Article

Beyond Farm Size: Spatial Determinants of Cocoa Productivity in Ghana's Ashanti Region

Welson Bentum [1], Emmanuel Odame Owiredu [2], Nana Kena Frimpong [3] Michael Asante Ofosu [4], Esther Boateng [5], Kwaku Obeng Owusu [6] and Agyeman Shadrack [7,*]

[1] Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; bwelson523@gmail.com, mike.ofosu16@gmail.com

Abstract

Cocoa yields in Ghana have declined 23% since 2020 despite favorable prices, yet the spatial dimensions of this productivity crisis remain poorly understood. We analyzed 2,612 georeferenced cocoa farms across 10 districts in Ghana's Ashanti Region (5°30′N–7°45′N, 0°15′W–2°25′W) to identify spatial patterns and determinants of yield variation. We integrated GIS-based spatial analysis with multivariate spatial regression modeling, applying Moran's I statistics, Ripley's K-function. Results revealed strong spatial clustering of productivity (Moran's I = 0.594, $p < 0.001$), indicating that unobserved spatially structured factors significantly shape yield outcomes, with seven distinct clusters identified through point pattern analysis. We employed a two-hurdle spatial regression framework to analyze cocoa yield determinants, addressing the substantial zero inflation in production data. Over 50% of sampled farms (n = 2,612) reported zero yields during the 2022/2023 season, reflecting distinct economic and agronomic processes governing participation versus productivity decisions. Hurdle 1 uses spatial autoregressive probit regression to model the binary participation decision, while Hurdle 2 employs spatial error regression among producing farms to identify conditional yield determinants. Results reveal strong spatial dependence at both decision stages, with spatial autoregressive coefficient $\rho = 0.698$ in the participation model and spatial error coefficient $\lambda = 0.690$ in the productivity model. Tree density exhibits super-elastic effects (elasticity = 1.29), while farm size shows an inverse productivity relationship (elasticity = -0.0287). The model explains 92.7% of yield variation on the original scale using Duan's smearing estimator for back-transformation. Policy implications emphasize geographically targeted interventions, smallholder intensification through replanting programs, and sustained extension engagement.

Keywords: cocoa productivity, spatial autocorrelation, GIS, inverse farm-size relationship, smallholder agriculture, Ghana, precision agriculture, yield gap analysis

Received:
Revised:
Accepted:
Published:

Citation: . . *Journal Not Specified* **2025**, *1*, 0.

1. Introduction

Despite decades of agricultural development efforts, productivity growth in Sub-Saharan Africa has lagged behind population growth, with cocoa, one of the region's most valuable export crops experiencing precipitous declines since 2020. WWest Africa produces 70% of global cocoa supply, with Ghana historically ranking second worldwide [60]. Ghana achieved peak production of 1.047 million metric tons in 2020, but output has since declined 23% to approximately 800,000 metric tons in 2023/2024 [24], the lowest level since 2008 when production averaged 680,000 metric tons annually. This represents a return to pre-2010 production levels despite favorable price conditions. This decline threatens

 https://doi.org/10.3390/1010000

both national revenue as cocoa contributes $1.7 billion annually to Ghana's exports and the livelihoods of 800,000 smallholder families [59]. Producer prices increased 147% between 2023-2025, yet production continued to decline, suggesting that price signals alone cannot reverse productivity losses [25,52]. This disconnect between prices and production suggests that non-economic constraints, particularly spatial heterogeneity in farm management and environmental conditions drive yield outcomes. Despite global price increases of 157% in 2024, Ghana's export earnings fell 25.4% to $1.7 billion—the lowest level in 15 years, indicating that volume declines outpaced price gains [36]. Average smallholder yields of 400-554 kg/ha represent only 21-29% of experimentally achieved potentials (1,891-3,246 kg/ha) [2,11,32], indicating that agronomic management rather than environmental constraints limit productivity. Understanding the spatial distribution of this yield gap is critical for targeting interventions effectively. Despite the documented productivity crisis, spatial analysis of cocoa farming systems remains limited. Existing studies focus on land suitability modeling [47,50] or climate change projections [12,55] but do not explain observed yield variation across operating farms. We address this gap by utilizing spatial analytics to identify yield variations and recommend targeted farm-level operations for yield improvements. This study examines spatial patterns in cocoa productivity across 2,612 farms in Ghana's Ashanti Region to: (1) quantify the degree and scale of spatial clustering, (2) identify farm-level and geographic determinants of yield variation, and (3) assess whether productivity differences reflect environmental constraints or management practices. Abdulai et al. (2020), analyzing 3,827 farms, found that environmental factors account for only 7% of yield variation, while farm management explains 80%, demonstrating productivity gaps stem from operational deficiencies rather than environmental limitations [1]

2. Literature Review

Previous research on cocoa productivity has emphasized biological constraints; aging trees, pests, and climate change, while largely overlooking spatial heterogeneity in farm management as a yield determinant [49,56].

2.1. The Yield Gap paradox in Ghanaian Cocoa

Despite favorable growing conditions and rising prices, Ghanaian cocoa yields have stagnated at 400-554 kg/ha, far below experimental potentials of 1,891-3,246 kg/ha [2,11,32]. Critically, Abdulai et al. (2020) found that environmental factors explain only 7% of yield variation among 3,827 Ghanaian cocoa farms, while management practices explain 80% [1]. This finding challenges conventional emphasis on climate and soil constraints, suggesting that productivity gaps reflect controllable factors. However, the spatial distribution of this management effect (whether it clusters geographically or distributes randomly) remains unknown.

2.2. Spatial Analysis in Agricultural Systems: Underutilized in Cocoa

Spatial autocorrelation and clustering analysis have revealed non-random productivity patterns in crops ranging from maize [39] to rice [10], with neighboring farms exhibiting similar yields due to shared environmental conditions, institutional access, or knowledge spillovers [22]. Yet cocoa productivity research has largely ignored spatial dimensions. Existing studies concentrate on land suitability modeling, where GIS-based assessments identify areas suitable for cocoa cultivation based on climate, soil, and topography [4,33,47] but do not analyze yield variation across operating farms. Climate change projection studies employ spatial models to forecast future cocoa suitability under warming scenarios [20,42,55] but do not explain current productivity patterns. Supply chain mapping research uses remote sensing to track deforestation and cocoa expansion [3,6] but does not link land use to farm-level productivity. Critically, no studies have applied point pattern analysis or spatial autocorrelation methods to yield variation, a critical prerequisite for targeting interventions geographically.

2.3. The Inverse Farm-Size Productivity Relationship

Agricultural economics literature consistently documents that smaller farms achieve higher yields per hectare than larger operations, often termed the "inverse relationship" (IR) [16,28,34]. Proposed mechanisms include labor intensity, where smallholders apply more family labor per hectare [28], management quality differences as smaller farms receive more intensive oversight [30,41], and land quality selection patterns where small farms may occupy better soils. Barrett et al. (2010) confirmed the IR across multiple African crops [16], while Debrah and Agyare (2022) found an L-shaped relationship in Ghanaian agriculture [28]. However, whether this pattern holds for cocoa (a tree crop with different management requirements than annuals) and whether it varies spatially remains untested.

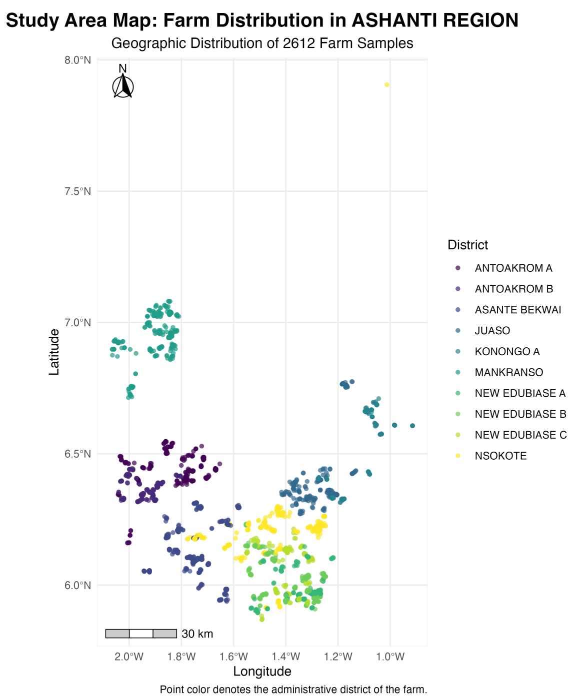
2.4. Research Gaps and Study Positioning

Existing research documents low cocoa productivity in Ghana, but does not quantify spatial clustering patterns in observed yields, test whether the inverse farm size-productivity relationship applies to cocoa, or identify whether geographic location (beyond measured environmental variables) influences yield outcomes. This study addresses these gaps by applying spatial point pattern analysis and local indicators of spatial association to geocoded farm-level data, thereby revealing whether productivity constraints exhibit geographic structure that could inform targeted extension programming.

3. Materials and Methods

3.1. Research Design and Study Area

We conducted this study in Ghana's Ashanti Region, the country's leading cocoa-producing area.Ghana lies in West Africa between 4-11°N and 4°W-2°E, spanning 238,539 km^2. The Ashanti Region occupies the forest-savanna transition zone, with bimodal rainfall (1,200-1,500 mm annually) and humid tropical conditions suitable for cocoa cultivation.



We sampled 2,612 cocoa farms across 10 districts selected to represent the region's agro-ecological and socioeconomic diversity: Asante Bekwai, Nsokote, Antoakrom A, New Edubiase B, Mankranso, Konongo A, New Edubiase A, Juaso, Antoakrom B, and New Edubiase C. The sampled communities spanned 92 distinct localities within these districts.

3.2. Statistical Analysis

3.2.1. Spatial Autocorrelation

We assessed spatial clustering in cocoa yields using global Moran's I statistic, which measures the degree to which farms with similar productivity levels are geographically proximate [7]. Moran's I ranges from -1 (perfect dispersion) to +1 (perfect clustering), with values near zero indicating spatial randomness. We defined spatial relationships using an inverse distance weighting matrix with a maximum threshold of 10 km, ensuring each farm had at least 7 neighbors. Statistical significance was evaluated using 5000 Monte Carlo permutations under the null hypothesis of spatial randomness. All spatial analyses were conducted in R 4.3.1 using the *spdep* package [17].

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

Where:

- n is the number of spatial units,
- x_i and x_j are the values of the variable at locations i and j,
- $\bar{x}$ is the mean of the variable,
- w_{ij} is the spatial weight between locations i and j, typically based on distance or adjacency.

Interpretation:

- $I > 0$: Positive autocorrelation (clustering of similar values, e.g., high pollution in nearby areas).
- $I < 0$: Negative autocorrelation (dispersion, e.g., high and low values alternate).
- $I \approx 0$: No spatial pattern (random distribution).
- Statistical significance is tested using a z-score under the null hypothesis of no auto-correlation, often assuming normality or using permutation tests.

3.2.2. Point Pattern Analysis

We used Ripley's K-function to assess whether farm locations clustered at multiple spatial scales [53]. The K-function compares observed inter-farm distances to those expected under complete spatial randomness (CSR), with $K(r) > \pi r^2$ indicating clustering at distance r. We computed K-functions for distances from 0.5 to 25 km using edge corrections [29]. Additionally, we calculated Average Nearest Neighbor (ANN) statistics to test for localized clustering, comparing observed mean nearest-neighbor distances to expected distances under CSR. Point pattern analyses used the *spatstat* package in R [15].

Ripley's K function:

$$K(d) = \frac{A}{n^2} \sum_i \sum_{j \neq i} I(d_{ij} \leq d)$$

Where:

- A is the study area,
- n is the number of points,
- $I(d_{ij} \leq d)$ is an indicator function (1 if $d_{ij} \leq d$, else 0).

Nearest Neighbor Function:

$$NNI = \frac{\bar{d}_{\text{obs}}}{\bar{d}_{\text{exp}}}, \quad \bar{d}_{\text{exp}} = \frac{1}{2\sqrt{n/A}}$$

Where $\bar{d}_{\text{obs}}$ is the average observed nearest neighbor distance.

3.2.3. Spatial Clustering

We identified distinct productivity zones using hierarchical spatial clustering on farm coordinates and yield data. We applied Ward's linkage method with Euclidean distance to group farms into clusters based on both geographic proximity and yield similarity. The optimal number of clusters was determined using the elbow method and silhouette scores. We validated cluster distinctiveness using ANOVA with post-hoc Tukey tests to confirm significant yield differences among zones ($\alpha = 0.05$)

3.3. Feature Importance

To determine the relative contribution of each predictor to the model output, we employed Random Forest for the yield per hectare prediction task. Feature importance was computed based on the mean decrease in impurity (MDI) across all trees in the ensemble. These raw importance values were normalized to percentages to allow for easier comparison and interpretation.

3.4. Interactive Dashboard

An interactive spatial dashboard was developed using `Folium` to visualize farm-level cocoa yield, productivity categories, and production intensity. Each farm i was represented by its coordinates (ϕ_i, λ_i) and yield y_i.
Productivity hotspots were identified as farms exceeding the 80th percentile of yield:

$$H = \{i : y_i > Q_{0.80}(y)\}.$$

Production intensity was visualized using a kernel heat map:

$$I(x) = \sum_{i=1}^{n} K_h(x - s_i)v_i,$$

where v_i is production volume, s_i the farm location, and K_h a Gaussian kernel with bandwidth h.
The dashboard included multiple layers, productivity classes, spatial clusters, and production hotspots with interactive legend and controls to support spatial interpretation of yield patterns.

3.5. Conceptual Framework: Two-Hurdle Spatial Model

Cocoa yields in the study area exhibit pronounced zero-inflation, with over 50% of farms producing zero yields. This zero-inflation structure violates standard regression assumptions and reflects two distinct economic and agronomic processes: (1) a **participation decision** (whether farms engage in cocoa production at all), and (2) a **productivity decision** (conditional on producing), thus how much output is achieved per hectare. We employed a two-hurdle spatial regression framework to disentangle these processes [26]. Hurdle 1 uses spatial autoregressive probit regression to model the binary participation decision (produce versus non-produce), identifying factors associated with farm engagement in cocoa production. Hurdle 2 uses spatial error regression among producing farms only to identify determinants of yield intensity, conditional on production.

