Estimating Sub-County Literacy Rates in Kenya's Coastal Region Using the Fay–Herriot Model

Yvonne A. Onyango[1*], Alexander Kasyoki[1], Nicholas Mwilu Mutothya[1] , Oscar Ngesa[1]

Department of Mathematics, Statistics and Physical Sciences, Taita Taveta University, Kenya

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100700511

Received: 22 July 2026; Accepted: 27 July 2026; Published: 06 August 2026

ABSTRACT

Reliable literacy estimates at lower administrative levels are essential for identifying educational inequalities and informing evidence-based planning. However, national household surveys are typically designed to generate reliable estimates at national and county levels, resulting in unstable direct estimates for smaller domains such as sub-counties. This study aimed to estimate sub-county literacy rates in Kenya's Coastal Region using the Fay–Herriot area-level small area estimation model and to examine the spatial distribution and precision of the resulting estimates. Literacy data were obtained from the 2022 Kenya Demographic and Health Survey, while auxiliary variables were derived from the 2019 Kenya Population and Housing Census. Direct literacy estimates and their sampling variances were first computed for each sub-county. A Fay–Herriot model was then fitted using three selected auxiliary variables: the proportion of persons living in urban areas, the proportion in the lower wealth category, and the proportion with basic education and above. The model generated literacy estimates for 29 sub-counties, which were subsequently mapped to assess geographic variation and estimate reliability. The results revealed substantial spatial variation in literacy levels across the Coastal Region. Higher literacy estimates were observed in Mvita, Voi, Changamwe and Mwatate, while lower estimates were recorded in Tana Delta, Tana North and Lunga Lunga. The Fay–Herriot model improved estimate precision, with most relative standard errors falling below 10%, indicating acceptable reliability for small-area statistics. However, estimates for Tana North, Tana Delta and Tana River exhibited comparatively higher uncertainty and should be interpreted with caution. The study concludes that the Fay–Herriot model provides reliable sub-county literacy estimates by combining survey and census information, thereby supporting more targeted educational planning and resource allocation in Kenya's Coastal Region.

keywords: Fay–Herriot model, Literacy Estimation, Small Area Estimation, Kenya

INTRODUCTION

Literacy is a core indicator of human development and remains central to educational planning, social inclusion, economic participation and the monitoring of development goals (UNESCO 2005, 2023). In the context of this study, literacy follows the operational definition used in the Kenya Demographic and Health Survey, where a person is classified as literate if they can read all or part of a simple sentence (Kenya National Bureau of Statistics and ICF 2022). Although this definition captures basic reading ability rather than the broader concept of functional literacy, it provides a consistent and survey-based measure for statistical estimation and comparison across population domains (Kirsch, 2001; UNESCO Institute for Statistics, n.d.).

Reliable literacy statistics are particularly important for countries seeking to reduce educational inequalities and improve the targeting of public resources (UNESCO 2023; World Bank 2022). In Kenya, national and county-level statistics provide useful information for broad policy planning, but they may conceal important variation at lower administrative levels. This limitation is especially relevant because many education interventions, including teacher deployment, adult literacy programmes, learning-resource allocation and infrastructure planning, require evidence at more local levels. When literacy estimates are available only at national or county level, pockets of educational disadvantage within smaller areas may remain hidden.

Kenya's Coastal Region provides an important setting for examining this issue. The region comprises the counties of Mombasa, Kwale, Kilifi, Tana River, Lamu and Taita Taveta, and contains diverse geographic, social and economic contexts (Kenya National Bureau of Statistics 2019a; 2019b). It includes densely populated urban centres, rural settlements, remote areas, islands and sparsely populated communities. These differences may influence access to schooling, learning resources and adult education opportunities. As a result, county-level averages may not adequately represent the distribution of literacy across sub-counties. Producing reliable estimates at this level is therefore important for identifying local disparities and supporting targeted educational planning.

A major challenge, however, is that household surveys are usually designed to produce reliable estimates at national or large subnational levels rather than for all small domains (Kish 1965; Skinner, Holt, and Smith 1989). Direct survey estimates for sub-counties may therefore be affected by small sample sizes, high sampling variability and unstable variance estimates (Rao and Molina 2015; Pfeffermann 2013). In such cases, the estimates can be difficult to interpret for planning purposes, particularly when the standard errors or relative standard errors are large. This problem creates a need for statistical methods that can improve the precision of small-domain estimates without requiring new large-scale data collection.

Small area estimation provides a framework for addressing this challenge. These methods combine direct survey information with auxiliary data from sources such as censuses, administrative records or geospatial datasets in order to produce more reliable estimates for small geographic or demographic domains (Ghosh and Rao 1994; Rao and Molina 2015; Pfeffermann 2013). The Fay–Herriot model is one of the most widely used area-level small area estimation approaches (Fay and Herriot 1979). It links direct estimates to area-level auxiliary variables while accounting for sampling error and unexplained between-area variation through random effects. By borrowing strength across areas, the model can improve the stability of estimates for domains where direct survey estimates are imprecise.

This paper applies the Fay–Herriot area-level model to estimate sub-county literacy rates in Kenya's Coastal Region (Fay and Herriot 1979; Rao and Molina 2015). Direct literacy estimates and their sampling variances are obtained from the 2022 Kenya Demographic and Health Survey (Kenya National Bureau of Statistics and ICF 2022), while auxiliary variables are constructed from the 2019 Kenya Population and Housing Census (Kenya National Bureau of Statistics 2019b). The modelling focuses on sub-county-level relationships between literacy and selected area-level covariates, including the proportion of persons living in urban areas, the proportion in the lower wealth category and the proportion with basic education and above. These variables are used to support model-based estimation and to improve the reliability of literacy estimates across the 29 sub-counties in the region.

The contribution of this paper is both methodological and practical. Methodologically, it demonstrates the application of the Fay–Herriot area-level model to literacy estimation at a lower administrative level in a setting where direct survey estimates may be unstable (Fay and Herriot 1979; Corral et al. 2020). Practically, it generates sub-county literacy estimates that can support more localized interpretation of educational inequalities in Kenya's Coastal Region. The paper also uses spatial mapping to present the geographic distribution of the Fay–Herriot literacy estimates and their associated precision, an approach that can improve the communication of local inequalities for policy use (United Nations Statistics Division, n.d.; MtWanguo and Mwange 2023). This visual component is important because maps can communicate local variation more clearly than tables alone and can help identify areas where targeted intervention may be required.

The paper is organized as follows. Section 2 describes the data sources and study area. Section 3 presents the Fay–Herriot model and the estimation procedure. Section 4 presents the empirical results, including the fitted model, sub-county literacy estimates and spatial maps. Section 5 discusses the findings in relation to educational planning and small area estimation. Section 6 concludes the paper and outlines implications for future research and policy use.

