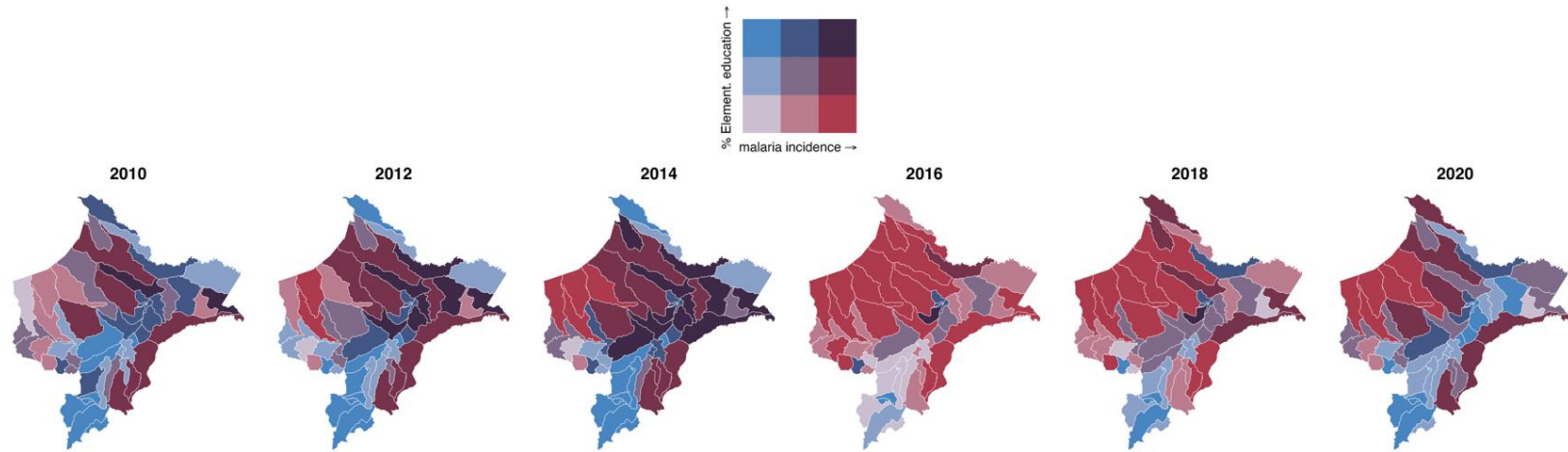


690 **Figure 5.** Predictions of socioeconomic indicators for education (Elementary education, a) and
 691 poverty (poverty status, b) in all districts of Loreto for the years 2010, 2012, 2014, 2016, 2018, and
 692 2020. Direct estimates correspond to values obtained from the DHS survey, while model-based
 693 predictions were generated using Me-3 and Mp-3, the best-performing models for education and
 694 poverty, respectively. UBIGEO: district-level geographic identifier.

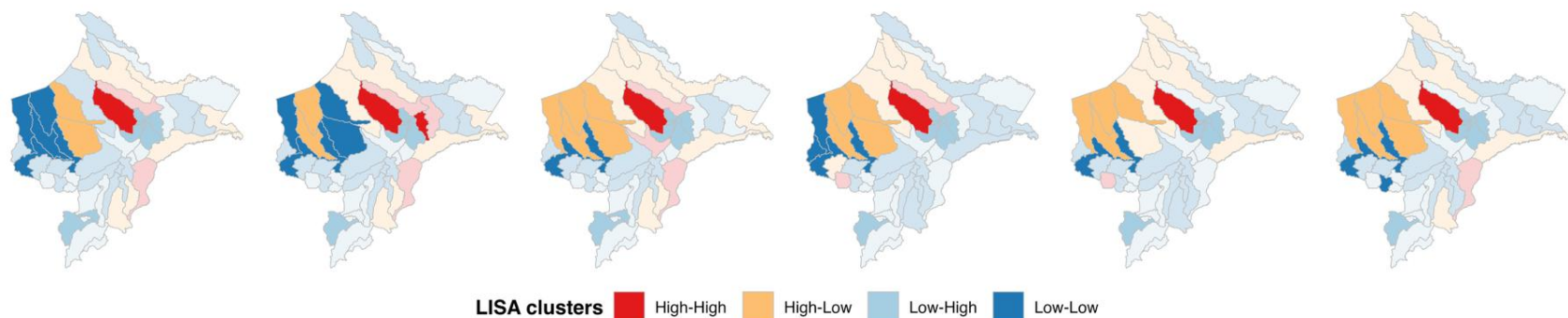


695 **Figure 6.** Spatial patterns of education and malaria incidence. a) Bivariate mapping of the relationship between the proportion of primary education
696 and the malaria index at the district level from 2010 to 2020. Both variables were categorized into tertiles, and the resulting combinations were used
697 to construct the categories displayed in the bivariate map. b) shows the clusters identified using the bivariate Local Moran's I test (malaria incidence
698 – primary education). A cluster of high malaria incidence and low primary-education levels is observed in the border area with Ecuador. Non-significant
699 clusters are shown as transparent. Panels a and b, values are displayed at two-year intervals for clarity; the full annual series is provided in the
700 supplementary figure 3.7.