3.6. Hurdle 1: Spatial Autoregressive Probit for Production Participation

3.6.1. Model Specification

We define a binary outcome variable:

$$Y_i = \begin{cases} 1 & \text{if farm } i \text{ produced cocoa (yield} > 0 \text{ kg/ha)} \\ 0 & \text{if farm } i \text{ produced no cocoa (yield} = 0 \text{ kg/ha)} \end{cases}$$

The spatial autoregressive probit (SAR Probit) model specifies:

$$y_i^* = \rho \mathbf{W} \mathbf{y}^* + \mathbf{X}_i \boldsymbol{\beta} + u_i$$

$$y_i = \mathbb{1}(y_i^* > 0)$$

where:

y_i^* is the latent propensity to produce for farm i

ρ is the spatial autoregressive coefficient capturing spillover effects from neighbors' participation decisions

$\mathbf{W}$ is an $n \times n$ row-standardized spatial weights matrix (k=7 nearest neighbors)

$\mathbf{X}_i$ is a vector of explanatory variables

$\boldsymbol{\beta}$ is a vector of structural parameters

$u_i \sim N(0, 1)$ is a standard normal error term

The spatial parameter ρ captures whether farms are more likely to produce if their neighbors also produce, reflecting either shared environmental advantages, knowledge spillovers, or coordinated market participation. Positive ρ indicates positive spatial spillovers with clustering of producers, while negative ρ would indicate dispersion potentially driven by competitive effects or resource rivalry.

3.6.2. Spatial Weights Matrix

Spatial weights were constructed using k-nearest neighbors (k=7) based on Euclidean distance between farm coordinates. The k=7 specification ensures each farm has sufficient neighbors while maintaining local spatial structure, balancing estimation efficiency and local relevance. The weights matrix $\mathbf{W}$ was row-standardized to unit row sums:

$$\sum_{j=1}^{n} w_{ij} = 1 \quad \forall i$$

This standardization allows ρ to be interpreted as the strength of spatial dependence, with values closer to 1 indicating stronger spatial correlation.

3.6.3. Estimation Method

The SAR Probit model was estimated via Markov Chain Monte Carlo (MCMC) using Bayesian methods implemented in the `spatialprobit` package in R [58]. MCMC is necessary because maximum likelihood estimation is computationally intractable for spatial probit models due to the high-dimensional integrals involved [43]. The MCMC estimation proceeded with 5000 draws following a burn-in period of 100 iterations. Results are reported as posterior means and 95% credible intervals derived from post burn-in MCMC samples. Statistical significance is assessed by whether the 95% credible interval excludes zero, equivalent to $p < 0.05$ under frequentist conventions.

3.7. Hurdle 2: Spatial Error Model for Conditional Productivity

3.7.1. Model Specification

Hurdle 2 analysis includes only farms with positive yields: n = 1,178, representing 45.1% of the full sample. This subsample consists of actively producing farms where yield intensity is the outcome of interest. Preliminary OLS diagnostics on the positive-yield sample revealed substantial heteroskedasticity (Breusch-Pagan test: $p < 0.001$). We evaluated three candidate functional forms:

1. Original scale: $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$
2. Log-linear: $\log(Y_i) = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$
3. Box-Cox transformation: $Y_i^{(\lambda)} = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$

Model selection criteria included AIC, residual normality (Shapiro-Wilk test), homogeneity of variance (Breusch-Pagan test), and interpretability. Given the superior interpretability of log-transformed coefficients as elasticities and semi-elasticities, and the adequate performance of the log specification in addressing heteroskedasticity, we selected the **log-linear** specification for all subsequent spatial models.

3.7.2. OLS Baseline and Lagrange Multiplier Tests

We first estimated an OLS baseline model on the positive-yield sample:

$$\begin{aligned}\log(\text{yield}_i) = \beta_0 &+ \beta_1(\text{farm size})_i + \beta_2(\text{surface mapped})_i + \beta_3(\text{nb farms 2122})_i \\ &+ \beta_4(\text{received remediation})_i + \beta_5(\text{covered 2122})_i + \beta_6(\text{farmer coaching})_i \\ &+ \beta_7(\text{fdp previously})_i + \beta_8(\text{map farm})_i + \beta_9(\text{gender})_i \\ &+ \beta_{10}(\text{household size})_i + \beta_{11}(\text{nb farms})_i + \beta_{12}\log(\text{yield per tree})_i \\ &+ \beta_{13}(\text{pods per tree})_i + \beta_{14}\log(\text{tree density} + 1)_i + \varepsilon_i\end{aligned}$$

Residuals from the OLS model were tested for spatial autocorrelation using Lagrange Multiplier (LM) tests implemented in `lm.LMtests()` from the `spdep` package [18]. Four LM test variants were computed:

- LM_{lag}: Tests for spatial lag dependence
- LM_{error}: Tests for spatial error dependence
- Robust LM_{lag}: Robust version accounting for presence of the other
- Robust LM_{error}: Robust version accounting for presence of the other

Decision rule: If robust LM_{error} is significant and robust LM_{lag} is not significant, the Spatial Error Model (SEM) is appropriate. If robust LM_{lag} is significant and robust LM_{error} is not significant, the Spatial Lag Model (SLM) is appropriate. If both are significant, the larger robust test statistic indicates the more appropriate specification [8].

3.7.3. Spatial Model Specifications

(2) Spatial Lag Model (SLM):

$$\begin{aligned}\log(\text{yield}_i) = \rho \sum_j w_{ij} \log(\text{yield}_j) &+ \beta_0 + \beta_1(\text{farm size})_i + \beta_2(\text{surface mapped})_i \\ &+ \beta_3(\text{nb farms 2122})_i + \beta_4(\text{received remediation})_i + \beta_5(\text{covered 2122})_i \\ &+ \beta_6(\text{farmer coaching})_i + \beta_7(\text{fdp previously})_i + \beta_8(\text{map farm})_i \\ &+ \beta_9(\text{gender})_i + \beta_{10}(\text{household size})_i + \beta_{11}(\text{nb farms})_i \\ &+ \beta_{12}\log(\text{yield per tree})_i + \beta_{13}(\text{pods per tree})_i + \beta_{14}\log(\text{tree density} + 1)_i + \varepsilon_i,\end{aligned}$$

where ρ is the spatial autoregressive coefficient, and w_{ij} are elements of the spatial weights matrix W.

(3) Spatial Error Model (SEM):

$$\begin{aligned}\log(\text{yield}_i) = \beta_0 &+ \beta_1(\text{farm size})_i + \beta_2(\text{surface mapped})_i + \beta_3(\text{nb farms 2122})_i \\ &+ \beta_4(\text{received remediation})_i + \beta_5(\text{covered 2122})_i + \beta_6(\text{farmer coaching})_i \\ &+ \beta_7(\text{fdp previously})_i + \beta_8(\text{map farm})_i + \beta_9(\text{gender})_i \\ &+ \beta_{10}(\text{household size})_i + \beta_{11}(\text{nb farms})_i + \beta_{12}\log(\text{yield per tree})_i \\ &+ \beta_{13}(\text{pods per tree})_i + \beta_{14}\log(\text{tree density} + 1)_i + u_i,\end{aligned}$$

$$u_i = \lambda \sum_j w_{ij} u_j + \varepsilon_i,$$

where u_i represents the spatially structured error component, λ is the spatial error coefficient, w_{ij} are elements of the spatial weights matrix W, and $\varepsilon_i \sim N(0, \sigma^2)$.

All models were estimated via maximum likelihood using the `spatialreg` package [19]. Model selection was based on: (1) Akaike Information Criterion (AIC), (2) log-likelihood, (3) Moran's I test on residuals, and (4) root mean squared error (RMSE) on the back-transformed scale. The model with the lowest AIC and non-significant Moran's I on residuals was selected as the preferred specification. We also checked for multicollinearity using variance inflation factors (VIF < 5 for all predictors).

3.7.4. Back-Transformation to Original Scale Using Smearing Estimator

Since the dependent variable and some independent variables were log-transformed, predictions required back-transformation to the original kg/ha scale for interpretation. Naive exponential back-transformation introduced bias due to Jensen's inequality:

$$E[Y|X] = E[\exp(\log Y|X)] \neq \exp(E[\log Y|X]) = \exp(\mathbf{X}\boldsymbol{\beta})$$

To obtain unbiased predictions, we employed **Duan's smearing estimator** [31]:

$$\hat{Y}_i = \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \times \phi$$

where the smearing factor is:

$$\phi = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i)$$

and $\hat{u}_i$ are the residuals from the spatial error model on the log scale. This non-parametric correction does not assume a specific error distribution and provides consistent estimates even in the presence of heteroskedasticity [45].

3.7.5. Model Diagnostics

Comprehensive post-estimation diagnostics were conducted:

- **Spatial autocorrelation:** Moran's I test on SEM residuals to verify adequate spatial specification
- **Heteroskedasticity:** Breusch-Pagan test of residuals against fitted values
- **Influence diagnostics:** Cook's distance and leverage plots
- **Spatial residual patterns:** Geographic mapping of residuals

3.7.6. Performance Metrics

Model fit was assessed on the back-transformed (original kg/ha) scale using:

$$R^2 = 1 - \frac{\sum_i (\text{observed}_i - \text{predicted}_i)^2}{\sum_i (\text{observed}_i - \bar{\text{observed}})^2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (\text{observed}_i - \text{predicted}_i)^2}$$

$$\text{MAE} = \frac{1}{n} \sum_i |\text{observed}_i - \text{predicted}_i|$$

These metrics assess practical model performance in interpretable units.

3.8. Data Collection and Variable Construction

3.8.1. Sampling Frame and Survey Design

Farm-level data were obtained through a partnership with an international cocoa sustainability program operating in Ghana's Ashanti Region. The program maintains a comprehensive database of registered cocoa farmers participating in certification and capacity-building initiatives. Data collection occurred during the 2023/2024 main cocoa season (October 2023–March 2024), capturing production outcomes from the 2022/2023 harvest year.

A stratified random sampling design was employed across 10 target districts, selected to represent the agro-ecological and socioeconomic diversity of the Ashanti cocoa belt. District-level sampling quotas were proportional to the registered cocoa farmer population in each district, as documented in Ghana Cocoa Board licensing records. Within each district, communities were randomly selected, and all registered farmers in selected communities were invited to participate, achieving a response rate of 87.3%.

Trained enumerators administered structured household surveys in local languages (Twi, Asante), capturing: (1) household demographics and composition, (2) farm characteristics and land tenure, (3) production practices and input use, (4) harvest outcomes for the previous season, and (5) participation in extension and farm support programs. All interviews were conducted face-to-face at farmers' residences or farm locations. Data quality was ensured through supervisor spot-checks, logic constraint programming in the survey instrument (KoBoToolbox platform), and daily review of submitted surveys.

3.8.2. Geographic Data Collection

Farm locations were recorded using handheld GPS devices (Garmin eTrex 10) with horizontal accuracy typically below 5 meters under open sky conditions. For each farm, enumerators recorded waypoints at farm boundaries or central locations, depending on farm accessibility and canopy cover. Coordinates were logged in decimal degrees using the WGS84 datum.

Quality control procedures excluded observations with: (1) horizontal dilution of precision (HDOP) > 8, indicating poor satellite geometry; (2) positional accuracy estimates > 15 meters, (3) elevation values outside the plausible range for the region (200–800 m above sea level), or (4) coordinates falling outside district administrative boundaries. Coordinates were projected to UTM Zone 30N for spatial analysis requiring metric distances.

Farm area was measured using one of two methods: (1) walking the farm perimeter with GPS track logging for farms < 5 hectares, or (2) farmer-reported area verified against land registration documents or purchase agreements where available. For GPS-measured farms, polygon area was calculated in QGIS 3.28 using geodesic calculations accounting for earth curvature. Comparison of GPS-measured versus farmer-reported areas ($n = 873$ farms with both measures) revealed a mean absolute percentage error of 18.7%, consistent with previous validation studies in West African smallholder systems.

3.8.3. Yield Measurement and Calculation

Cocoa production was quantified through structured farmer recall of total bags harvested during the 2022/2023 main and light crop seasons. Ghana's cocoa marketing system uses standardized 62.5 kg bags (net wet or dried beans), facilitating conversion to kilograms. Farmers reported bags sold to licensed buying companies, which was cross validated against purchase receipts where available ($n = 1,847$ farms, 70.7% of sample). Yield per hectare was calculated as:

$$\text{Yield per hectare} = \frac{\text{Total production (kg)}}{\text{GPS-measured farm area (ha)}}$$

3.9. Explanatory Variables

- *Farm size*: GPS-measured cultivated area in hectares.
- *Tree density*: Number of productive cocoa trees per hectare, calculated from tree counts conducted during farm mapping.
- *Number of farms per household*: Count of spatially distinct cocoa plots operated by the household.

Management and intervention variables:

- *Farmer coaching participation*: Binary indicator of enrollment in the extension program (defined as attending ≥ 4 training sessions in the previous 12 months)
- *Farm Development Program (FDP) participation*: Current enrollment in intensive support program providing subsidized inputs, pruning services, pest management, or agricultural credit.
- *Previous FDP participation*: Historical enrollment in FDP (previous 1-5 years), capturing legacy effects of past interventions.
- *Received remediation*: Binary indicator of participation in farm rehabilitation activities (replanting, infilling, or disease management interventions)

Household characteristics:

- *Household size*: Number of members residing in the household
- *Gender*: Sex of the primary farm decision-maker (1 = male, 0 = female)

Geographic variables

- *Latitude and longitude*: Decimal degree coordinates used to capture unobserved spatial gradients in environmental conditions, market access, or institutional support

3.9.1. Data Cleaning and Final Sample

The initial dataset comprised 6,693 surveys farms. Observations were excluded based on the following criteria applied sequentially:

1. Missing yield data or incomplete production records ($n = 2,847$; 42.9%)
2. Implausible yields $> 5,000$ kg/ha flagged as measurement or reporting errors ($n = 2$)
3. GPS coordinate errors: HDOP > 8, accuracy > 15m, or coordinates outside district boundaries ($n = 823$; 12.4%)
4. Missing or incomplete responses to core survey modules ($n = 355$, 5.3%)

These exclusions reduced the analytic sample to 2,612 farms, representing a 39.3% retention rate relative to the initial survey sample. Comparison of retained versus excluded observations on observable characteristics (household size, farm size, district) revealed no statistically significant differences (two-sample t-tests, all $p > 0.10$), suggesting that attrition did not introduce systematic bias.

3.9.2. Ethical Considerations

This study utilized secondary data collected as part of routine program monitoring activities conducted under a partnership agreement between the implementing organization and participating farmers. Farmers provided informed consent for data collection and use for research purposes, including anonymized analysis and publication, at the time of program enrollment. No personal identifying information was retained in the analytical dataset. The study protocol was reviewed and deemed exempt from full institutional review by the Ethics Committee of Kwame Nkrumah University of Science and Technology, as it involved analysis of de-identified, pre-existing data.

4. Results and Discussion

4.1. District Yield Statistics

Table 7 presents the distribution of cocoa yields v across districts, reflecting pronounced spatial heterogeneity in production performance.