Data Sources And Study Area

Study area

The study was conducted in Kenya's Coastal Region, which comprises six counties: Mombasa, Kwale, Kilifi, Tana River, Lamu and Taita Taveta (Kenya National Bureau of Statistics 2019a). The Coastal Region was considered appropriate for this study because it presents marked demographic, socio-economic and geographical diversity, providing a useful setting for analysing sub-county differences in literacy outcomes (Kenya National Bureau of Statistics, 2019a, 2019b). The region comprises highly urbanised areas such as Mombasa, alongside rural communities, remote inland settlements, sparsely populated areas and island populations. Such variation is important because differences in settlement patterns, socio-economic conditions and geographic accessibility may shape opportunities for schooling, availability of learning resources and participation in adult literacy programmes.

According to the 2019 Kenya Population and Housing Census, the total populations of Mombasa, Kwale, Kilifi, Tana River, Lamu and Taita Taveta counties were 1,208,333; 866,820; 1,453,787; 315,943; 143,920; and 340,671, respectively (Kenya National Bureau of Statistics 2019a). These figures describe the broader demographic context of the study area. However, the analytical focus of this paper is not the total population, but persons aged 15–49 years, consistent with the age coverage used in the literacy analysis.

The small area, or domain of estimation, was the sub-county. The study covered 29 sub-counties: Changamwe, Jomvu, Likoni, Kisauni, Mvita, Nyali, Kinango, Lunga Lunga, Matuga, Msambweni, Samburu, Chonyi, Ganze, Kaloleni, Kauma, Kilifi North, Kilifi South, Magarini, Malindi, Rabai, Tana Delta, Tana North, Tana River, Lamu East, Lamu West, Mwatate, Taita, Taveta and Voi. These sub-counties formed the area-level units for the Fay–Herriot model and the spatial presentation of literacy estimates.

Survey data

The main survey data source was the 2022 Kenya Demographic and Health Survey (KDHS) (Kenya National Bureau of Statistics and ICF 2022). The KDHS was used to compute direct literacy estimates and their sampling variances for each sub-county in Kenya's Coastal Region. Although the KDHS is designed primarily to provide reliable estimates at national and county levels, sub-county direct estimates were derived in this study to serve as the response variable in the Fay–Herriot area-level model.

The women's Individual Recode file and men's Recode file were imported, restricted to the six counties of Kenya's Coastal Region and harmonized into a combined analytical dataset. To maintain consistency across the two files and with the literacy target population, the analysis was restricted to respondents aged 15–49 years, consistent with the KDHS age coverage for the individual recode files (Kenya National Bureau of Statistics and ICF 2022). The literacy variable was recoded as a binary outcome. Respondents who could read a whole sentence or part of a sentence were classified as literate, while those who could not read were classified as illiterate. Missing and non-standard responses were excluded from the estimation process.

Survey-weighted direct estimates were computed to account for the complex KDHS design, including primary sampling units, stratification and sampling weights (Kalton 1983; Skinner, Holt, and Smith 1989; Kenya National Bureau of Statistics and ICF 2022). For each sub-county, the direct literacy estimate, standard error and sampling variance were obtained. These sampling variances were then incorporated into the Fay–Herriot model as known area-level variances.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII July 2026

Auxiliary and spatial data

Auxiliary variables were constructed from the 2019 Kenya Population and Housing Census (Kenya National Bureau of Statistics 2019b). The census data were restricted to the six coastal counties and to persons aged 15–49 years in order to align the auxiliary information with the KDHS literacy target population. The initial set of candidate auxiliary variables included measures of educational attainment, sex composition, age structure, wealth status, internet access, urban residence, household size, population density, employment status and employment type. These variables were summarized mainly as sub-county-level proportions so that they could be linked to the direct literacy estimates.

The final Fay–Herriot model used three selected census-based auxiliary variables: the proportion of persons living in urban areas, the proportion of persons in the lower wealth category and the proportion of persons with basic education and above. These variables were selected because they were theoretically relevant to literacy outcomes and provided a suitable balance between explanatory power, model parsimony and acceptable multicollinearity.

Spatial data were used to support the mapping of literacy estimates and their precision across sub-counties. The mapping process required harmonizing sub-county names between the statistical estimation dataset and the administrative boundary data, aggregating duplicated spatial units where necessary and merging the final estimates with the spatial boundary file. Population-density information from WorldPop was explored as a spatially derived auxiliary source (WorldPop 2019), but it was not retained in the final model. Therefore, the fitted model should be interpreted as a standard area-level Fay–Herriot model rather than a spatial Fay–Herriot model.

Construction of the area-level modelling dataset

The final modelling dataset was created by merging the KDHS-based direct literacy estimates with the selected KPHC-based auxiliary variables at the sub-county level. For each of the 29 sub-counties, the dataset contained the direct literacy proportion, the corresponding sampling variance and the selected auxiliary variables. This structure is appropriate for the Fay–Herriot area-level model because the response variable is an area-level direct estimate and the sampling variance is treated as known (Fay and Herriot 1979; Rao and Molina 2015).

Estimates are reported as proportions in the modelling tables and as percentages in figures and maps for ease of interpretation. The resulting area-level dataset provided the basis for fitting the Fay–Herriot model and for producing spatial maps of sub-county literacy estimates and relative standard errors.

Fay–Herriot Model And Estimation Procedure

Direct estimates and sampling variances

Let $\hat{Y}_i$ denote the direct survey estimate of the literacy rate for sub-county i, where i = 1, 2, …, m, and m = 29 is the number of sub-counties in Kenya's Coastal Region. The direct estimates were obtained from the 2022 Kenya Demographic and Health Survey using survey-weighted estimation to account for sampling weights, stratification and primary sampling units (Kenya National Bureau of Statistics and ICF 2022; Skinner, Holt, and Smith 1989). For each sub-county, the corresponding sampling variance was denoted by D_i..

The direct estimates served two purposes in the analysis. First, they provided the baseline design-based literacy estimates for the sub-counties. Second, they were used as the response variable in the Fay–Herriot area-level model. The sampling variances D_i.were treated as known area-level variances in the model. This is consistent with the standard Fay–Herriot framework, where direct survey estimates and their sampling variances are combined with auxiliary information to improve small-area estimation (Fay and Herriot 1979; Rao and Molina 2015).

Because one sub-county, Kauma, had a zero estimated sampling variance from the direct survey estimation, a small positive lower bound was applied before fitting the Fay–Herriot model. Specifically, the zero sampling variance was replaced by the smallest positive sampling variance observed among the remaining sub-counties. This adjustment was made to avoid treating the Kauma direct estimate as perfectly precise, which would be unrealistic in small-domain survey estimation where direct variance estimates may be unstable (Rao and Molina 2015; Prasad and Rao 1990). The adjustment retained Kauma in the analysis while ensuring that all area-level sampling variances used in the Fay–Herriot model were strictly positive.

Selection of auxiliary variables

Auxiliary variables were constructed from the 2019 Kenya Population and Housing Census and aggregated at sub-county level (Kenya National Bureau of Statistics 2019b). The candidate variables included demographic, socio-economic, educational and settlement-related characteristics expected to be associated with literacy. The selection of auxiliary variables was guided by both statistical performance and theoretical relevance.