a

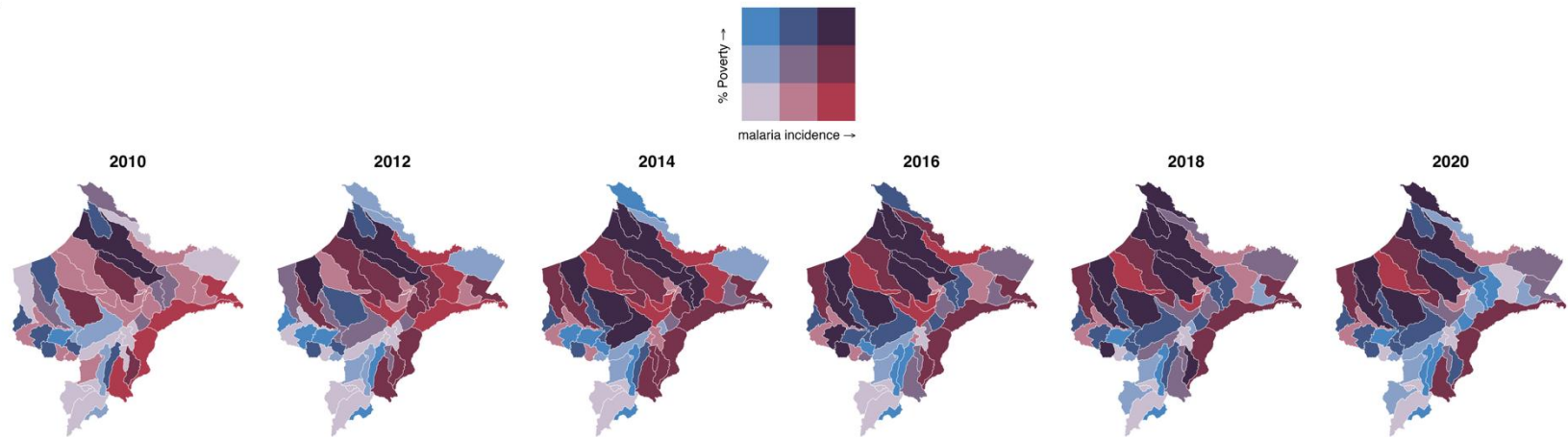


b

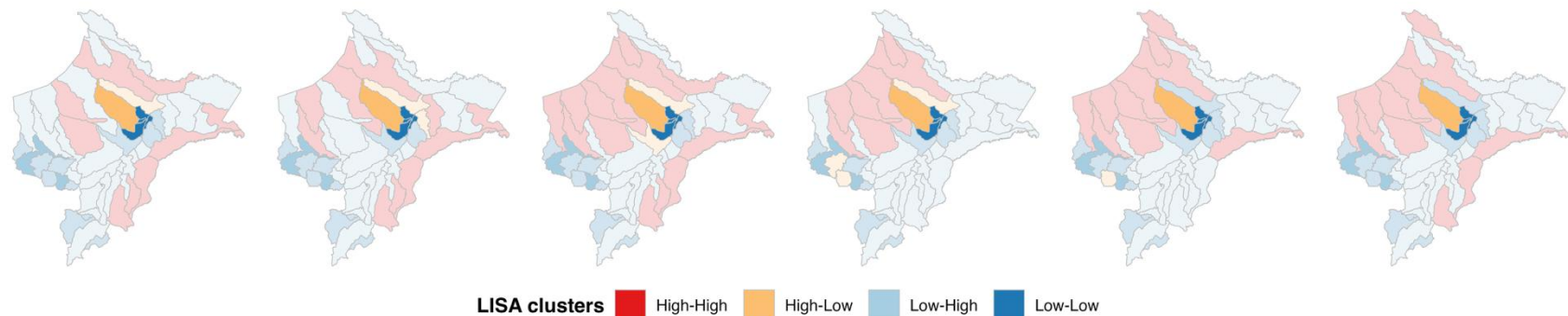


701 **Figure 7.** Spatial patterns of poverty and malaria incidence. a) Bivariate mapping of the relationship between the proportion of poverty and the
702 malaria index at the district level from 2010 to 2020. Both variables were categorized into tertiles, and the resulting combinations were used to
703 construct the categories displayed in the bivariate map. b) shows the clusters identified using the bivariate Local Moran's I test (malaria incidence –
704 poverty). Non-significant clusters are shown as transparent. Panels a and b, values are displayed at two-year intervals for clarity; the full annual
705 series is provided in the supplementary figure 3.8.
706

a



b



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)