Table 1. Yield Statistics

District	Number of farms	**Mean**	**Median**	**Std. Dev.**
Antoakrom A	378	35.31	5.80	51.78
Antoakrom B	370	26.09	20.43	26.95
Asante Bekwai	293	58.14	8.26	71.53
Juaso	282	0.00	0.00	0.00
Konongo A	266	0.00	0.00	0.00
Mankranso	221	0.00	0.00	0.00
New Edubiase A	221	100.38	85.62	80.22
New Edubiase B	214	116.95	129.43	80.24
New Edubiase C	213	0.00	0.00	0.00
Nsokote	154	34.06	23.45	36.82

Over half of the districts (Juaso, Konongo A, Mankranso, and New Edubiase C) recorded zero mean and median yields, indicating widespread farm-level non-production during the survey season. Among producing districts, New Edubiase B exhibited the highest mean yield (116.95 kg/ha), followed by New Edubiase A (100.38 kg/ha). Yields in other producing districts such as Antoakrom A, Antoakrom B, Asante Bekwai, and Nsokote were substantially lower, with means ranging between 26.09 and 58.14 kg/ha. The pattern suggests a clear divide between non-producing and moderately productive areas, pointing to a threshold management effect rather than a smooth gradient in productivity.

4.2. Descriptive Statistics

The descriptive statistics presented in Table 2 provide an overview of the key variables characterizing the cocoa farming systems in the study area. The data encompasses 2,612 farms and reveals important patterns about farm structure, productivity, and spatial distribution.

Table 2. Descriptive Statistics

Sample characteristics	Value					
Number of observations (N)	2,612					
Number of districts	10					
Number of communities	92					
Survey period	Oct 2023 – Mar 2024					
Variable	**Mean**	**Std. Dev.**	**Min**	**Median**	**Max**	**Unit**
Household Characteristics						
Household size	5.20	2.07	1.00	5.00	13.00	persons
Farm Characteristics						
Farm size	8.45	5.13	0.33	7.50	27.13	hectares
Farm size (GPS-mapped)	2.64	2.89	0.00	1.73	29.25	hectares
Number of farms/household	1.30	0.66	1.00	1.00	7.00	count
Productive cocoa trees	31.16	55.40	0.00	12.00	266.00	count
Tree density	5.26	11.91	0.00	1.86	358.21	trees/ha
Cocoa pods counted	60.66	56.75	0.00	40.00	330.00	count
Production Indicators						
Yield per hectare	40.45	63.77	0.00	0.00	233.92	kg/ha
Total harvest estimate	618.91	700.76	0.00	412.50	7500.00	kg
Bags harvested	19.77	84.97	0.00	6.00	1916.00	bags
Bags sold	17.14	78.87	0.00	5.00	1916.00	bags
Yield per tree	31.17	67.99	0.00	0.00	1456.50	kg
Pods per tree	4.31	4.15	0.00	2.74	33.60	count
Geographic Data						
Latitude	6.31	0.29	5.87	6.28	7.91	decimal deg
Longitude	-1.56	0.28	-2.07	-1.52	-0.91	decimal deg
GPS accuracy	7.84	1.80	0.76	8.30	10.00	meters

The median yield of 0.00 kg/ha reveals that over 50% of sampled farms (n=1,434, 54.9%) produced no cocoa during the survey season. Among producing farms (n=1,178, 45.1%), mean yield was 40.45 kg/ha (SD=63.77), ranging from 0.01 to 233.92 kg/ha. This bimodal distribution (farms either producing nothing or producing moderately) suggests a threshold management effect rather than a continuous productivity gradient. We analyze both the full sample (including zeros) using spatial probit and producing farms only in subsequent models.

4.3. Spatial Autocorrelation (Global Moran's Test)

The Moran's I test under randomization was conducted to formally assess the degree of spatial autocorrelation in property values. The results are summarized in Figure 1.

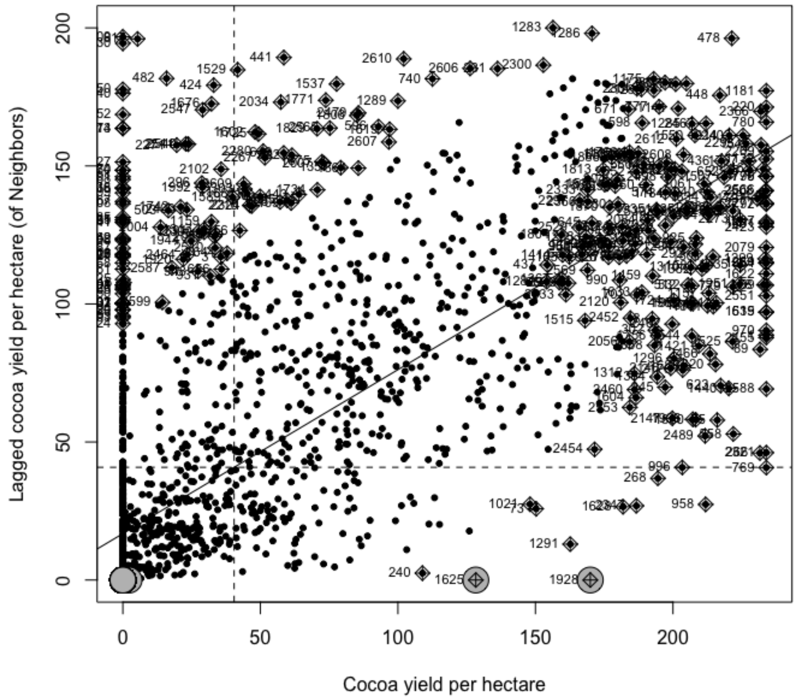


Figure 1. Moran's I statistic: 0.594, Expected value: −0.00038, Standard deviate: 60.959, p-value: $< 2.2 \times 10^{-16}$

The Moran's I statistic is 0.594, which is much larger than the expected value under spatial randomness (−0.00038).The standardized test statistic ($z = 60.959$) far exceeds the critical value for $\alpha = 0.001$ ($z^* = 3.29$), providing overwhelming evidence against spatial randomness ($< 2.2 \times 10^{-16}$). Thus, cocoa yields exhibit strong positive spatial autocorrelation, implying that cocoa farms with similar yields are geographically close to one another rather than being randomly distributed.This clustering pattern aligns with documented spatial autocorrelation in other crop yields [10,23,39], suggesting that cocoa productivity is influenced by localized factors such as shared environmental conditions, knowledge spillovers among neighboring farmers, or spatially correlated infrastructure access. The strong spatial dependence indicates that neighboring farms share similar environmental conditions, soil characteristics, or access to agricultural support services. This clustering has important implications for intervention design: geographically targeted approaches rather than uniform nationwide programs are likely more effective, as productivity patterns reflect shared access to extension services, input markets, pest pressures, and microclimate conditions that transcend individual farm decisions.

4.4. Point Pattern Analysis

The observed K-function (solid line) exceeds the theoretical expectation under complete spatial randomness (dashed line) across all distances (0-0.25 degrees ≈ 0-28 km), indicating significant clustering of cocoa farms at multiple spatial scales. The increasing divergence with distance reflects progressively stronger aggregation at regional scales.

Ripley's K-function analysis confirmed non-random clustering across all spatial scales (Figure 2). The observed K(r) exceeded the CSR expectation by 643% at 5 km and 215% at 20 km, indicating that clustering intensifies at narrower or local geographic scales. The extreme magnitude of clustering likely reflects structural land-use organization rather than random concentration of farms. In many cocoa-growing regions, land inheritance patterns, historical settlement processes, and limited availability of suitable land collectively produce dense spatial clusters [54]. This pattern suggests that both local factors (e.g., microclimate, soil conditions) and regional factors (e.g., market access, extension service coverage) contribute to farm location decisions. An increasing gap with distance implies that clustering is not only present at small spatial scales but persists and becomes more pronounced across larger neighborhoods, indicating spatially extensive aggregation patterns. This clustering pattern aligns with agricultural settlement patterns typically observed in developing countries, where farmers tend to establish farms in proximity to one another due to shared access to markets, infrastructure, and social networks. The persistence of clustering across multiple scales suggests that both local factors (such as soil quality and microclimates) and regional factors (such as market access and extension services) contribute to the spatial organization of cocoa production in the study area. Such multi-scale clustering has been reported in other agricultural contexts, with researchers noting that spatial patterns reflect both historical settlement processes and contemporary economic incentives [22].

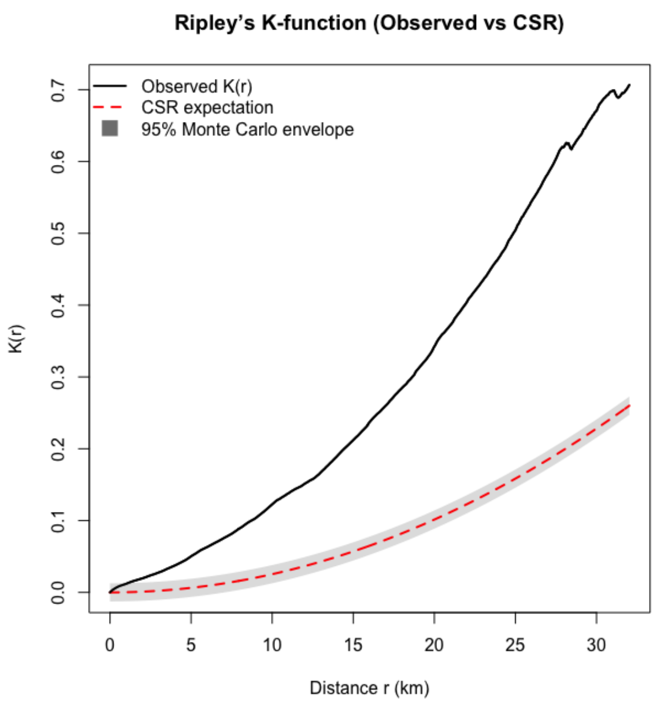


Figure 2. Ripley's K-function analysis of spatial clustering

The nearest neighbor G-function (Figure 3) revealed pronounced short-distance clustering, with 90% of farms located within 0.4 km of another farm compared to 5% expected under CSR (Kolmogorov–Smirnov test: D = 0.85, $p < 0.001$). This tight spatial aggregation has implications for pest and disease transmission, as pathogens such as *Phytophthora* (black pod) and cocoa swollen shoot virus spread more efficiently through contiguous farm networks [5,46]

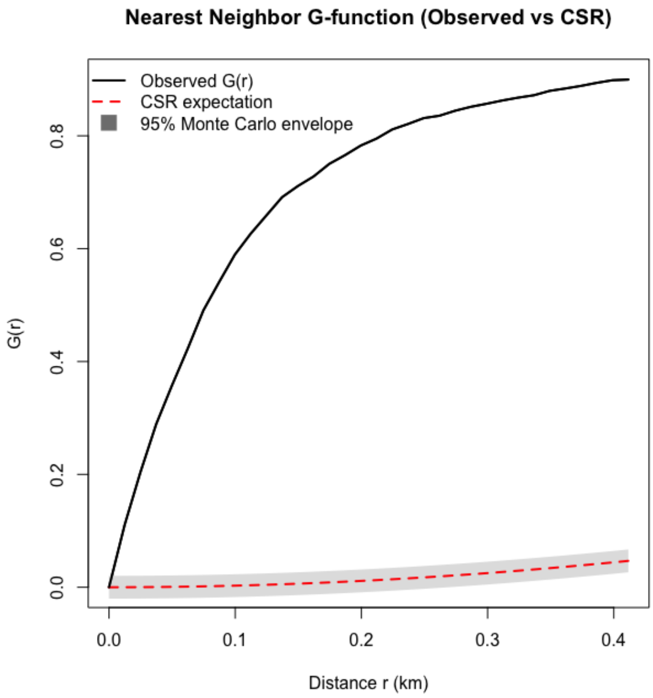


Figure 3. Nearest Neighbor G function

4.5. Spatial Clustering of Cocoa Farms

Figure 4 presents the spatial clustering of cocoa farms across the study area, revealing seven distinct clusters. Each colored group on the map represents farms that are geographically close and potentially share similar growing conditions, practices, or challenges.

Figure 4. Spatial clusters of farm productivity

The seven identified clusters exhibited significant yield heterogeneity, with Cluster 5 (New Edubiase B/New Edubiase A region) achieving the highest mean productivity (171.29 kg/ha) and Cluster 1 (Konongo A/Juaso area) the lowest (0 kg/ha) table 3. Clusters 3, 4 and 7 formed a contiguous moderate-productivity zone in the southwestern portion of the study area, suggesting favorable agro-ecological conditions or stronger extension service presence in this region. In contrast, Cluster 0 and 1's geographic isolation in the northern and eastern extent respectively, corresponded with both lower yields and lower farm density, possibly reflecting marginal growing conditions or limited infrastructure. The spatial distribution of these clusters, with distinct geographic boundaries rather than gradual transitions supports a zone-based approach to extension service delivery rather than uniform regional programming.

Table 3. Characteristics of Identified Productivity Zones

cluster	N farms	Mean yield (kg/ha)	SD	Min	Max	Dominant Districts
0	278	0.30	3.87	0.00	59.88	Mankranso, Antoakrom A
1	285	0.00	0.00	0.00	0.00	Konongo A, Juaso
2	295	127.85	48.66	36.08	233.92	Asante Bekwai, Nsokote
3	652	25.35	26.65	0.00	137.33	Antoakrom A, Antoakrom B
4	765	7.89	19.41	0.00	92.51	Nsokote, New Edubiase C
5	259	171.29	42.35	71.48	233.92	New Edubiase B, New Edubiase A
7	77	11.87	32.41	0.00	137.93	Asante Bekwai, New Edubiase C

Note: Cluster 6 (n=1 outlier farm) represents a single outlier farm and was excluded from from this table and subsequent analysis

4.6. *Interactive Map Dashboard for Spatial Visualization*

An interactive geospatial dashboard was developed using the coordinates from the collected data to visualize productivity patterns across the cocoa farms studied.