Correlation analysis was used to examine the association between the direct literacy estimates and each candidate auxiliary variable. Multicollinearity was assessed using the variance inflation factor. Candidate models were then compared using model fit and parsimony criteria, including the Akaike information criterion (AIC), Bayesian information criterion (BIC) and adjusted coefficient of determination. Variables that introduced serious multicollinearity were avoided because they could produce unstable coefficient estimates and weaken model interpretation.

The final Fay–Herriot model included three auxiliary variables: the proportion of persons living in urban areas, the proportion of persons in the lower wealth category and the proportion of persons with basic education and above. These variables were retained because they provided an appropriate balance between explanatory value, parsimony, acceptable multicollinearity and substantive relevance to literacy outcomes.

Fay–Herriot area-level model

The Fay–Herriot model was used because the outcome of interest was available as area-level direct literacy estimates with corresponding sampling variances. The model is appropriate in this setting because it does not require full population-level microdata for all domains, but instead combines direct estimates with area-level auxiliary variables (Fay and Herriot 1979; Chambers and Clark 2012; Rao and Molina 2015). The model was specified in two stages. The first stage is the sampling model expressed using Eqn. 1

$$\hat{Y}_i = \theta_i + e_i, \quad e_i \sim N(0, D_i) \tag{1}$$

where $\hat{Y}_i$ is the direct literacy estimate for sub-county (i), θ_i is the true but unknown literacy rate, e_i is the sampling error and D_i is the sampling variance of the direct estimate.

The second stage is the linking model expressed using Eqn. 2.

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \qquad u_i \sim N(0, \delta_u^2) \tag{2}$$

where $\mathbf{x}_i$ is the vector of auxiliary variables for sub-county (i), $\boldsymbol{\beta}$ is the vector of regression coefficients, u_i is the area-specific random effect and σ_u^2 is the between-area variance component. The sampling errors e_i and area random effects u_i were assumed to be independent. Combining the sampling and linking models gives the area-level Fay–Herriot model as shown in Eqn. 3.

$$\hat{Y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i \tag{3}$$

For this study, the fitted model was specified as:

$$\text{literacy\_prop}_i = \beta_0 + \beta_1(\text{prop\_urban}_i) + \beta_2(\text{prop\_lower\_wealth}_i) + \beta_3(\text{prop\_basic\_edu\_above}_i) + u_i + e_i$$

Here, literacy_prop_i is the direct literacy proportion for sub-county (i), prop_urban_i is the proportion of persons living in urban areas, $\text{prop\_lower\_wealth}_i$ is the proportion of persons in the lower wealth category and $\text{prop\_basic\_edu\_above}_i$ is the proportion of persons with basic education and above.

Estimation of Fay–Herriot model parameters

The unknown variance component σ_u^2 was estimated using restricted maximum likelihood. Restricted maximum likelihood was preferred because it accounts for the estimation of fixed effects when estimating variance components and is commonly used in mixed-model small area estimation (Rao and Molina 2015). The regression coefficients were estimated jointly with the variance component under the Fay–Herriot model.

The model was fitted using the fh() function from the emdi package in R, which supports the estimation, diagnostic assessment and mapping of regionally disaggregated small-area indicators (Kreutzmann et al., 2019). The direct literacy estimate was specified as the response variable, the selected auxiliary variables were included as fixed-effect predictors, and the sub-county identifier defined the model domains. The corresponding sampling variances were estimated from the KDHS survey data and treated as fixed, known inputs in the Fay–Herriot model.

Empirical best linear unbiased prediction

After estimating the model parameters, empirical best linear unbiased predictors were obtained for each sub-county. The Fay–Herriot estimator combines the direct estimate with the regression-synthetic component, with the balance between the two determined by the relative size of the between-area variance and the sampling variance of the direct estimate (Fay and Herriot 1979; Ghosh and Rao 1994; Rao and Molina 2015).

The empirical best linear unbiased predictor for sub-county (i) was computed using Eqn. 4.

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{Y}_i + (1 - \hat{\gamma}_i)\mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \qquad (4)$$

Where $\hat{\gamma}_i$ and $\hat{Y}_i$ are the shrinkage factor and direct literacy estimate for sub-county (i) respectively. The shrinkage factor $\hat{\gamma}_i$, is further derived using Eqn 5

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + D_i} \qquad (5)$$

From Eqn 5 above, when the direct estimate for a sub-county has a small sampling variance, $\hat{\gamma}_i$ becomes larger and the Fay–Herriot estimate remains closer to the direct estimate. When the direct estimate has a large sampling variance, $\hat{\gamma}_i$ becomes smaller and the model-base regression component receives greater weight. This borrowing of strength across areas is the key mechanism through which the Fay–Herriot model improves the stability of small-area estimates (Fay and Herriot 1979; Rao and Molina 2015).

Measures of uncertainty, precision and model diagnostics

The uncertainty of the Fay–Herriot estimates was assessed using mean squared error (MSE), standard error (SE) and relative standard error (RSE). The standard error (shown in Eqn. 6) for each sub-county estimate was obtained as the square root of the estimated mean squared error (Prasad and Rao 1990; Rao and Molina 2015):

$$SE\left(\hat{\theta}_i^{EBLUP}\right) = \sqrt{MSE\left(\hat{\theta}_i^{EBLUP}\right)} \qquad (6)$$

Where the relative standard error was then computed as:

$$RSE_i = \frac{SE\left(\hat{\theta}_i^{EBLUP}\right)}{\hat{\theta}_i^{EBLUP}} \times 100 \tag{7}$$

Where the RSE was used to assess precision because it expresses the standard error relative to the magnitude of the estimate and determined using Eqn 7. Lower relative standard errors indicate more precise estimates. In the presentation of results, estimates were reported as proportions in tables and as percentages in figures and maps for ease of interpretation (Rao and Molina 2015).

Model diagnostics were conducted to assess the adequacy of the fitted Fay–Herriot model. The assessment focused on the distribution of residuals and the relationship between residuals and fitted values. A normal Q-Q plot was used to examine whether the residuals followed an approximately normal distribution, while a residuals-versus-fitted-values plot was used to assess whether the residuals showed systematic patterns. The Shapiro–Wilk test was also used as a formal check of residual normality. These diagnostics were used to determine whether the model assumptions were reasonably satisfied and whether any sub-counties were poorly fitted by the selected auxiliary variables. Since the Fay–Herriot model is an area-level model, the resulting associations were interpreted at sub-county level and not as individual-level causal relationships.

Spatial presentation of estimates

The final Fay–Herriot estimates and their relative standard errors were merged with sub-county boundary data to produce spatial maps. Sub-county names in the statistical estimation dataset and the administrative boundary data were harmonized before mapping. The maps were used to show the geographic distribution of estimated literacy rates and the spatial pattern of estimate precision across Kenya's Coastal Region. The spatial outputs were used for interpretation and communication of results; the fitted model itself was a standard area-level Fay–Herriot model rather than a spatial Fay–Herriot model (MtWanguo and Mwange 2023).

Ethical considerations

This study involved secondary analysis of anonymized data from the 2022 Kenya Demographic and Health Survey, the 2019 Kenya Population and Housing Census and publicly available geospatial sources. No participants were recruited or contacted directly during the study. Access to the KDHS data was obtained through authorization from the DHS Program, and all data were used in accordance with the applicable conditions of use. The original KDHS data collection received the relevant ethical approvals, and informed consent was obtained from participants during the survey. Confidentiality was maintained throughout the analysis, and the results are reported only at aggregated sub-county level.

EMPIRICAL RESULTS

Selection of auxiliary variables

The final Fay–Herriot model was fitted using three auxiliary variables derived from the 2019 Kenya Population and Housing Census: the proportion of persons in the lower wealth category, the proportion of persons living in urban areas and the proportion of persons with basic education and above (Kenya National Bureau of Statistics 2019b) as shown in Table 1. These variables were selected after comparing alternative candidate models using statistical diagnostics and substantive relevance to literacy outcomes.

Although the candidate model that included basic education, lower wealth, urban residence and professional employment produced the lowest AIC and BIC values, it was not selected because it showed serious multicollinearity, with a maximum variance inflation factor of 22.3976. The selected model, consisting of basic education, lower wealth and urban residence, provided a better balance between model fit, parsimony and interpretability. It had a maximum variance inflation factor of 4.5558 and an adjusted (R^2) of 0.5227. The selected variables were also theoretically meaningful because literacy is expected to be associated with education, socio-economic status and urban–rural differences.

Table 1. Selected auxiliary variables

Variable	Description	VIF	Correlation
prop_lower_wealth	Proportion in lower wealth category	4.5558	-0.5910
prop_urban	Proportion living in urban places	4.1693	0.3513
prop_basic_edu_above	Proportion with basic education and above.	1.2155	0.5871

The correlations show that the proportion with basic education and above was positively associated with direct literacy estimates, while the proportion in the lower wealth category was negatively associated with literacy. Urban residence had a weaker positive bivariate association with literacy, but it was retained because of its statistical relevance and its importance in explaining geographical differences in education access and literacy outcomes.

Fitted Fay–Herriot model and the diagnostics

The Fay–Herriot model was fitted using the direct literacy estimates as the response variable and the selected auxiliary variables as predictors as shown in Table 2. The sampling variances of the direct estimates were incorporated as known area-level variances, and the model was estimated using restricted maximum likelihood (Fay and Herriot 1979; Rao and Molina 2015).

Table 2. Fay–Herriot model coefficients

Parameter	Coefficient	Standard error	Test statistic	p-value
Intercept	-1.3192	1.0074	-1.3095	0.1904
prop_lower_wealth	-0.3719	0.1086	-3.4262	0.0006***
prop_urban	-0.1606	0.0782	-2.0545	0.0399*
prop_basic_edu_above	2.4555	1.0179	2.4122	0.0159*

The coefficient for the proportion in the lower wealth category was negative and statistically significant. This indicates that sub-counties with higher proportions of persons in the lower wealth category tended to have lower literacy estimates, after controlling for the other auxiliary variables. The coefficient for the proportion with basic education and above was positive and statistically significant, suggesting that literacy rates were higher in sub-counties with a larger share of persons who had attained at least basic education.

Urban residence had a negative and statistically significant coefficient in the multivariable model despite its positive bivariate correlation with literacy, suggesting that its effect was conditional on wealth and educational attainment rather than reflecting a direct negative relationship. The estimated area-level variance component of 0.0046 indicated residual variation between sub-counties after accounting for the selected covariates. With a log-likelihood of 33.7164, AIC of −57.4328, BIC of −50.5963, adjusted (R^2) of 0.5227 and Fay–Herriot (R^2) of 0.5695, the model showed moderate explanatory power, which was reasonable for the 29 sub-county domains.

Diagnostic checks were conducted to assess the adequacy of the fitted Fay–Herriot model. The normal Q–Q plot of the standardized residuals showed that most observations followed the theoretical normal reference line reasonably well, although departures were evident in the lower tail of the distribution (Fig. 1a). This finding was consistent with the Shapiro–Wilk test, which indicated a statistically significant departure from normality (W = 0.6375, p < 0.001). The deviations were, however, largely confined to a few sub-counties rather than being widespread across all domains.

The standardized residuals-versus-fitted-values plot showed that the residuals were generally centred around zero without a clear systematic pattern across the fitted values (Fig. 1b). Most standardized residuals fell within the reference limits of ±2, although Tana North, Tana River and Tana Delta had relatively large negative residuals. This indicates that literacy variation in these sub-counties was not fully explained by the selected auxiliary variables. Overall, the model provided an acceptable fit for most sub-counties, although the estimates for the identified domains should be interpreted with consideration of their associated standard errors, mean squared errors and relative standard errors.

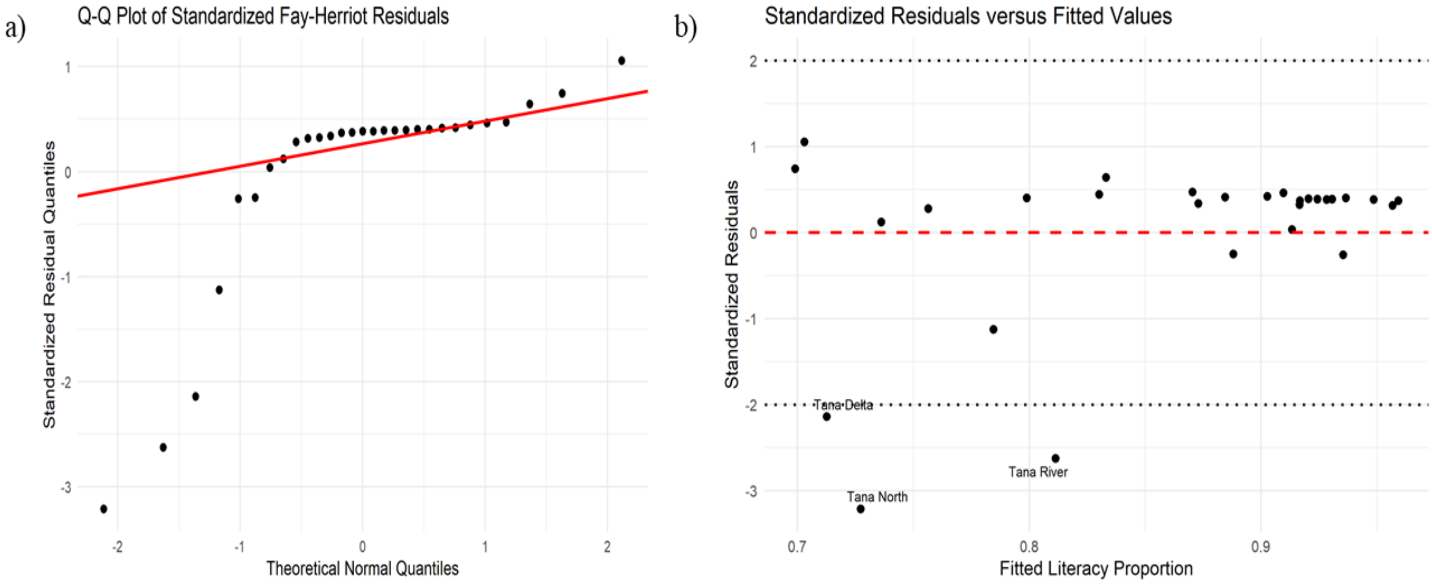

Figure 1. (a) Q–Q plot of standardized Fay–Herriot residuals; (b) standardized residuals versus fitted values.