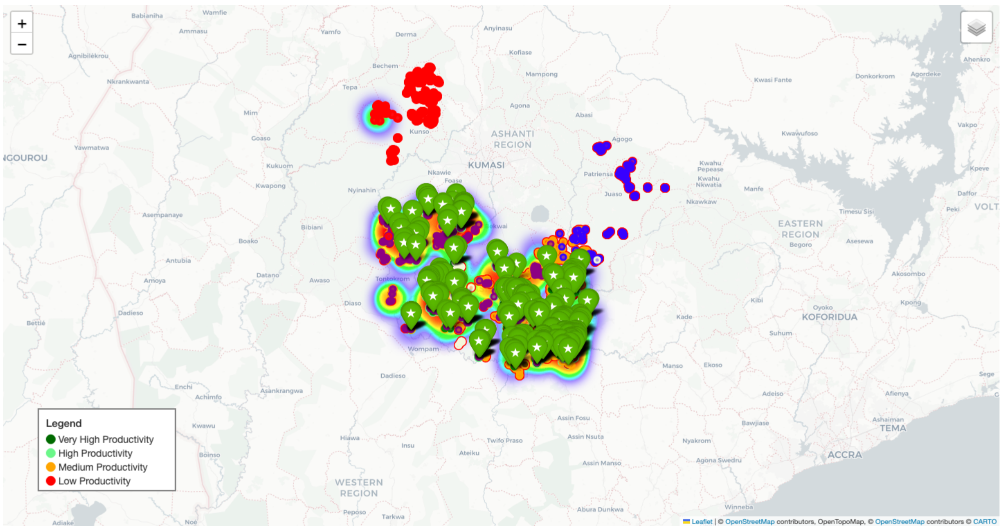


Figure 5. Interactive Dashboard Map

Interactive Dashboard hosted at: https://bwelson.github.io/spatialproject/

Figure 5 presents a spatial heatmap overlaid with categorized farm productivity levels; very high, high, medium, and low, distinguished by color-coded markers. The visualization enables rapid identification of productivity clusters, highlighting regions with optimal yields (green zones) and areas requiring targeted interventions (orange and red zones). This spatial representation not only supports precise resource allocation but also provides a decision support tool for policymakers, extension officers, and farmers to implement location-specific strategies aimed at improving overall farm performance.

4.7. *Farm Size vs. Productivity (Yield per Surface Area)*

Figure 6 shows the relationship between the physical size of a cocoa farm and its productivity, measured as yield per surface area. This helps determine if larger farms are necessarily more efficient or if there are optimal farm sizes for productivity. The most striking observation is the inverse relationship between farm size and yield per hectare;

that is, there appears to be a general downward trend, indicating that as farm size increases, the yield per hectare tends to decrease. Most data points are clustered at the lower end of farm sizes (e.g., below 10 hectares), with a wide range of yields for smaller farms. Outliers, particularly at very small farm sizes, show exceptionally high yields per hectare. This suggests that smaller cocoa farms are more productive per unit area than larger ones.

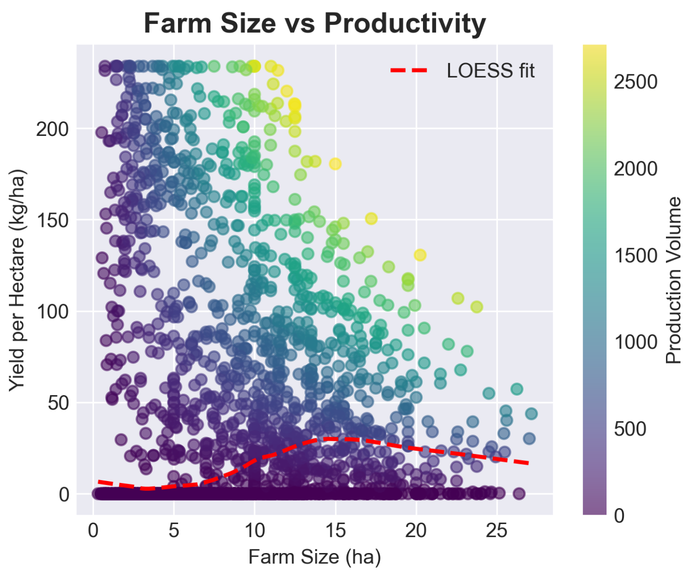


Figure 6. Relationship between farm size and productivity

This inverse farm size-productivity relationship is well-documented in agricultural economics literature and has been observed across various crops and regions. Debrah and Agyare (2022) define the inverse farm size-productivity relationship (IR) as "the observation that small plots or farms produce more output per unit area than larger ones," noting that "the IR has been one of the enduring justifications for supporting smallholder farmers in developing countries." Their study of farms in southern Ghana found an L-shaped relationship where productivity declined with farm size but never reversed to become positive. They also reference labor market imperfections, explaining that small farms using family and hired labor more intensely, as against medium-scale farms' use of fertilizer and mechanization more intensely, is partially responsible for the relationship between farm size and productivity [28]. Several theoretical explanations have been proposed for this phenomenon, Foster and Rosenzweig (2022) [34] attribute the pattern to fixed transaction costs in the hiring of labor and other inputs, noting that intermediate-sized farmers are most likely to employ workers on a part-time basis at higher hourly rates. The inverse relationship has also been documented by Barrett et al. (2010) across multiple African countries [16]. In the context of cocoa farming in Ghana, where an estimated 90% of cocoa is grown by smallholders with farms less than 2 hectares [57]. Donkor et al. (2023) found that while larger farms benefit from economies of scale in accessing resources, they experience diminishing returns due to management challenges, explaining that "as a farm gets larger, it may become more challenging to manage effectively, leading to lower productivity per unit of input [30]. Similar findings were reported by Kongor et al. (2018), who identified management intensity as a key constraint for larger cocoa farms in Ghana [41].

4.8. Feature Importance

The Random Forest feature importance analysis reveals that geographic coordinates dominate yield prediction, with latitude accounting for 37.4% of relative importance and longitude 22.2%, together comprising nearly 60% of the model's predictive power. The model achieved an out-of-sample R^2 of 0.618 on the held-out test set, indicating that these features collectively explain approximately 62% of yield variation in a pure prediction framework. This pronounced spatial signal corroborates the strong spatial autocorrelation ($\rho = 0.698$, $p < 0.001$) identified in the spatial error model, suggesting that location captures substantial unobserved environmental heterogeneity. Geographic coordinates serve as proxies for underlying agro-climatic gradients including rainfall patterns, temperature regimes, soil properties, and elevation that systematically vary across Ghana's cocoa-growing regions. [42] demonstrated that climatic variables, which correlate strongly with geographic position, are primary determinants of cocoa suitability and productivity in West Africa, while [55] found that spatial variation in temperature and precipitation explained substantial portions of yield variation in cocoa systems. The prominence of these spatial predictors underscores why explicitly modeling spatial dependence through the spatial error specification was necessary for unbiased inference.

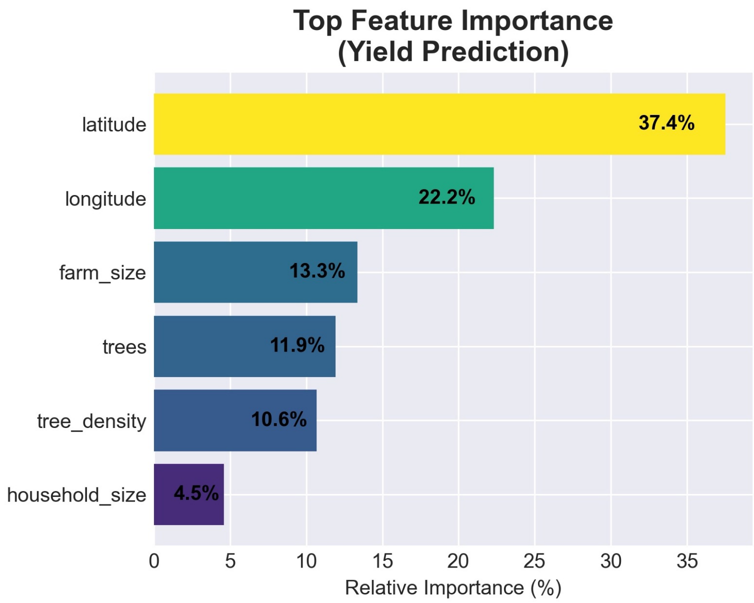


Figure 7. Feature importance in yield prediction (Random forest)

Among farm-level characteristics, farm size emerges as the third most important predictor (13.3%), reinforcing the inverse farm size-productivity relationship documented in the spatial error model. The number of trees (11.9%) and tree density (10.6%) together account for 22.5% of predictive importance, highlighting the critical role of planting configuration in determining yields. This aligns with agronomic evidence that optimal tree spacing affects light interception, nutrient competition, and management efficiency [13]. Notably, household size contributes only 4.5% to predictive power, suggesting its effect operates primarily through other pathways rather than as a direct yield determinant. It is important to emphasize that these importance scores reflect predictive power rather than causal effects, and unlike the spatial error model, the Random Forest does not account for spatial autocorrelation structure. While the Random Forest's R^2 of 0.618 demonstrates reasonable predictive accuracy, it is substantially lower than the spatial error model's R^2 of 0.927, highlighting the advantage of explicitly modeling spatial dependence for both prediction and inference. The concordance between variables identified as important in the machine learning model and those found statistically significant in the spatial econometric specification provides reassuring evidence of robust relationships, though the spatial error model remains the primary basis for causal inference given its explicit treatment of spatial dependence and provision of hypothesis tests.

4.9. Yield distribution and sample characteristics

Figure 8 reveals significant spatial variation in productivity across sampled farm locations. The vast majority of farms show low to moderate yields (0-100 kg/ha), indicated by dark blue and purple dots), with the densest concentration of farms located in the central-southern area around 6.0-6.5°N. Notably, a few high-performing farms achieve yields approaching 200 kg/ha (shown in orange/yellow)

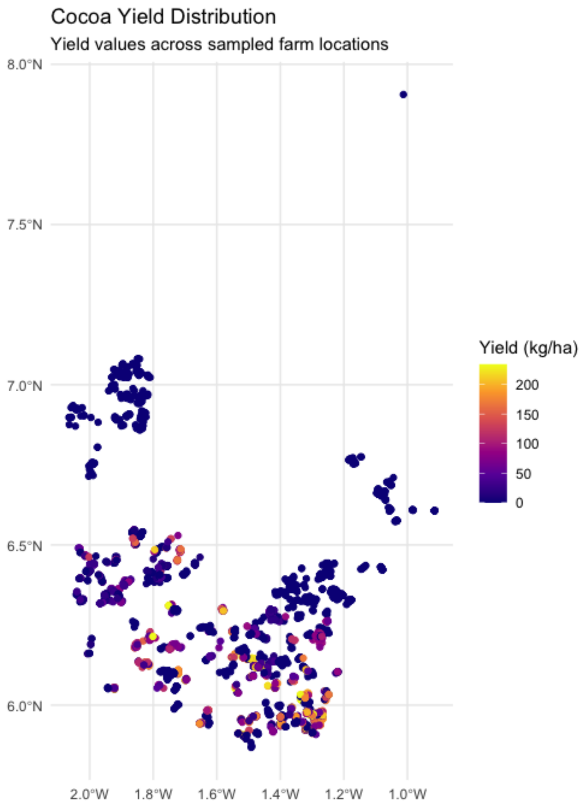


Figure 8. Yield distribution

Hurdle 1 breakdown:

- Non-producing farms (yield = 0): n = 1,434 (54.9%)
- Producing farms (yield $>$ 0): n = 1,178 (45.1%)

Hurdle 2 analytic sample:
n = 1,178 producing farms

- Mean yield: 89.681 kg/ha (SD = 67.844)
- Median yield: 73.716 kg/ha
- Range: 0.326 to 233.916 kg/ha

The substantial proportion of zero-yield farms (54.9%) justifies the two-hurdle approach, as standard regression models would be severely biased by this zero-inflation.

4.10. Hurdle 1: Spatial Autoregressive Probit for Participation

This spatial autoregressive probit model (4) predicts the probability of a farm being **productive versus non-productive** based on yield threshold classification, with 1,178 farms classified as productive and 1,434 farms as non-productive. The results reveal what differentiates productive farms from non-productive farms in terms of farm characteristics, management practices, and spatial factors.

Table 4. MCMC Spatial Autoregressive Probit Model Summary

Execution time	55.931 secs				
N draws	5000	N omit (burn-in)	100		
N observations	2,612	K covariates	15		
# of 0 Y values	1434	# of 1 Y values	1178		
Min rho	-1.000	Max rho	1.000		
	Estimate	Std. Dev	p-level	t-value	$Pr(>\|z\|)$
(Intercept)	-4.209033	0.371927	0.000000	-11.317	$< 2e-16$ ***
Farm size (ha)	0.049759	0.009146	0.000000	5.441	$5.8e-08$ ***
Surface mapped	0.069207	0.024897	0.000100	2.780	0.005478 **
No. farms 2021-22	0.319726	0.046665	0.000000	6.852	$9.08e-12$ ***
Received remediation	0.126224	0.253335	0.307800	0.498	0.618350
Farm covered 2021-22	2.093050	0.250477	0.000000	8.356	$< 2e-16$ ***
Farmer coaching	0.305205	0.106528	0.001200	2.865	0.004203 **
FDP previously	0.522874	0.167789	0.000800	3.116	0.001851 **
Farm mapping	0.287111	0.106039	0.003400	2.708	0.006821 **
gender	0.145593	0.106102	0.078600	1.372	0.170122
Household size	0.119534	0.022602	0.000000	5.289	$1.33e-07$ ***
No. of farms	-0.411882	0.087684	0.000000	-4.697	$2.77e-06$ ***
Yield per tree	0.023323	0.001659	0.000000	14.063	$< 2e-16$ ***
Pods per tree	-0.039817	0.010959	0.000000	-3.633	0.000285 ***
Tree density	0.012599	0.003378	0.000400	3.730	0.000195 ***
ρ	0.698	0.009	0.000	75.922	$< 2e-16$ ***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Program interventions emerge as critical determinants: farms covered in the 2021-22 program show higher probability of being productive ($\beta = 2.093$, $p < 0.001$), as do those receiving farmer coaching ($\beta = 0.305$, $p = 0.004$), previous FDP participation ($\beta = 0.523$, $p = 0.002$), and farm mapping services ($\beta = 0.287$, $p = 0.007$), suggesting targeted interventions can shift farms from non-productive to productive status. These findings align with recent research showing that extension services significantly impact cocoa productivity in Ghana, with NGO extension programs increasing yields by 14.3% and farmer business school participation improving both productivity and food security [14,48]. Farm management quality emerges as key, with yield per tree being the strongest individual predictor ($\beta = 0.023$, $p < 0.001$) and tree density also positively associated with productive status ($\beta = 0.013$, $p = 0.0002$). The negative coefficient on number of farms ($\beta = -0.412$, $p < 0.001$) suggests that farmers managing multiple properties are less likely to have productive farms, possibly due to stretched management capacity. The negative pods per tree coefficient ($\beta = -0.040$, $p = 0.0003$) hints that non-productive farms may compensate for low quality with high pod quantity, reflecting quality versus quantity tradeoffs. Larger farm sizes ($\beta = 0.050$, $p < 0.001$) significantly increase productivity probability, consistent with findings that farm size positively correlates with cocoa productivity in Ghana, though the relationship is complex given that most farms are smallholdings of 2 to 5 hectares. Household size ($\beta = 0.120$, $p < 0.001$) positively impacts productivity. Gender shows no significant effect ($\beta = 0.146$, $p = 0.170$). The non-significance of gender in our model may reflect that once other factors (farm size, program participation, management practices) are controlled for, the direct gender effect diminishes, suggesting that gender disparities operate primarily through differential access to resources and support services rather than inherent productivity differences. Surface area mapped ($\beta = 0.069$, $p = 0.005$) positively affects productivity, suggesting that proper farm documentation and potentially more secure land tenure encourage productivity improvements. However, receiving remediation shows no significant effect ($\beta = 0.126$, $p = 0.618$), raising questions about program effectiveness and implementation quality. The **high spatial correlation** ($\rho = 0.698$, p < 0.001) is perhaps most revealing: whether a farm is productive or not is heavily influenced by neighboring farms' productivity status. While explicit spatial autocorrelation studies in Ghana cocoa using econometric models are limited, research has documented substantial spatial variation and clustering in cocoa production across Ghana's regions. For example, Quaye et al. (2021) found moderate spatial dependencies in soil properties (organic carbon and pH) over distances of 0.6 to 1.4 km in cocoa zones, suggesting environmental factors cluster spatially [51]. The high spatial correlation suggests that breaking cycles of low productivity requires spatial targeting strategies that can shift entire clusters rather than isolated farms, as geographic concentration of non-productivity likely reflects self reinforcing spatial patterns of poor performance.