Direct and Fay–Herriot literacy estimates

The Fay–Herriot estimates were obtained as empirical best linear unbiased predictors. They combine the direct survey estimates, the regression-synthetic component and the area-level random effects (Fay and Herriot 1979; Rao and Molina 2015). Table 3 presents the Fay–Herriot literacy estimates, mean squared errors, standard errors and relative standard errors for the 29 sub-counties.

Table 3. Direct and Fay–Herriot literacy estimates by sub-county

Sub-county	**Direct estimate**	**FH estimate**	**FH MSE**	**FH SE**	**FH RSE (%)**	**Absolute difference**
Mvita	0.9936	0.9931	0.000035	0.0059	0.5918	0.0005
Voi	0.9700	0.9692	0.000115	0.0107	1.1083	0.0008
Changamwe	0.9659	0.9656	0.000072	0.0085	0.8763	0.0003
Mwatate	0.9571	0.9572	0.000244	0.0156	1.6319	0.0001
Taveta	0.9389	0.9360	0.000467	0.0216	2.3086	0.0029
Nyali	0.9358	0.9354	0.000186	0.0136	1.4589	0.0005
Msambweni	0.9345	0.9342	0.000233	0.0153	1.6331	0.0003
Taita	0.9321	0.9341	0.000386	0.0196	2.1030	0.0020
Likoni	0.9304	0.9292	0.000122	0.0110	1.1868	0.0012
Rabai	0.9280	0.9274	0.000367	0.0192	2.0666	0.0006
Kauma	0.9119	0.9111	0.000035	0.0059	0.6453	0.0008
Kilifi South	0.9092	0.9094	0.000160	0.0126	1.3893	0.0003
Lamu East	0.9099	0.9084	0.001080	0.0329	3.6174	0.0014
Kisauni	0.8975	0.8993	0.000439	0.0210	2.3303	0.0018
Kaloleni	0.8883	0.8860	0.000186	0.0136	1.5397	0.0023
Matuga	0.8791	0.8759	0.001811	0.0426	4.8588	0.0032
Malindi	0.8590	0.8704	0.001010	0.0318	3.6511	0.0114

Kilifi North	0.8448	0.8662	0.001146	0.0338	3.9080	0.0214
Lamu West	0.8638	0.8651	0.000651	0.0255	2.9496	0.0012
Magarini	0.8780	0.8551	0.000637	0.0252	2.9522	0.0229
Samburu	0.8543	0.8453	0.002119	0.0460	5.4450	0.0090
Jomvu	0.8170	0.8381	0.001489	0.0386	4.6042	0.0210
Ganze	0.7709	0.7585	0.000844	0.0291	3.8300	0.0124
Tana River	0.6515	0.7527	0.003235	0.0569	7.5559	0.1012
Kinango	0.7388	0.7420	0.000909	0.0302	4.0641	0.0032
Chonyi	0.6996	0.7082	0.001159	0.0340	4.8075	0.0086
Lunga Lunga	0.6373	0.6880	0.001714	0.0414	6.0171	0.0506
Tana North	0.5255	0.6464	0.003474	0.0589	9.1184	0.1209
Tana Delta	0.5148	0.5996	0.002237	0.0473	7.8880	0.0848

The Fay–Herriot estimates show clear variation in literacy across sub-counties in Kenya's Coastal Region. The highest model-based literacy estimates were observed in Mvita, Voi, Changamwe and Mwatate. These sub-counties had estimated literacy rates above 95%. In contrast, the lowest estimates were observed in Tana Delta, Tana North and Lunga Lunga, indicating lower literacy levels in these areas relative to the rest of the region.

The relative standard errors indicate that the Fay–Herriot estimates were generally precise. Most sub-counties had RSE values below 10%, suggesting acceptable precision for small-area interpretation. The relatively higher RSE values for Tana North, Tana Delta and Tana River indicate that the estimates for these areas should be interpreted with greater caution.After applying the lower-bound adjustment to the zero sampling variance for Kauma, the Fay–Herriot model was refitted. The adjustment removed the unrealistic zero standard error and zero relative standard error for Kauma while preserving the overall spatial pattern of literacy estimates across the Coastal Region. The estimates for Kauma should therefore be interpreted as model-supported estimates with adjusted uncertainty rather than as estimates with perfect precision.

The Fay–Herriot estimates remained close to the direct estimates in sub-counties where the latter were already precise, whereas larger adjustments occurred in areas with greater sampling variability. The largest upward adjustments were observed in Tana North, Tana River, Tana Delta and Lunga Lunga. For example, the estimate for Tana North increased from 0.5255 under direct estimation to 0.6464 under the Fay–Herriot model, while that of Tana River increased from 0.6515 to 0.7527. Similarly, the estimates for Tana Delta and Lunga Lunga increased from 0.5148 to 0.5996 and from 0.6373 to 0.6880, respectively. This pattern reflects the shrinkage property of the Fay–Herriot estimator, whereby precise direct estimates remain close to their survey-based values, while less precise estimates receive greater influence from the regression model and auxiliary information (Fay and Herriot 1979; Rao and Molina 2015). The mean relative standard error declined from 4.12% under direct estimation to 3.32% under the Fay–Herriot model, representing an overall reduction of 19.46%. Relative efficiency was calculated as the ratio of the original direct sampling variance to the mean squared error of the corresponding Fay–Herriot estimate. Kauma was excluded from this calculation because its original sampling variance was effectively zero and therefore did not provide a meaningful basis for comparison. Across the remaining 28 sub-counties, all relative-efficiency values exceeded one, with a mean of 1.23 and a median of 1.08, indicating that the direct sampling variances were, on average, 1.23 times the corresponding Fay–Herriot mean squared errors. The largest efficiency gains were recorded in Tana River and Tana North, with relative-efficiency values of 2.45 and 1.98, respectively. These results provide quantitative evidence that the Fay–Herriot model produced more precise sub-county literacy estimates than direct estimation.

Precision and spatial distribution of the Fay–Herriot estimates

The Fay–Herriot estimator produced a mean estimate of 0.8589, a mean standard error of 0.0265, a mean MSE of 0.0009 and a mean RSE of 3.3151%. Its maximum RSE was 9.1184%, which was below the maximum RSE observed for the direct estimates; 15.7852. These results indicate that the Fay–Herriot model improved estimate stability while retaining a link to the direct survey information. Although a detailed comparison with direct, synthetic and composite estimators is not the central focus of this manuscript, the precision results show that the Fay–Herriot estimates provide a reliable basis for presenting sub-county literacy patterns. The mean RSE of 3.3151% indicates that the model-based estimates are sufficiently precise for spatial interpretation and for identifying broad geographic patterns in literacy across the region.