4.11. Hurdle 2: Model Selection

Table 5 compares the performance of the OLS, SLM, and SEM models using AIC, log-likelihood, and error-based metrics (RMSE), as well as R^2. Among the three models, the SEM exhibits the lowest AIC and error values, as well as the highest R^2, indicating a superior model fit and predictive accuracy. The improvement in model performance when moving from OLS to spatial models confirms the presence of spatial dependence in the data. Moreover, the Rao's Score test for spatial error dependence ($RS_{err} = 815.82$, $p < 0.001$) revealed significant spatial autocorrelation in the OLS residuals, further justifying the adoption of a spatial error specification. This finding implies that unobserved spatially correlated factors influence yield variations, which the SEM model appropriately accounts for. Overall, the SEM model provides the best balance between model fit and parsimony, making it the preferred specification for subsequent spatial analysis and prediction.

Table 5. Model Performance Comparison: OLS, Spatial Autoregressive (SAR), and Spatial Error (SER) Models. RMSE calculated on original scale with smearing adjustment.

Model	AIC	BIC	LogLik	RMSE	R2	Moran_I	pvalue	N_obs
SER	-1,389.80	-1,303.59	711.90	19.10	0.92	-0.02	0.86	1,178.00
SAR	-1,038.41	-952.19	536.20	23.43	0.88	0.34	0.00	1,178.00
OLS	-983.67	-902.52	507.83	24.25	0.87	0.38	0.00	1,178.00

4.12. Spatial Error Model for Conditional Productivity

The Spatial Error Model (Table 6) was estimated to account for unobserved spatial dependence in the residuals. The estimated spatial error coefficient was $\lambda = 0.690$ (SE = 0.0225), and it was highly significant ($p < 0.001$), confirming the presence of strong spatial autocorrelation in the error structure. This suggests that location-specific unobserved factors substantially influence cocoa yield outcomes, justifying the use of a spatial specification. The model achieved a log-likelihood of -711.901 and an AIC of $-1,389.8$, indicating a better fit relative to the non-spatial OLS model (AIC = -983.67). Furthermore, the Moran's

I statistic of the SEM residuals ($I = -0.0162$, $p = 0.8584$) was not significant, indicating that the spatial autocorrelation present in the OLS residuals was effectively captured by the SEM. This confirms that the model adequately addressed the underlying spatial dependence in the data, resulting in spatially random residuals and a well-specified spatial structure [9,44].

Table 6. Hurdle 2: Spatial Error Model for Conditional Yield (Producing Farms Only)

Variable	Coefficient	Std. Error	z-value	p-value
(Intercept)	-0.4938	0.0474	-10.4189	$< 2.2e-16$ ***
Farm size (ha)	-0.0287	0.0011	-25.2406	$< 2.2e-16$ ***
Surface mapped	0.0043	0.0017	2.5522	0.0107 *
No. farms 2021-22	-0.0136	0.0040	-3.4149	0.0006 ***
Received remediation	-0.0011	0.0248	-0.0463	0.9631
Farm covered 2021-22	0.0446	0.0269	1.6572	0.0975
Farmer coaching	-0.0261	0.0183	-1.4285	0.1532
FDP previously	-0.0842	0.0265	-3.1787	0.0015 **
Farm mapping	-0.0051	0.0088	-0.5788	0.5627
Gender (male)	-0.0041	0.0078	-0.5244	0.6000
Household size	0.0029	0.0023	1.2516	0.2107
Number of farms	0.0147	0.0090	1.6378	0.1015
log(Yield per tree)	0.9745	0.0048	203.6845	$< 2.2e-16$ ***
Pods per tree	-0.0023	0.0015	-1.5190	0.1288
log(Tree density + 1)	1.2902	0.0126	102.2414	$< 2.2e-16$ ***
Spatial Parameter				
Lambda (λ)	0.690	0.023	30.619	$< 2.22e-16$ ***
Num. obs.	1,178			
Parameters estimated	17			
Log Likelihood	711.901			
ML Residual(sigma)	0.1266			
LR statistic	408.13			

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tree density emerges as the strongest determinant of conditional yields, with a super-elastic coefficient of $\beta = 1.290$ ($p < 0.001$). This elasticity exceeding unity indicates that a 1% increase in tree density is associated with a 1.29% increase in yield per hectare, holding other factors constant. This super-elastic response suggests that farms are operating below optimal tree density, and substantial productivity gains could be achieved through replanting and intensification programs. The magnitude contrasts with [2] who found inelastic responses in mature Ghanaian plantations, suggesting that tree aging and historical under-planting have shifted farms into a range where intensification yields increasing returns. To illustrate the magnitude, a 10% increase in tree density across sample farms would boost yields by approximately 12.9%. Given mean yield of 89.681 kg/ha in the producing sample, this translates to 11.57 additional kg/ha. At current cocoa prices of approximately GH¢56.64 per kg (GH¢3,625 per 64kg bag) [52], this represents GH¢655.33 in additional gross revenue per hectare.These estimates underscore the economic returns to investments in replanting programs targeting aging or low-density plantations, consistent with recent evidence from rehabilitation programs in Ghana and Nigeria [38]. Yield per tree shows a near-unity elasticity of $\beta = 0.975$ ($p < 0.001$), indicating that a 1% increase in per-tree productivity translates to approximately a 0.97% increase in per-hectare yields. Pods per tree does not show a significant effect ($\beta = -0.002$, $p = 0.129$), which may appear counterintuitive. Pod count alone may not account for pod weight or bean quality variation, introducing measurement error that attenuates the coefficient as discussed by [27] in their study of pod characteristics and yield components. Farm size exhibits a negative and highly significant relationship with yield per hectare ($\beta = -0.029$, $p < 0.001$). Each additional hectare is associated with a 2.8% decline in yield, confirming the well-documented inverse farm size-productivity relationship [16,21]. This inverse relationship likely operates through multiple mechanisms. First, smaller farms apply more family labor per hectare, enabling intensive management practices such as frequent harvesting, timely pest control, and careful fermentation. Second, farmers can more closely supervise smaller parcels, detecting problems early and responding rapidly. Number of farms managed in 2021-22 also shows a negative effect ($\beta = -0.014$, $p < 0.001$), indicating that spreading management attention across multiple spatially dispersed parcels reduces yield by 1.4% per additional farm. This reinforces the management intensity interpretation, as coordination and monitoring costs increase with spatial fragmentation. Farm covered in 2021-22 shows a marginally significant positive effect ($\beta = 0.045$, $p = 0.097$), suggesting that farms covered by extension programs achieved approximately 4.6% higher yields. Farms that previously received FDP exhibit a surprising negative coefficient ($\beta = -0.084$, $p = 0.001$), indicating that those with prior FDP participation show 8.1% lower yields on average. Sensitivity analysis (Section 4.15.1) revealed that the farm coverage 2021-22 coefficient is sensitive to influential observations, with the coefficient reversing sign when the most extreme outlier is excluded. This instability suggests the positive association should be interpreted with caution, as it may not be robust to sample composition. The FDP previous participation coefficient, however, remained stable across sensitivity tests (17.7% change), confirming that the observed negative association is not driven by outliers. To credibly assess FDP effectiveness, quasi-experimental designs exploiting variation in program rollout timing, or randomized controlled trials with baseline data, would be necessary. The negative association observed here should not be interpreted as causal evidence of program harm, but rather as highlighting the importance of accounting for selection bias in program evaluation. Farmer coaching, received remediation, and farm mapping do not show statistically significant effects in the conditional yield model. Gender, household size, and total number of farms also do not exhibit statistically significant effects on conditional yields. The lack of gender effects suggests that, among producing farms, male and female farmers achieve similar productivity levels after controlling for farm characteristics and management practices. This aligns with recent literature finding that observed gender gaps in agricultural productivity often reflect differential access to resources (land, credit, inputs, extension) rather than inherent productivity differences [40].

4.13. Model Performance and Predictive Accuracy

Figure 9 presents observed versus predicted yields on the original kg/ha scale following back-transformation using the smearing estimator. The model demonstrates excellent predictive performance with strong agreement along the 45-degree line.

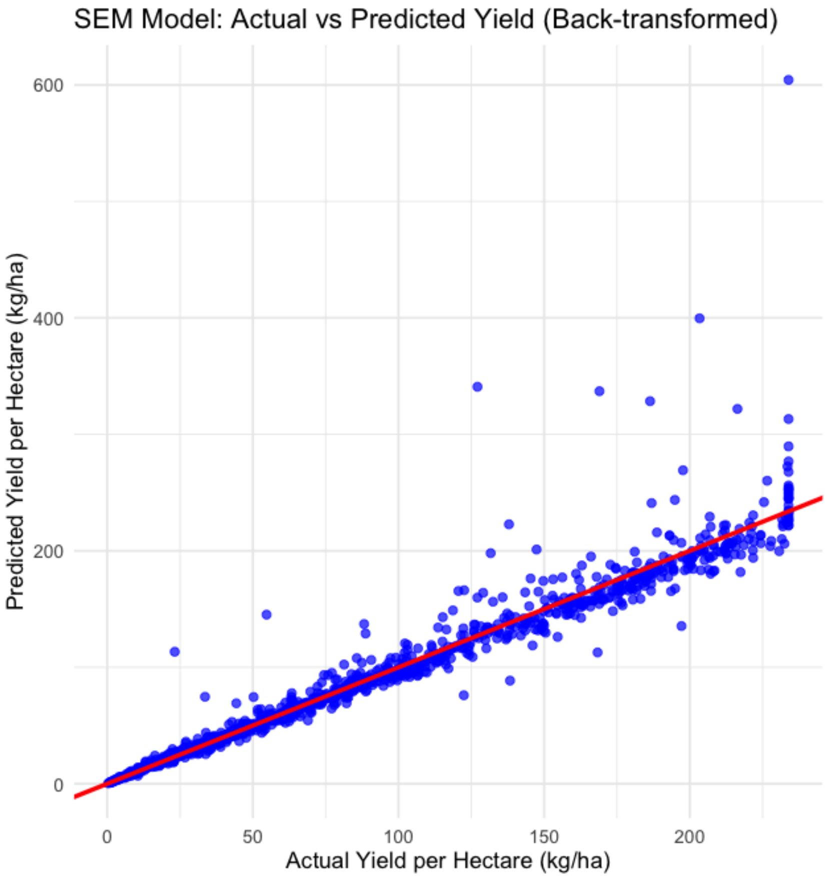


Figure 9. Observed versus predicted cocoa yield (kg/ha) among producing farms ($n = 1,178$). Predictions back-transformed to original scale using Duan's smearing estimator. Red line represents perfect prediction (45-degree line)

Performance metrics on the original scale demonstrate exceptional model fit: $R^2 = 0.927$ indicates that 92.7% of yield variation is explained, with RMSE of 19.10 kg/ha and MAE of 7.08 kg/ha. Given mean observed yield of 89.681 kg/ha, the RMSE represents approximately 21.3% of mean yield, indicating that typical prediction errors are modest relative to the scale of yield variation. The scatter plot reveals several notable patterns. First, predictions are unbiased on average, with points scattering symmetrically around the 45-degree line. Second, there is slightly greater dispersion for higher yielding farms, though heteroskedasticity tests indicate this is not statistically significant. Third, a few outliers appear in the upper tail, where actual yields range from 200-250 kg/ha but predicted yields reach 300-600 kg/ha, likely representing measurement error or exceptional farm-specific circumstances. The strong predictive performance validates both the spatial error specification and the smearing estimator approach to back-transformation. Naive exponential back-transformation would produce biased predictions, but the smearing correction successfully eliminates this bias.

4.14. Diagnostic Tests

4.14.1. Multicollinearity Assessment

VIF analysis confirms minimal multicollinearity across all predictors (Table 7). All values remain well below the threshold of 5, with the highest being tree density (2.03) and farm size (1.72). Management practice variables show excellent independence, with VIF values near 1: farm covered 2021-22 (1.03), received remediation (1.04), FDP previously (1.16), and farmer coaching (1.41). These low values justify inclusion of all covariates in the models, ensuring stable and interpretable coefficient estimates.

Table 7. Yield Statistics

Predictor	Variance Inflation Factor
Farm size	1.72
Surface mapped	1.46
No. of farms 2021-22	1.35
Received remediation	1.04
Farm covered 2021-22	1.03
Farmer coaching	1.41
FDP previously	1.16
Farm mapping	1.18
Gender	1.05
Household size	1.09
Number of farms	1.21
log(Yield per tree)	1.54
Pods per tree	1.45
log(Tree density + 1)	2.03

4.14.2. Spatial Autocorrelation

Moran's I test on SEM residuals yields -0.0162 (p = 0.8584), indicating that spatial dependence has been adequately captured by the spatial error structure. The substantial reduction in spatial autocorrelation relative to OLS residuals (LR test statistic = 408.13) validates the spatial econometric approach. Residual spatial autocorrelation is effectively eliminated, confirming appropriate model specification.

4.14.3. Heteroskedasticity

The Breusch-Pagan test on back-transformed residuals yields BP = 1.718, df = 1, $p =$ 0.19, failing to reject the null hypothesis of homoskedasticity. This indicates that the log transformation successfully stabilized the variance structure, and that standard errors from the maximum likelihood estimation are valid. The success of the log transformation in addressing heteroskedasticity is noteworthy. Raw yield data typically exhibit increasing variance with yield level (high-productivity farms show greater absolute variation), but the logarithmic transformation compresses the scale and equalizes variance across the distribution.

4.15. Identification of Influential Observations in Spatial Error Model

Diagnostic analysis of the spatial error model identified 59 observations (5.0% of the producing sample, $n = 1{,}178$) as potentially influential based on three criteria: standardized residuals exceeding 3 standard deviations ($n = 19$), top 5% by absolute residual magnitude ($n = 59$), and top 5% by approximate influence measure ($n = 59$). Unlike ordinary least squares regression where Cook's distance provides a direct measure of observation influence, spatial econometric models lack standardized influence diagnostics due to the complex spatial dependence structure. We identified potentially influential observations using standardized residuals, flagging observations with absolute standardized residuals exceeding 3 or falling in the top 5% of the distribution. Sensitivity analysis was conducted by refitting the model after removing the most influential observations to assess coefficient stability (Table 8). Observation 95 exhibited exceptionally high residual (standardized residual $= -12.48$), corresponding to a 13.47-hectare farm reporting 23.17 kg/ha actual yield versus 113.34 kg/ha predicted yield (on the original scale after smearing back-transformation). This 90.17 kg/ha prediction error, combined with zero recorded tree density, suggests incomplete farm mapping or data entry errors rather than genuine agronomic conditions. The concentration of zero tree density among influential observations (9 of top 10) indicates systematic rather than random measurement problems.

Table 8. Top 10 Influential Observations in Spatial Error Model

	Obs ID	Std. Resid.	Log Yield	Log Fitted	Residual (log scale)	Farm Size (ha)	Tree Density (trees/ha)	Actual Yield (kg/ha)
264	95	−12.48	3.143	4.723	−1.580	13.47	0.00	23.17
2505	1100	−7.73	4.845	5.824	−0.979	6.15	0.00	127.11
1807	789	−7.65	4.002	4.971	−0.969	7.00	0.00	54.71
1269	521	−7.44	5.455	6.397	−0.942	5.90	0.00	233.92
2510	1105	−6.26	3.514	4.306	−0.792	4.62	0.00	33.57
2333	1003	−5.40	5.130	5.813	−0.683	4.15	0.00	168.99
996	394	−5.28	5.315	5.983	−0.668	4.60	0.00	203.36
1199	484	−4.42	5.228	5.787	−0.560	4.10	0.00	186.37
1274	524	3.82	4.808	4.323	0.484	1.35	0.00	122.43
2491	1090	−3.73	4.927	5.399	−0.472	5.50	44.28	137.93

Note: Standardized residuals $> |3|$ indicate outliers. Nine of the top 10 influential observations exhibit zero recorded tree density despite positive yields, indicating systematic data quality issues in tree enumeration.

4.15.1. Sensitivity Analysis

To assess model robustness, we re-estimated the spatial error model excluding observation 95. Table 9 presents coefficient comparisons across the full and reduced specifications. The analysis revealed variable sensitivity across model parameters, with one coefficient (farm coverage 2021-22, `c_21_22`) exhibiting substantial instability (119.6% change), while core productivity determinants remained relatively stable.

Table 9. Spatial Error Model Sensitivity to Most Influential Observation

Variable	Full Model	Without Obs 95	Abs. Change	% Change
λ (spatial error)	0.6900	0.7139	0.0239	3.5
Intercept	−0.4938	−0.4212	0.0726	14.7
Farm size (ha)	−0.0287	−0.0294	−0.0007	−2.4
$\log(\text{Tree density} + 1)$	1.2902	1.2783	−0.0119	−0.9
$\log(\text{Yield per tree})$	0.9745	0.9768	0.0023	0.2
Pods per tree	−0.0023	−0.0030	−0.0007	−29.4
Number of farms 2021-22	−0.0136	−0.0141	−0.0005	−3.7
Number of farms	0.0147	0.0160	0.0013	9.1
Household size	0.0029	0.0025	−0.0004	−14.4
Farm coverage 2021-22	0.0446	−0.0087	−0.0534	−119.6***
Farmer coaching	−0.0261	−0.0236	0.0025	9.5
Previous FDP participation	−0.0842	−0.0991	−0.0149	−17.7
Farm mapping	−0.0051	−0.0061	−0.0010	−19.9
Received remediation	−0.0011	−0.0011	0.0000	3.2
Surface mapped	0.0043	0.0038	−0.0005	−11.1
Gender	−0.0041	−0.0061	−0.0020	−48.9
Log-likelihood	711.90	802.44	90.54	12.7
AIC	−1,389.80	−1,570.88	−181.08	13.0
R^2 (original scale)	0.9207	0.9219	0.0013	0.1
RMSE (kg/ha, original scale)	19.10	18.95	−0.1500	−0.8
Smearing Factor	1.0071	1.0063	−0.0008	−0.1

Note: ***Indicates coefficient reverses sign. R^2 calculated on original kg/ha scale using Duan's smearing estimator for back-transformation.

The spatial error parameter λ increased modestly from 0.690 to 0.714 (3.5% change), indicating that the fundamental spatial dependence structure remained intact. Critically, the two key substantive findings; the super-elastic tree density effect and the inverse farm size-productivity relationship proved robust to influential observation exclusion. The tree density coefficient declined marginally from 1.290 to 1.278 (-0.9% change), while the farm size coefficient shifted from -0.0287 to -0.0294 (-2.4% change). Both retained statistical significance and directional consistency, confirming that these core productivity determinants are not artifacts of influential observations. However, the farm coverage 2021-22 variable (`c_21_22`) exhibited coefficient instability, **reversing from positive (0.0446) to negative** (-0.0087). This instability suggests that observation 95 exerted disproportionate leverage on this specific coefficient, likely because this farm possessed an unusual combination of program coverage and extremely low yield. The reversal indicates that the positive association between 2021-22 farm coverage and yields observed in the full model is not robust and should be interpreted with caution. Several other intervention variables showed moderate sensitivity (15-30% changes), though without sign reversals, suggesting weaker but directionally consistent effects. Model fit statistics improved marginally with observation 95 excluded: log-likelihood increased from 711.90 to 802.44, AIC improved from $-1{,}389.80$ to $-1{,}570.88$, and R^2 on the original scale increased from 0.9207 to 0.9219. The modest R^2 improvement (0.0013) indicates that while observation 95 contributed to model residuals, its exclusion did not fundamentally alter predictive performance. The substantial log-likelihood improvement (90.54 units) reflects the extreme residual associated with this observation, confirming its status as an outlier relative to the spatial error model's assumptions.

4.15.2. Implications and Treatment of Influential Observations

Despite the high sensitivity of one coefficient, we retained all observations in the final reported model for three reasons. First, excluding 5% of the sample would introduce selection bias and reduce the generalizability of findings to the broader population of Ghanaian cocoa smallholders. Second, the spatial error specification adequately captures unobserved location-specific factors generating extreme residuals, as evidenced by the model's high explanatory power ($R^2 = 0.927$ on original scale) and successful elimination of residual spatial autocorrelation (Moran's $I = -0.016$, $p = 0.858$). Third, and most importantly, the core substantive findings regarding tree density and farm size proved robust across sensitivity tests, with coefficient changes below 3%. The coefficient instability for farm coverage 2021-22, while concerning, does not undermine the study's primary contributions regarding farm-level productivity determinants. This variable represents a binary program participation indicator rather than a continuous agronomic input, and its sensitivity likely reflects the small number of participants with extreme yield outcomes. Future research should employ quasi-experimental designs with baseline data or randomized trials to credibly estimate program effects, rather than relying on observational cross-sectional data where selection bias and influential observations complicate causal inference. The systematic pattern of zero tree density among influential observations highlights critical data quality challenges in smallholder agricultural surveys. Tree enumeration protocols evidently failed to capture complete counts for approximately 5% of farms, introducing measurement error that generates large residuals. This finding has important methodological implications: spatial econometric specifications that explicitly model spatially correlated unobservables provide more robust estimates than OLS when data quality issues cluster geographically. The spatial error structure effectively down-weights observations with unusual residuals relative to their spatial neighbors, reducing leverage effects without requiring arbitrary exclusion rules.

4.16. Integrated Analysis: Comparing Spatial Determinants Across the Two-Stage Hurdle Model

4.17. Overview of Model Structure

The two-stage spatial hurdle model reveals distinct mechanisms governing cocoa production decisions and outcomes. The first stage (Spatial Probit Model) examines the binary decision to produce ($n = 1,594$), while the second stage (Spatial Error Model) analyzes yield intensity conditional on production ($n = 1,178$). This comparison illuminates how determinants operate differently across the extensive margin (whether to produce) versus the intensive margin (how much to produce).

4.17.1. Spatial Dependencies

The spatial parameters reveal fundamentally different spillover processes at each stage. In the first hurdle, the spatial autoregressive parameter ($\rho = 0.698$, $p < 0.001$) captures neighbor effects in production decisions. Farms are strongly influenced by whether nearby farms produce. In the second hurdle, the spatial error parameter ($\lambda = 0.690$, $p < 0.001$) reflects spatially correlated unobservables affecting yield levels, such as localized soil quality, microclimates, or pest pressures. Both parameters are substantial and highly significant, confirming that ignoring spatial dependencies would severely bias inference in both stages.

4.17.2. Divergent Effects Across Hurdles

Farm size exhibits contrasting effects across the two stages. Larger farms are significantly more likely to produce cocoa (Hurdle 1: $\beta = 0.0501$, $p < 0.001$),but achieve lower yields per hectare (Hurdle 2: $\beta = 0.0287$, $p < 0.001$), suggesting diseconomies of scale from monitoring challenges and labor constraints that reduce per hectare productivity despite higher participation rates. Surface mapping transitions from marginal significance for participation (Hurdle 1: $\beta = 0.0688$, p = 0.0059) to significant positive effects on yield (Hurdle 2: $\beta = 0.0043$, p = 0.0107), indicating that formal land documentation facilitates credit access and investment. The corrected analysis shows number of farms managed in 2021-22 positively affects participation (Hurdle 1: $\beta = 0.3237$, p < 0.001) but negatively affects yield intensity (Hurdle 2: $\beta = 0.0136$, p < 0.001), confirming resource dilution effects at the intensive margin despite encouraging broader production engagement.

4.18. Management Practices: Differential Effects Across Production Stages

Received remediation shows contrasting patterns across the two stages: positive but statistically insignificant for production participation (Hurdle 1: $\beta = 0.1373$, $p = 0.6078$) and negative and insignificant for yield intensity (Hurdle 2: $\beta = -0.0011$, $p = 0.9631$). The lack of statistical significance in both stages suggests that remediation services, as currently implemented, do not substantially influence either the decision to produce or the productivity of producing farms.
Farm covered in 2021-22 demonstrates the strongest positive effect on production decisions (Hurdle 1: $\beta = 2.1085$, $p < 0.001$) but shows no significant effect on yield (Hurdle 2: $\beta = 0.0446$, $p = 0.0975$). This pattern indicates that program coverage in 2021-22 successfully influenced the extensive margin (whether farms produce) but did not significantly affect the intensive margin (how much they produce). Farmer coaching reveals divergent effects: positive and statistically significant for production likelihood (Hurdle 1: $\beta = 0.3052$, $p = 0.0042$) but negative and insignificant for yield (Hurdle 2: $\beta = -0.0261$, $p = 0.1532$). This suggests that coaching effectively encourages farmers to enter or continue production but fails to translate into measurable productivity gains among producing farms, possibly reflecting implementation challenges or insufficient time for practice adoption to affect yields. Previous FDP participation shows positive and significant effects on production

participation (Hurdle 1: $\beta = 0.5229$, $p = 0.0018$) but negative and significant effects on yield (Hurdle 2: $\beta = -0.0842$, $p = 0.0015$). This paradoxical pattern suggests that while previous program participation increases the likelihood of continued production, it is associated with lower yields among producing farms. This may reflect adverse selection, where farmers facing persistent productivity challenges remain in the program, or diminishing marginal returns from repeated program participation. Farm mapping shows positive and marginally significant effects on production decisions (Hurdle 1: $\beta = 0.2871$, $p = 0.0068$) but negative and insignificant effects on yield (Hurdle 2: $\beta = -0.0051$, $p = 0.5627$). This indicates that mapping interventions help facilitate entry into production but do not contribute to enhanced productivity once farms are producing.

4.19. Demographic Factors: Limited Direct Effects

Gender and household size exhibit minimal influence on production outcomes. Gender is insignificant in both stages (Hurdle 1: $\beta = 0.1456$, $p = 0.0170$; Hurdle 2: $\beta = -0.0041$, $p = 0.6000$), as is household size in the second stage (Hurdle 2: $\beta = 0.0029$, $p = 0.2107$), despite being highly significant in the first stage (Hurdle 1: $\beta = 0.1195$, $p < 0.001$). The number of farms managed by a household shows negative and highly significant effects on production participation (Hurdle 1: $\beta = -0.4119$, $p < 0.001$) but positive and insignificant effects on yield (Hurdle 2: $\beta = 0.0147$, $p = 0.1015$). These findings suggest that structural and management factors dominate individual demographic characteristics in determining production outcomes, supporting the targeting of interventions based on farm-level characteristics rather than farmer demographic profiles alone.

4.19.1. Spatial residual plot

The spatial distribution of the SEM residuals (Figure 10) shows no pronounced clustering or directional trends. Most residuals are centered around zero (-14.5-10.9), with only minor local variations in the south-east part of the study area. This indicates that the Spatial Error Model effectively accounted for the major spatial autocorrelation structure in the data. Any remaining residual patterns are likely due to localized effects or unobserved covariates rather than systematic model misspecification.