Spatial maps were produced to show the distribution of Fay–Herriot literacy estimates across the 29 sub-counties in Kenya's Coastal Region. Sub-county names in the estimation and administrative boundary datasets were harmonized, duplicated spatial units were aggregated where necessary, and the final estimates were merged with the boundary file to ensure consistency between statistical and geographic units (MtWanguo and Mwange 2023). The map revealed clear spatial variation in literacy. Higher estimates were concentrated in more urbanized or better-connected areas, including Mvita, Voi, Changamwe and Mwatate, while lower levels were observed in Tana Delta, Tana North, Lunga Lunga, Chonyi, Kinango and Tana River. These patterns indicate that county-level averages may conceal important sub-county disparities.

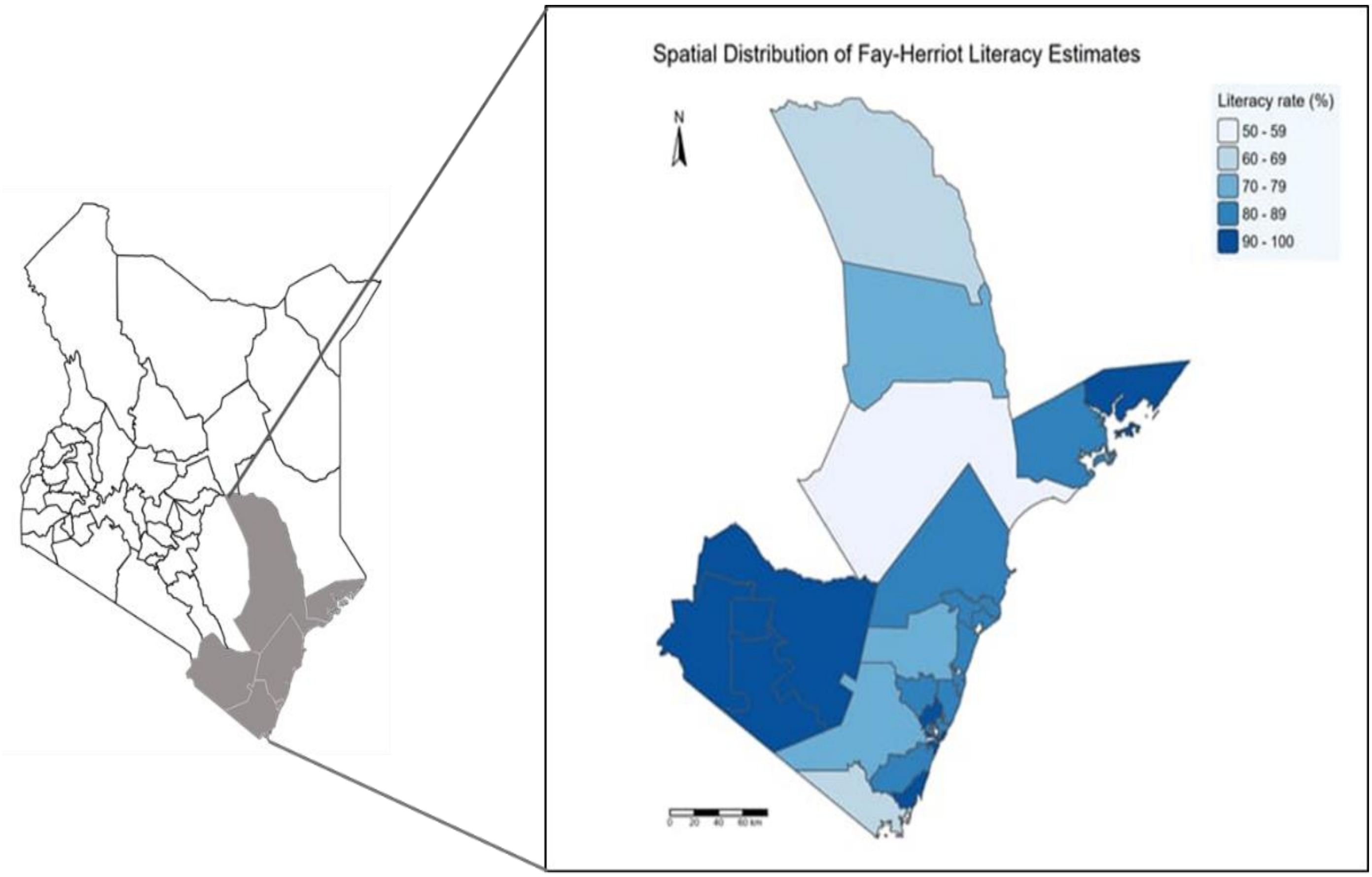

Figure 2. Spatial distribution of Fay-Herriot literacy estimates across sub-counties in Kenya's Coastal region

The spatial results are useful because they translate the model-based estimates into a form that can support local interpretation. Areas with lower estimated literacy rates may require closer policy attention, especially where low literacy coincides with rurality, poverty, remoteness or limited educational infrastructure (UNESCO 2023; World Bank 2022). However, the map should be interpreted together with the corresponding RSE map so that both the estimated level of literacy and the reliability of the estimate are considered.

A second map was produced to show the spatial distribution of Fay–Herriot relative standard errors. This map complements the literacy estimate map by identifying where the model-based estimates are more or less precise. Sub-counties with lower RSE values indicate more reliable estimates, while those with higher RSE values indicate greater uncertainty.