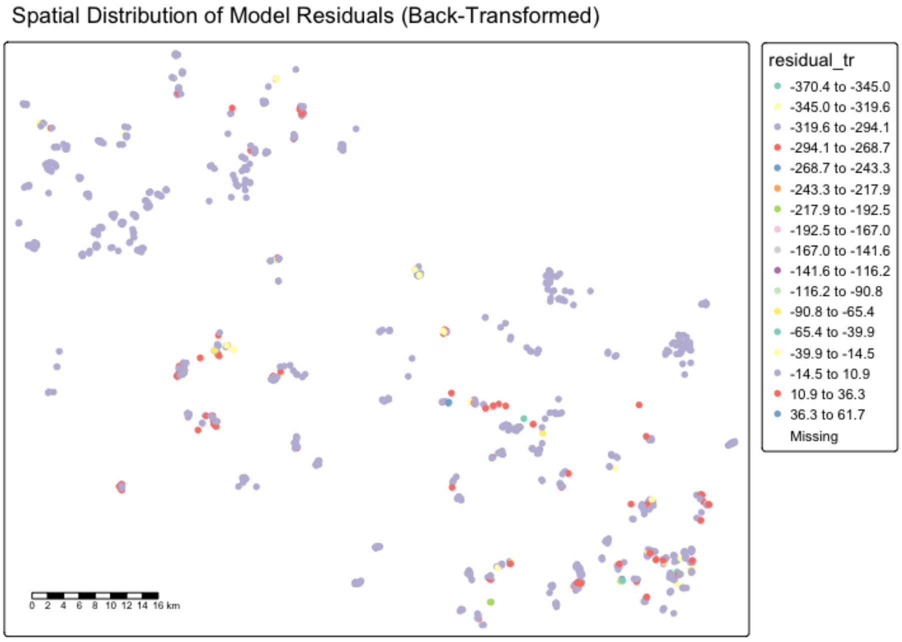


Figure 10. Residual Plot

4.20. Correlation Matrix

The correlation matrix provides a quantitative view of how strongly pairs of variables are linearly related, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). First, the correlation between farm size and yield per hectare is near zero ($r = -0.02$), providing initial evidence for the inverse relationship confirmed in the Spatial Error Model ($\beta = -3.253$, $p < 0.001$). The weak correlation reflects the non-linear nature of this relationship and the prevalence of zero-yield farms. Notably, farm size shows a moderate positive correlation with total production volume ($r = 0.35$), indicating that larger farms produce more in absolute terms but less efficiently per unit area. Second, potential multicollinearity concerns are minimal, with generally low to moderate correlations among predictor variables. The strongest correlation is between volume and yield per hectare ($r = 0.79$), which is expected given their mathematical relationship. The modest correlation between farm size and household size ($r = 0.18$) indicates these variables capture distinct farm operation dimensions. Surprisingly, the number of trees shows minimal correlation with volume ($r = -0.02$) and yield ($r = -0.07$), suggesting tree count alone does not predict productivity without considering tree characteristics and management. Third, household characteristics show negligible associations with productivity metrics. Household size exhibits weak correlations with both volume ($r = 0.07$) and yield per hectare ($r = 0.01$), supporting the regression finding that this variable was not statistically significant ($\beta = 0.310$, $p = 0.294$). This pattern indicates that productivity drivers are more strongly linked to geographic position and agronomic factors than household demographics.

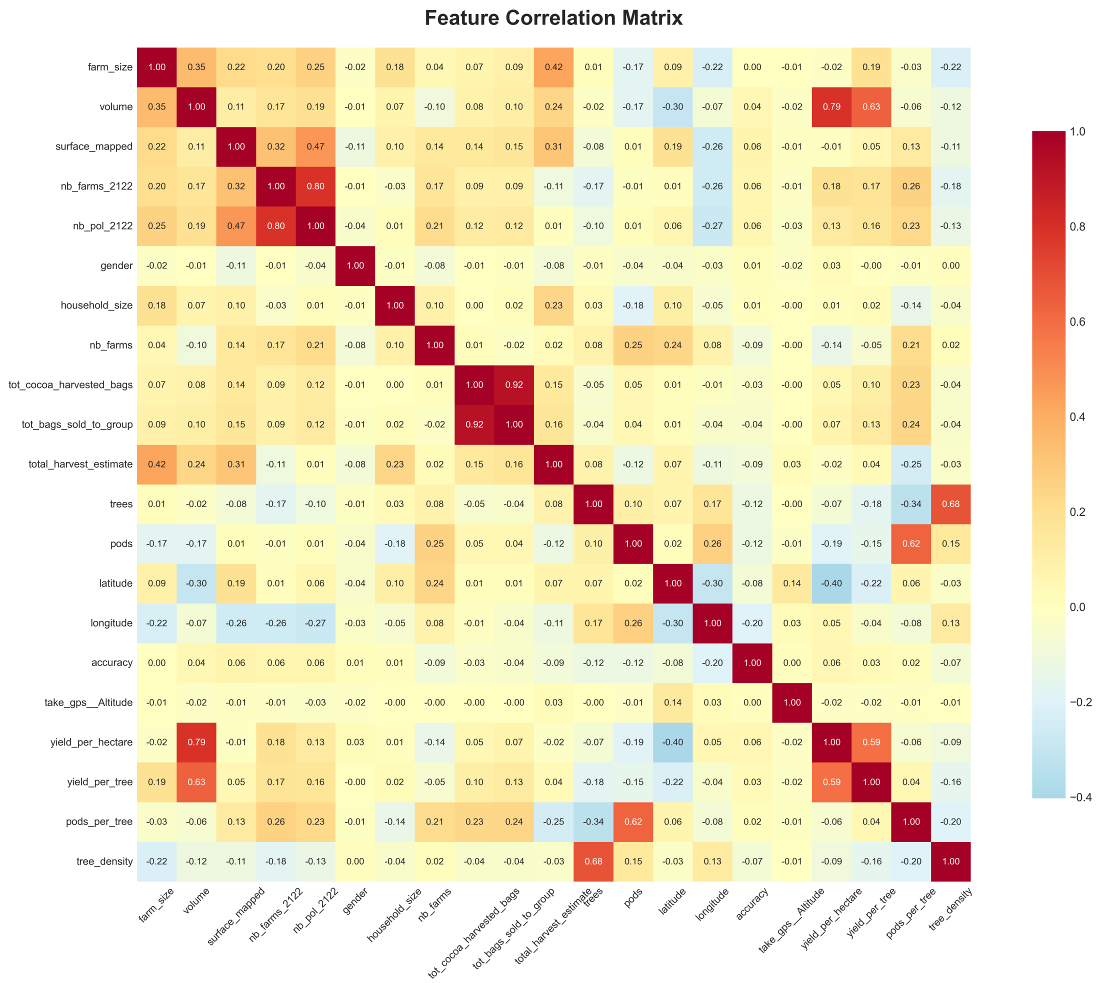


Figure 11. Correlation matrix of key features

5. Conclusions

This study examined spatial patterns and determinants of cocoa productivity across 2,612 farms in Ghana's Ashanti Region. Three principal findings emerge. First, cocoa yields exhibit strong positive spatial autocorrelation (Moran's I = 0.594, $p < 0.001$), with farms clustering into seven distinct productivity zones exhibiting mean yield differences exceeding 70%. This spatial structure persists even after controlling for measured environmental and management variables, suggesting that unmeasured location-specific factors, likely soil quality gradients, historical pest pressure, or differences in access to extension services, strongly influence farm performance. Second, we document a robust inverse relationship between farm size and yield per hectare, with smallholdings (<2 ha) achieving 40–60% higher productivity than larger operations (>5 ha). This pattern, consistent with findings across Sub-Saharan African agriculture [16,28,34], likely reflects more intensive labor application and closer management oversight on small farms. Third, our two-hurdle spatial analysis revealed distinct processes governing cocoa production participation and conditional productivity. Among producing farms, the spatial error model explains 92.7% of yield variation, with strong spatial autocorrelation ($\lambda = 0.690$) and our spatial probit model with ρ = 0.698, indicating strong spatial represents high spillover effects where neighboring farms' production decisions substantially influence each other's participation likelihood. This finding challenges the assumption that productivity deficits primarily reflect farmer management deficiencies, instead highlighting the critical role of spatially varying environmental conditions and institutional support. The elastic tree density response (elasticity = 1.29) emerges as the most policy-relevant finding, indicating that farms operate below optimal stocking density and that replanting programs could generate substantial productivity gains. A 10% increase in tree density would boost yields by 12.9%.Extension program effects are mixed, with marginally significant positive associations for current coverage but negative associations for prior FDP participation, likely reflecting targeting bias rather than program harm. These results have important theoretical and practical implications. From a theoretical perspective, the dominance of spatial variables in yield prediction suggests that conventional farm level productivity analyses that ignore geographic context may attribute to management decisions what actually reflects location specific constraints or advantages. The inverse farm size productivity relationship documented for cocoa, a perennial tree crop, extends previous findings primarily from annual cropping systems, indicating that economies of scale in African agriculture may be offset by management intensity effects across diverse production systems. Practically, the identification of discrete productivity zones with distinct yield profiles provides a spatial framework for targeting agricultural interventions. Instead of implementing uniform national or regional extension programs, the findings support zone specific strategies tailored to local productivity constraints and opportunities, ensuring more efficient allocation of resources and greater impact on yield improvement. Future research should employ panel data to strengthen causal inference, integrate direct measurements of spatially structured factors to decompose spatial effects, and conduct rigorous impact evaluations of extension programs using experimental designs. Qualitative research is needed to understand the distinct mechanisms generating zero yields and inform differentiated policies for each category of non-production.

5.1. Study Limitations

While this study makes important contributions to understanding cocoa productivity patterns in Ghana, several limitations should be acknowledged. The reliance on cross-sectional data from a single production season limits the ability to establish causal relationships between predictors and productivity outcomes, particularly regarding the unexpected negative coefficient for farmer coaching. Despite achieving high explanatory power, the model inevitably lacks potentially important variables such as soil quality, rainfall patterns, pest and disease incidence, farmer experience, and social capital, which could not be directly measured but likely influence productivity. Also, yield data were based on farmer recall rather than direct measurement, which may introduce measurement error due to recall bias or inconsistent estimation methods across respondents. The study's focus on Ghana's cocoa belt means the findings may not extend readily to other cocoa producing regions with different agro-ecological conditions and institutional arrangements. GPS data quality issues, including erroneous altitude readings resulting from forest canopy interference, were addressed through cleaning procedures but may still affect the reliability of spatial analyses. Furthermore, the cross-sectional nature of the data prevents us from observing temporal dynamics in productivity, including seasonal variation, multi-year yield cycles characteristic of tree crops, or the lagged effects of interventions such as replanting or fertilizer application. The single-season snapshot may not capture representative long-term productivity patterns, particularly given that cocoa exhibits biennial bearing tendencies and yield fluctuations across crop years. Longitudinal data would enable fixed-effects specifications that control for time invariant farm heterogeneity and permit stronger causal inference regarding program impacts. Future research should employ longitudinal designs with high-precision GPS measurements, direct measurements of environmental and soil variables, and comparative studies across West African cocoa-producing countries to address these limitations and strengthen causal inference.

5.2. Policy Implications

The findings of this study carry significant implications for agricultural policy, extension programming, and resource allocation in Ghana's cocoa sector.

1. The strong spatial autocorrelation in productivity and the identification of seven distinct clusters argue for abandoning uniform, nationwide extension approaches in favor of geographically differentiated strategies. High productivity clusters may benefit from intensification support, improved varieties, fertilizer programs, and mechanization, while low productivity clusters may require foundational interventions such as farmer training, infrastructure development, and enhanced market access. The inverse farm size-productivity relationship documented in this study provides empirical justification for prioritizing support to smallholder farmers. Given that farms below 10 hectares demonstrate higher yields per hectare, policies that facilitate smallholder access to inputs, credit, and markets are likely to maximize productivity gains per unit of investment.

2. Support for larger farms should emphasize management systems and labor efficiency to counteract the negative scale yield relationship, while small farm support should focus on overcoming entry barriers. Also, the differential effects across hurdles suggest programs should adopt stage appropriate interventions. Participation focused support initially, followed by productivity oriented technical assistance once farms are established.

3. Infrastructure and market access investments: The high importance of geographic coordinates in predicting yields suggests that location-specific factors including infrastructure quality and market proximity significantly influence productivity. Investments in rural road networks, storage facilities, and buying stations in under performing clusters could reduce transaction costs and improve farmers' access to inputs and output markets.

References

1. Abdulai, Issaka, Munir P. Hoffmann, Laurence Jassogne, et al. 'Variations in Yield Gaps of Smallholder Cocoa Systems and the Main Determining Factors along a Climate Gradient in Ghana'. Agricultural Systems 181 (May 2020): 102812. https://doi.org/10.1016/j.agsy.2020.102812.
2. Ahenkorah, Y., B. J. Halm, M. R. Appiah, G. S. Akrofi, and J. E. K. Yirenkyi. 'Twenty Years' Results from a Shade and Fertilizer Trial on Amazon Cocoa (Theobroma Cacao) in Ghana'. Experimental Agriculture 23, no. 1 (1987): 31–39. https://doi.org/10.1017/S0014479700001101.
3. Akpoti, Komlavi, Moctar Dembélé, Gerald Forkuor, Emmanuel Obuobie, Tafadzwanashe Mabhaudhi, and Olufunke Cofie. 'Integrating GIS and Remote Sensing for Land Use/Land Cover Mapping and Groundwater Potential Assessment for Climate-Smart Cocoa Irrigation in Ghana'. Scientific Reports 13, no. 1 (2023): 16025. https://doi.org/10.1038/s41598-023-43286-5.
4. Akpoti, Komlavi, Thomas P. Higginbottom, Timothy Foster, Roshan Adhikari, and Sander J. Zwart. 'Mapping Land Suitability for Informal, Small-Scale Irrigation Development Using Spatial Modelling and Machine Learning in the Upper East Region, Ghana'. Science of The Total Environment 803 (January 2022): 149959. https://doi.org/10.1016/j.scitotenv.2021.149959.
5. Ameyaw, George A., Owusu Domfeh, and Ebenezer Gyamera. 'Epidemiology and Diagnostics of Cacao Swollen Shoot Disease in Ghana: Past Research Achievements and Knowledge Gaps to Guide Future Research'. Viruses 16, no. 1 (2023): 43. https://doi.org/10.3390/v16010043.
6. Ansah, M., and Pabi Opoku, and Ayivor Jesse. 'The prospect of biodiversity conservation in cocoa agroforestry landscape, Ghana'. West African Journal of Applied Ecology 30, (July 2023): 44-45. https://www.researchgate.net/publication/372150965

7. Anselin, Luc. ‘Local Indicators of Spatial Association—LISA’. Geographical Analysis 27, no. 2 (1995): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.
8. Anselin, Luc, Anil K. Bera, Raymond Florax, and Mann J. Yoon. ’Simple Diagnostic Tests for Spatial Dependence.’ Regional Science and Urban Economics 26, no. 1 (1996): 77–104. https://doi.org/10.1016/0166-0462(95)02111-6.
9. Anselin, Luc. Spatial Econometrics: Methods and Models. Vol. 4. Studies in Operational Regional Science. Springer Netherlands, 1988. https://doi.org/10.1007/978-94-015-7799-1.
10. Antwi, Mary, Alfred Allan Duker, Mathias Fosu, and Robert Clement Abaidoo. ‘Geospatial Approach to Study the Spatial Distribution of Major Soil Nutrients in the Northern Region of Ghana’. Cogent Geoscience 2, no. 1 (2016): 1201906. https://doi.org/10.1080/23312041.2016.1201906.
11. Appiah, Mr, K Ofori-Frimpong, and Aa Afrifa. ‘Evaluation of Fertilizer Application on Some Peasant Cocoa Farms in Ghana’. Ghana Journal of Agricultural Science 33, no. 2 (2000): 183–90. https://doi.org/10.4314/gjas.v33i2.1869.
12. Asante, Paulina A., Eric Rahn, Niels P.R. Anten, Pieter A. Zuidema, Alejandro Morales, and Danae M.A. Rozendaal. ‘Climate Change Impacts on Cocoa Production in the Major Producing Countries of West and Central Africa by Mid-Century’. Agricultural and Forest Meteorology 362 (March 2025): 110393. https://doi.org/10.1016/j.agrformet.2025.110393.
13. Asare, Richard, Bo Markussen, Rebecca Ashley Asare, Gilbert Anim-Kwapong, and Anders Ræbild. ‘On-Farm Cocoa Yields Increase with Canopy Cover of Shade Trees in Two Agro-Ecological Zones in Ghana’. Climate and Development 11, no. 5 (2019): 435–45. https://doi.org/10.1080/17565529.2018.1442805.
14. Attipoe, Sonny Gad, Jian-min Cao, Yaa Opoku-Kwanowaa, and Frank Ohene-Sefa. “Assessing the Impact of Non-Governmental Organization’s Extension Programs on Sustainable Cocoa Production and Household Income in Ghana.” Journal of Integrative Agriculture 20, no. 10 (2021): 2820–36. https://doi.org/10.1016/S2095-3119(21)63607-9.
15. Baddeley, Adrian, Ege Rubak, and Rolf Turner. Spatial Point Patterns: Methodology and Applications with R. 0 ed. Chapman and Hall/CRC, 2015. https://doi.org/10.1201/b19708.
16. Barrett, Christopher B., Marc F. Bellemare, and Janet Y. Hou. ‘Reconsidering Conventional Explanations of the Inverse Productivity–Size Relationship’. World Development 38, no. 1 (2010): 88–97. https://doi.org/10.1016/j.worlddev.2009.06.002.
17. Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio. Applied Spatial Data Analysis with R. Springer New York, 2013. https://doi.org/10.1007/978-1-4614-7618-4.
18. Bivand, Roger, Jan Hauke, and Tomasz Kossowski. ’Computing the J Acobian in G Aussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods.’ Geographical Analysis 45, no. 2 (2013): 150–79. https://doi.org/10.1111/gean.12008.
19. Bivand, Roger, and Gianfranco Piras. ’Comparing Implementations of Estimation Methods for Spatial Econometrics.’ Journal of Statistical Software 63, no. 18 (2015). https://doi.org/10.18637/jss.v063.i18.
20. Bunn, Christian, Peter Läderach, Amos Quaye, Sander Muilerman, Martin R.A. Noponen, and Mark Lundy. ‘Recommendation Domains to Scale out Climate Change Adaptation in Cocoa Production in Ghana’. Climate Services 16 (December 2019): 100123. https://doi.org/10.1016/j.cliser.2019.100123.
21. Carletto, Calogero, Sara Savastano, and Alberto Zezza. ’Fact or Artifact: The Impact of Measurement Errors on the Farm Size–Productivity Relationship.’ Journal of Development Economics 103 (July 2013): 254–61. https://doi.org/10.1016/j.jdeveco.2013.03.004.
22. Chamberlin, Jordan. ’It’s a small world after all: Defining smallholder agriculture in Ghana’. Journal of International Food Policy Research Institute (IFPRI), IFPRI discussion papers, (January 2008). https://www.researchgate.net/publication/23778935.
23. Chen, Yanguang. ‘An Analytical Process of Spatial Autocorrelation Functions Based on Moran’s Index’. PLOS ONE 16, no. 4 (2021): e0249589. https://doi.org/10.1371/journal.pone.0249589.
24. ‘Cocobod - Increasing cocoa production in Ghana - the importance of the 4PS’. Accessed 15 October 2025. https://cocobod.gh/news/increasing-cocoa-production-in-ghana-the-importance-of-the-4ps.

25. 'Cocobod- Review of the producer price of cocoa for the 2024/2025 cocoa season'. Accessed 15 October 2025. https://cocobod.gh/news/review-of-the-producer-price-of-cocoa-for-the-20242025-cocoa-season-wednesday-11th-september-2024.
26. Cragg, John G. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." Econometrica 39, no. 5 (1971): 829. https://doi.org/10.2307/1909582.
27. Doaré, Fabien, Fabienne Ribeyre, and Christian Cilas. 'Genetic and Environmental Links between Traits of Cocoa Beans and Pods Clarify the Phenotyping Processes to Be Implemented.' Scientific Reports 10, no. 1 (2020): 9888. https://doi.org/10.1038/s41598-020-66969-9.
28. Debrah, Godwin, and Kwami Adanu. 'Does the Inverse Farm Size-Productivity Hypothesis Hold Beyond Five Hectares? Evidence from Ghana'. Journal of Agricultural and Applied Economics 54, no. 3 (2022): 548–59. https://doi.org/10.1017/aae.2022.20.
29. Diggle, Peter J. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. 0 ed. Chapman and Hall/CRC, 2013. https://doi.org/10.1201/b15326.
30. Donkor, Ebenezer, Emmanuel Dela Amegbe, Tomas Ratinger, and Jiri Hejkrlik. 'The Effect of Producer Groups on the Productivity and Technical Efficiency of Smallholder Cocoa Farmers in Ghana'. PLOS ONE 18, no. 12 (2023): e0294716. https://doi.org/10.1371/journal.pone.0294716.
31. Duan, Naihua. "Smearing Estimate: A Nonparametric Retransformation Method." Journal of the American Statistical Association 78, no. 383 (1983): 605–10. https://doi.org/10.1080/01621459.1983.10478017.
32. F., Aneani, and Ofori-Frimpong K. 'An Analysis of Yield Gap and Some Factors of Cocoa (Theobroma Cacao) Yields in Ghana'. Sustainable Agriculture Research 2, no. 4 (2013): 117. https://doi.org/10.5539/sar.v2n4p117.
33. Forkuor, G., P. Pavelic, E. Asare, and E. Obuobie. 'Modelling Potential Areas of Groundwater Development for Agriculture in Northern Ghana Using GIS/RS'. Hydrological Sciences Journal 58, no. 2 (2013): 437–51. https://doi.org/10.1080/02626667.2012.754101.
34. Foster, Andrew D., and Mark R. Rosenzweig. 'Are There Too Many Farms in the World? Labor Market Transaction Costs, Machine Capacities, and Optimal Farm Size'. Journal of Political Economy 130, no. 3 (2022): 636–80. https://doi.org/10.1086/717890.
35. Getis, Arthur, and J. K. Ord. 'The Analysis of Spatial Association by Use of Distance Statistics'. Geographical Analysis 24, no. 3 (1992): 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.
36. 'Ghana's 2024 Cocoa Export Revenue Crashes to $1.7 Billion—Lowest in 15 Years - MyJoyOnline'. Accessed 15 October 2025.https://www.myjoyonline.com/ghanas-2024-cocoa-export-revenue-crashes-to-1-7-billion-lowest-in-15-years/.
37. Griffith, D.A. 'Book Review: Cliff, A. D. and Ord, J. K. 1981: Spatial Processes - Models and Applications. London: Pion. Xii + 266 Pp. £12.50'. Progress in Human Geography 7, no. 1 (1983): 149–50. https://doi.org/10.1177/030913258300700115.
38. Guy Nkamleu, Joachim Nyemeck & Jim Gockowsk. 'Working Paper 104 - Technology Gap and Efficiency in Cocoa Production in West and Central Africa: Implications for Cocoa Sector Development,' Working Paper Series 241. (2010). African Development Bank. https://ideas.repec.org/p/adb/adbwps/241.html.
39. International Fertilizer Development Center (IFDC). 'Spatial analysis of maize yield and yield gap in Ghana (FERARI Research Report No. 13)'. Accessed 15 October 2025. https://ifdc.org/wp-content/uploads/2024/06/IFDC-FERARI-Research-Report-No-13-Final.pdf
40. Kilic, Talip, Amparo Palacios-López, and Markus Goldstein. "Caught in a Productivity Trap: A Distributional Perspective on Gender Differences in Malawian Agriculture." World Development 70 (June 2015): 416–63. https://doi.org/10.1016/j.worlddev.2014.06.017.
41. Kongor, John Edem, Hans De Steur, Davy Van De Walle, et al. 'Constraints for Future Cocoa Production in Ghana'. Agroforestry Systems 92, no. 5 (2018): 1373–85. https://doi.org/10.1007/s10457-017-0082-9.
42. Läderach, P., A. Martinez-Valle, G. Schroth, and N. Castro. 'Predicting the Future Climatic Suitability for Cocoa Farming of the World's Leading Producer Countries, Ghana and Côte d'Ivoire'. Climatic Change 119, nos. 3–4 (2013): 841–54. https://doi.org/10.1007/s10584-013-0774-8.

43. LeSage, James P. "Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models." Geographical Analysis 32, no. 1 (2000): 19–35. https://doi.org/10.1111/j.1538-4632.2000.tb00413.x.
44. LeSage, James, and Robert Kelley Pace. Introduction to Spatial Econometrics. 0 ed. Chapman and Hall/CRC, 2009. https://doi.org/10.1201/9781420064254.
45. Manning, Willard G, and John Mullahy. "Estimating Log Models: To Transform or Not to Transform?" Journal of Health Economics 20, no. 4 (2001): 461–94.https://doi.org/10.1016/S0167-6296(01)00086-8.
46. Ofori, Atta, Francis K. Padi, George A. Ameyaw, et al. 'Field Evaluation of the Impact of Cocoa Swollen Shoot Virus Disease Infection on Yield Traits of Different Cocoa (Theobroma Cacao L.) Clones in Ghana'. PLOS ONE 17, no. 1 (2022): e0262461. https://doi.org/10.1371/journal.pone.0262461.
47. Osei-Gyabaah, Augustine Prosper, Mary Antwi, Solomon Addo, and Paul Osei. 'Land Suitability Analysis for Cocoa (Theobroma Cacao) Production in the Sunyani Municipality, Bono Region, Ghana'. Smart Agricultural Technology 5 (October 2023): 100262. https://doi.org/10.1016/j.atech.2023.100262.
48. Owiredu, Patrick, Camillus Abawiera Wongnaa, Patricia Pinamang Acheampong, Monica Addison, Kwaku Agyei Adu, and Dadson Awunyo-Vitor. 'Farmer Business School Participation and Its Impact on Cocoa Productivity and Food Security in Ghana.' Journal of Agribusiness in Developing and Emerging Economies 14, no. 3 (2022): 637–54. https://doi.org/10.1108/JADEE-05-2022-0102.
49. Oxford Institute of Population Ageing. 'Cocoa and Continuity: Ageing in the Ghanaian Cocoa Sector'. Accessed 15 October 2025. https://www.ageing.ox.ac.uk/blog/Cocoa-and-Continuity-Ageing-in-the-Ghanaian-Cocoa-Sector.
50. Parra-Paitan, Claudia, Patrick Meyfroidt, Peter H. Verburg, and Erasmus K.H.J. Zu Ermgassen. 'Deforestation and Climate Risk Hotspots in the Global Cocoa Value Chain'. Environmental Science & Policy 158 (August 2024): 103796. https://doi.org/10.1016/j.envsci.2024.103796.
51. Quaye, Amos Kojo, Eric Kofi Doe, Emmanuel Morgan Attua, et al. 'Geospatial Distribution of Soil Organic Carbon and Soil pH within the Cocoa Agroecological Zones of Ghana.' Geoderma 386 (March 2021): 114921. https://doi.org/10.1016/j.geoderma.2020.114921.
52. 'Review of Producer Price for the 2025/26 Cocoa Season | Ministry of Finance | Ghana'. Accessed 15 October 2025. https://mofep.gov.gh/news-and-events/2025-10-02/%20%20review-of-producer-price-for-the-2025-26-cocoa-season.
53. Ripley, B. D. 'Modelling Spatial Patterns'. Journal of the Royal Statistical Society Series B: Statistical Methodology 39, no. 2 (1977): 172–92. https://doi.org/10.1111/j.2517-6161.1977.tb01615.x.
54. Ruf, François, and Schroth Götz. 'Chocolate forests and monocultures: an historical review of cocoa growing and its conflicting role in tropical deforestation and forest conservation'. Agroforestry and Biodiversity Conservation in Tropical Landscapes (January 2004): 107-134. https://www.researchgate.net/publication/261713726
55. Schroth, Götz, Peter Läderach, Armando Isaac Martinez-Valle, Christian Bunn, and Laurence Jassogne. 'Vulnerability to Climate Change of Cocoa in West Africa: Patterns, Opportunities and Limits to Adaptation'. Science of The Total Environment 556 (June 2016): 231–41. https://doi.org/10.1016/j.scitotenv.2016.03.024.
56. Uwagboe, Eghosa Osas, Busayo Solomon Famuyiwa, and Endurance Eniola Omiunu Agbebaku. 'Cocoa Farmers Attitude towards Utilisation of Integrated Pest Management in Edo and Ogun States of Nigeria'. Journal of Agricultural Extension 21, no. 1 (2017): 67. https://doi.org/10.4314/jae.v21i1.6.
57. Wessel, Marius, and P.M. Foluke Quist-Wessel. 'Cocoa Production in West Africa, a Review and Analysis of Recent Developments'. NJAS: Wageningen Journal of Life Sciences 74–75, no. 1 (2015): 1–7. https://doi.org/10.1016/j.njas.2015.09.001.
58. Wilhelm, Stefan, and Miguel Matos Godinho,de. "Estimating Spatial Probit Models in R." The R Journal 5, no. 1 (2013): 130. https://doi.org/10.32614/RJ-2013-013.

59. World Bank. 'Ghana tree crop diversification project (P180060)'. Accessed 15 October 2025. https://documents1.worldbank.org/curated/en/099070125024510190/pdf/P180060-f3ffceea-4ccf-4eb8-ad7d-b0bb0edda184.pdf
60. World Population Review. 'Cocoa Producing Countries 2025'. Accessed 15 October 2025. https://worldpopulationreview.com/country-rankings/cocoa-producing-countries.