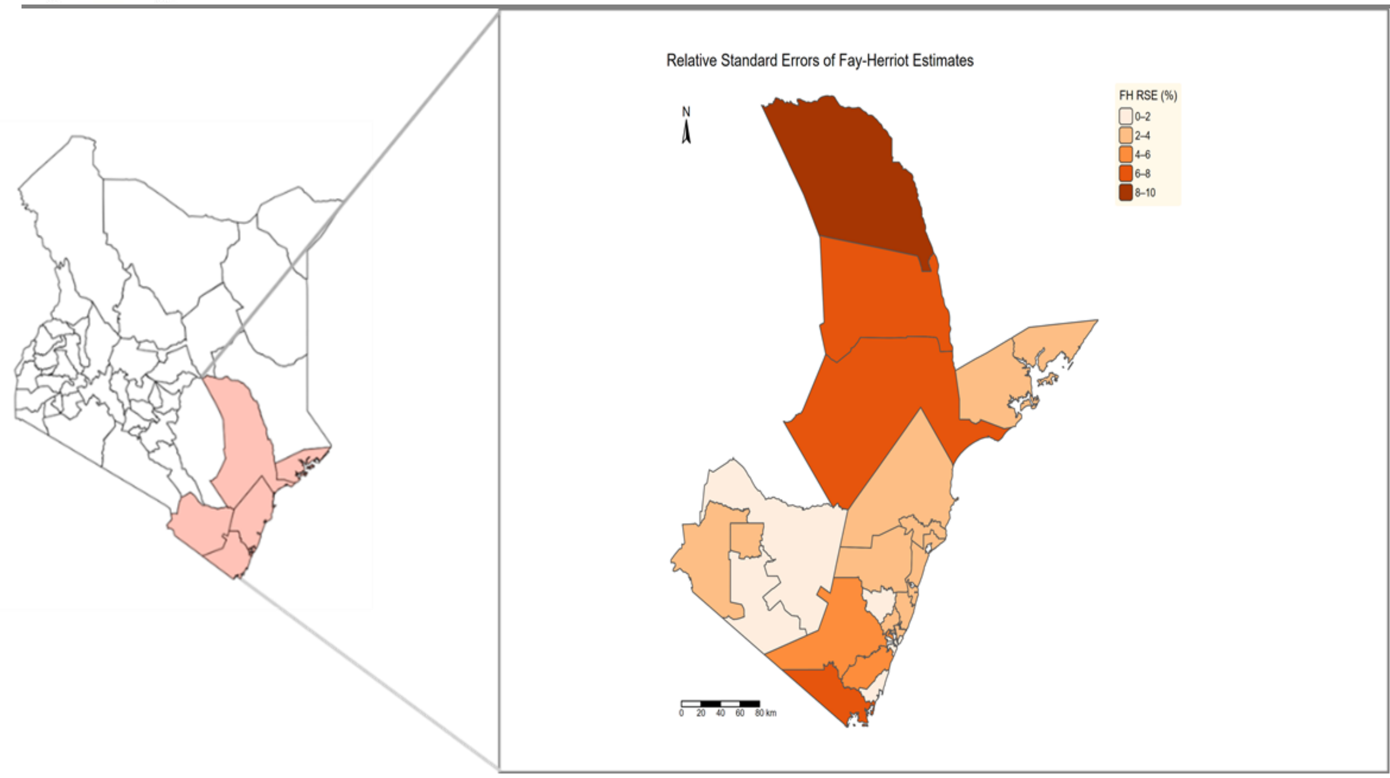

Figure 3. Spatial distribution of Fay–Herriot relative standard errors across sub-counties in Kenya's Coastal Region

The RSE map indicated that most sub-counties had relatively low estimation uncertainty, demonstrating the improved precision achieved through the Fay–Herriot model. However, Tana North, Tana Delta and Tana River exhibited comparatively higher RSE values than most other sub-counties and should therefore be interpreted with greater caution, although their estimates remain useful for broad spatial analysis. Interpreting the literacy and RSE maps together is essential because the reliability of an estimate influences the strength of the evidence it provides; for example, a low literacy estimate with a high RSE carries less certainty than a similarly low estimate with a low RSE. Overall, the spatial analysis confirmed that the Fay–Herriot model supports both numerical and geographic interpretation of literacy inequalities by providing policy-relevant information on the distribution of literacy levels across sub-counties while simultaneously indicating the precision of the corresponding estimates.

DISCUSSION

The findings demonstrate substantial variation in literacy rates across sub-counties in Kenya's Coastal Region. Higher Fay–Herriot estimates were observed in Mvita, Voi, Changamwe and Mwatate, while lower estimates were recorded in Tana Delta, Tana North and Lunga Lunga. These differences show that county-level and regional averages may conceal important local disparities, particularly in geographically and socio-economically diverse areas (Kenya National Bureau of Statistics 2019a). The sub-county estimates therefore provide more localized evidence for planning teacher deployment, adult literacy programmes, school infrastructure, learning materials and community-based education initiatives. Without such disaggregated information, planners may overlook areas where literacy outcomes are comparatively low and where additional policy attention may be required (World Bank 2022; UNESCO 2023).

Spatial mapping strengthened the interpretation and communication of the model-based estimates (MtWanguo and Mwange 2023). The literacy map showed the geographic distribution of estimated literacy levels, while the relative standard error map indicated the precision of the estimates. Reading the two maps together is important because a low literacy estimate with a low relative standard error provides stronger evidence for intervention than a similar estimate accompanied by greater uncertainty. Most Fay–Herriot estimates had relative standard errors below 10%, indicating that they were sufficiently precise for broad sub-county comparisons. Nevertheless, estimates with comparatively higher uncertainty should be interpreted cautiously, particularly when informing major planning or resource-allocation decisions. Small area estimates are therefore most useful when both the point estimates and their associated uncertainty measures are considered (United Nations Statistics Division, n.d.; Pfeffermann 2013).

The results also confirm the value of the Fay–Herriot model for education statistics. Although direct estimates are design-based and derived from observed survey data, they may be unstable in domains with small samples or high sampling variability (Kish 1965; Rao and Molina 2015). Tana North, Tana River, Tana Delta and Lunga Lunga, for example, recorded relatively high direct relative standard errors. The Fay–Herriot model addressed this limitation by retaining the direct survey estimates and their sampling variances while borrowing strength from census-based auxiliary information (Fay and Herriot 1979; Corral et al. 2020). The positive association between basic education attainment and literacy and the negative association between lower-wealth status and literacy were substantively plausible, reflecting the importance of schooling and socio-economic resources in developing and retaining reading skills (Kirsch 2001; UNESCO 2005, 2023; World Bank 2022). Urban residence, however, had a positive bivariate association with literacy but a negative coefficient after controlling for wealth and educational attainment. This change should not be interpreted as evidence that urban residence directly reduces literacy. Rather, it may reflect shared explanatory information, conditional relationships among the covariates or specific local characteristics of the Coastal Region. The coefficients should therefore be interpreted as predictive area-level relationships rather than causal effects.

The initial zero relative standard error for Kauma required specific methodological treatment. Because a zero RSE would imply unrealistic perfect precision, a lower-bound adjustment was applied to its sampling variance before fitting the final model (Prasad and Rao 1990; Rao and Molina 2015). This adjustment retained Kauma in the analysis while preventing its direct estimate from being treated as uncertainty-free. After refitting the model, Kauma had a positive uncertainty measure, while the general spatial pattern of literacy estimates remained unchanged. Its result should therefore be interpreted as a model-supported estimate with adjusted uncertainty rather than as an estimate with perfect precision.

The diagnostic findings indicate that the fitted model was useful but did not capture all variation in literacy. Most residuals were concentrated around zero, suggesting a reasonable fit for many sub-counties, although the normality assumption was not fully satisfied and a few areas had relatively large negative residuals. These departures indicate that some locally important influences may not have been adequately represented by the selected auxiliary variables. The results for the affected sub-counties should therefore be interpreted alongside their standard errors, mean squared errors and relative standard errors.

This study was restricted to the 29 sub-counties of Kenya's Coastal Region; consequently, the findings should not be interpreted as nationally representative of all Kenyan sub-counties. Although the selected auxiliary variables were supported by statistical screening, theoretical relevance and acceptable multicollinearity, the final model was limited to three predictors. Potentially important determinants, including school infrastructure, teacher availability, household learning environments and digital-literacy indicators, were either unavailable consistently at sub-county level or were not retained during model selection. Future research could extend the analysis to all Kenyan sub-counties, incorporate additional education-access, school-resource, accessibility and geospatial variables, and compare the conventional Fay–Herriot model with unit-level, Bayesian and spatial small area estimation approaches (Jean et al. 2016; WorldPop 2019).

The study therefore collectively demonstrates that the Fay–Herriot model provides a practical and statistically grounded method for producing lower-level literacy indicators where direct survey estimates are limited (Fay and Herriot 1979; Pfeffermann 2013; Rao and Molina 2015). Recent education-focused research in low- and middle-income countries similarly demonstrates the value of small area estimation for producing geographically disaggregated education indicators from household survey and auxiliary data (Wu, Dharamshi, and Wakefield 2026). The main contribution of the present study is not merely the production of sub-county estimates, but the provision of estimates accompanied by measures of uncertainty. This enables planners to distinguish between areas where low literacy is strongly indicated and those where further data collection or validation may be necessary. By integrating household-survey estimates, census-based auxiliary information and spatial mapping, the approach improves the visibility of local educational inequalities and provides a stronger basis for equitable and evidence-based targeting of literacy interventions (United Nations Statistics Division, n.d.; Corral et al. 2020).

CONCLUSION AND IMPLICATIONS

The study achieved its objectives of developing a Fay–Herriot area-level model for estimating sub-county literacy rates in Kenya's Coastal Region and presenting the estimates spatially. By combining direct literacy estimates from the 2022 Kenya Demographic and Health Survey with auxiliary variables from the 2019 Kenya Population and Housing Census, the model produced more stable and informative estimates for the 29 sub-counties studied. The findings revealed variation in literacy levels across the region. Higher estimated literacy rates were observed in Mvita, Voi, Changamwe and Mwatate, while lower rates were recorded in Tana Delta, Tana North, Tana River and Lunga Lunga. These differences demonstrate that regional and county averages can conceal important inequalities between sub-counties. The selected model also showed that the proportion of the population with basic education or above, lower-wealth status and level of urbanization were important predictors of local literacy outcomes. The spatial maps further met the study objective of communicating both the distribution and reliability of the estimates. The literacy map identified areas with high and low literacy, while the relative standard error map showed where estimates were more or less precise. Most sub-counties had acceptable precision, although estimates for Tana North, Tana Delta and Tana River required caution. The findings confirm that the Fay–Herriot model is useful for producing sub-county education indicators where direct survey estimates may be unstable. The results can support targeted adult literacy programmes, resource allocation, teacher deployment and local education planning, provided that literacy estimates are interpreted together with their precision measures.

Declarations

Conflict of interest

The authors declare that they have no competing or conflicting interests that could have influenced the conduct, analysis or reporting of this study.

Data availability statement

The 2022 Kenya Demographic and Health Survey data used in this study are available from the DHS Program upon registration and approval of a data-access request. The 2019 Kenya Population and Housing Census microdata are available from the Kenya National Bureau of Statistics subject to its data-access procedures and conditions. The derived sub-county estimates and analytical code are available from the corresponding author upon reasonable request, subject to the terms and restrictions governing the original datasets.

REFERENCES

1. Chambers, R. L., & Clark, R. G. (2012). An introduction to model-based survey sampling with applications. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198566625.001.0001
2. Chandra, H., Chambers, R., & Salvati, N. (2012). Small area estimation of proportions in business surveys. Journal of Statistical Computation and Simulation, 82(6), 783–795. https://doi.org/10.1080/00949655.2011.554834
3. UNESCO Institute for Statistics. (n.d.). Reporting indicator 4.6.1. UNESCO Institute for Statistics. https://uis.unesco.org/sites/default/files/medias/fichiers/2025/08/4.6.1_Reporting-indicator-4.6.1.pdf

4. Fay, R. E., III, & Herriot, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. Journal of the American Statistical Association, 74(366), 269–277. https://doi.org/10.1080/01621459.1979.10482505
5. Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. Statistical Science, 9(1), 55–93. https://doi.org/10.1214/ss/1177010647
6. Corral Rodas, P. A., Molina, I., & Nguyen, M. C. (2020). Pull your small area estimates up by the bootstraps (Policy Research Working Paper No. 9256). World Bank. https://doi.org/10.1596/1813-9450-9256
7. Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790–794. https://doi.org/10.1126/science.aaf7894
8. Kalton, G. (1983). Introduction to survey sampling. Sage Publications. https://doi.org/10.4135/9781412984683
9. Kenya National Bureau of Statistics. (2019a). 2019 Kenya population and housing census: Volume I: Population by county and sub-county. Kenya National Bureau of Statistics.
10. Kenya National Bureau of Statistics. (2019b). 2019 Kenya population and housing census: Volume IV: Distribution of population by socio-economic characteristics. Kenya National Bureau of Statistics.Kenya National Bureau of Statistics, & ICF. (2023). Kenya demographic and health survey 2022: Volume 1. Kenya National Bureau of Statistics and ICF.
11. Kirsch, I. S. (2001). The international adult literacy survey: Understanding what was measured. ETS Research Report Series, 2001(2), i–61. https://doi.org/10.1002/j.2333-8504.2001.tb01867.x
12. Kish, L. (1965). Survey sampling. John Wiley & Sons.
13. Mtwanguo, M. C., & Mwange, C. (2023). Use of geographic information systems (GIS) to assess literacy in rural communities in Kenya. In Proceedings of the FIG Working Week 2023. International Federation of Surveyors. https://www.fig.net/resources/proceedings/fig_proceedings/fig2023/
14. Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science, 28(1), 40–68. https://doi.org/10.1214/12-STS395
15. Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85(409), 163–171. https://doi.org/10.1080/01621459.1990.10475320
16. Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software, 91(7), 1–33. https://doi.org/10.18637/jss.v091.i07
17. Pratesi, M. (2016). Analysis of poverty data by small area estimation. John Wiley & Sons.
18. Rao, J. N. K., & Molina, I. (2015). Small area estimation (2nd ed.). Wiley. https://doi.org/10.1002/9781118735855
19. Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). Analysis of complex surveys. Wiley.
20. UNESCO. (2005). Literacy for life: EFA global monitoring report 2006. UNESCO.
21. UNESCO. (2023). Global education monitoring report 2023: Technology in education: A tool on whose terms? UNESCO.
22. United Nations Statistics Division. (n.d.). Toolkit on using small area estimation for SDGs. United Nations. https://www.unsdglearn.org/resources/toolkit-on-using-small-area-estimation-for-sdgs/
23. World Bank. (2022, June 7). Kenya Economic Update: Lessons from Kenya's education reforms. World Bank. https://www.worldbank.org/en/country/kenya/publication/lessons-from-kenya-s-education-reforms
24. WorldPop. (2019). WorldPop global population data. University of Southampton. https://www.worldpop.org
25. Wu, Y., Dharamshi, A., & Wakefield, J. (2026). Small area estimation of education levels in low- and middle-income countries. Annals of Applied Statistics, 20(1), 833–855. https://doi.org/10.1214/25-AOAS2135