



OPEN ACCESS

EDITED BY

Aqeel Ahmad,
University of Florida, United States

REVIEWED BY

Martín Alfredo Legarreta-González,
Monterrey Institute of Technology and
Higher Education (ITESM), Mexico
Selfia Bintariningtyas,
Sebelas Maret University, Indonesia

*CORRESPONDENCE

Welson Bentum
✉ wbentum1@st.knust.edu.gh

RECEIVED 06 June 2026

REVISED 13 July 2026

ACCEPTED 17 July 2026

PUBLISHED 13 August 2026

CITATION

Bentum W, Owiredo EO, Frempong NK,
Ofosu MA, Boateng E, Owusu KO,
Shadrack A, Asiedu EA, Addai-Henne S,
Pels WA and Dwamena HA (2026)
Beyond farm size: spatial determinants
of cocoa productivity in Ashanti
Region, Ghana.
Front. Agron. 8:1901636.
doi: 10.3389/fagro.2026.1901636

COPYRIGHT

© 2026 Bentum, Owiredo, Frempong,
Ofosu, Boateng, Owusu, Shadrack,
Asiedu, Addai-Henne, Pels and Dwamena.
This is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Beyond farm size: spatial determinants of cocoa productivity in Ashanti Region, Ghana

Welson Bentum^{1*}, Emmanuel Odame Owiredo¹,
Nana Kena Frempong¹, Michael Asante Ofosu²,
Esther Boateng¹, Kwaku Obeng Owusu¹, Agyeman Shadrack¹,
Elvis Anyimadu Asiedu¹, Sandra Addai-Henne¹,
Wilhemina Adoma Pels¹ and Harriet Achiaa Dwamena³

¹Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, ²Department of Mathematics and Statistics, Auburn University, Auburn, AL, United States, ³CSIR-Crops Research Institute, Kumasi, Ghana

Cocoa yields in Ghana have declined 23% since 2020 despite favorable prices, yet the spatial dimensions of this productivity crisis remain under-researched. This study 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, and 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. A two-hurdle spatial regression approach was employed to analyze cocoa yield determinants, while addressing the substantial zero inflation in production data. More than 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. In Hurdle 1, a spatial autoregressive probit model was applied to estimate the binary participation decision, while Hurdle 2 employed spatial error regression among producing farms to identify conditional yield determinants. The 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 92.07% of yield variation on the original scale using Duan's smearing estimator for backtransformation. Policy implications emphasize geographically targeted interventions, smallholder intensification through replanting programs, and sustained extension engagement.

KEYWORDS

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

1 Introduction

Despite decades of agricultural development efforts, productivity growth in Sub-Saharan Africa has lagged behind population growth Wessel and Quist-Wessel (2015); Asante et al. (2022), with cocoa, one of the region's most valuable export crops, exhibiting persistently low yields Abdulai et al. (2020) and experiencing an acute production decline since its 2020/2021 peak (Ghana Cocoa Board (COCOBOD), 2024a). West Africa produces about 64% of global cocoa supply, with Ghana ranked second worldwide (Asante et al., 2022). Ghana achieved peak production of 1,047,000 metric tons in 2020 (Ghana Cocoa Board (COCOBOD), 2024a), but output declined 49.3% to approximately 531,000 metric tons in 2023/2024 United States Department of Agriculture Foreign Agricultural Service (2025), representing the lowest level since 2008 when production averaged 680,000 metric tons annually. This decline reflects a return to pre-2010 production levels despite favorable price conditions. This decline threatens both national revenue as cocoa contributes \$1.7 billion annually to Ghana's exports and the livelihoods of 800,000 smallholder families (World Bank, 2022). Producer prices increased 147% between 2023–2025, yet production continued to decline, suggesting that price signals alone cannot reverse productivity losses (Ghana Cocoa Board (COCOBOD), 2024b; Ministry of Finance, Ghana, 2025). This disconnect between prices and production suggests that non-economic constraints, particularly spatial heterogeneity in farm management and environmental conditions drive yield outcomes (Abdulai et al., 2020). 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 gains¹. Average smallholder yields of 400–554 kg/ha represent only 21–29% of experimentally achieved potentials (1,891–3,246 kg/ha) (Ahenkorah et al., 1987; Appiah et al., 2000; Aneani and Ofori-Frimpong, 2013), 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 modelling (Osei-Gyabaah et al., 2023; Parra-Paitan et al., 2024) or climate change projections (Schroth et al., 2016; Asante et al., 2025) 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.

2 Literature review

2.1 The yield gap paradox in Ghanaian cocoa

Previous research on cocoa productivity in Ghana and Sub-Saharan Africa has primarily emphasized biological constraints, including aging trees, pests, and climate change, while largely overlooking spatial heterogeneity in farm management as a yield determinant (Kongor et al., 2018; Schroth et al., 2016; Asitoakor et al., 2022; Abdulai et al., 2020). 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 (Ahenkorah et al., 1987; Appiah et al., 2000; Aneani and Ofori-Frimpong, 2013). 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%. 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, with neighboring farm households exhibiting similar yields due to shared environmental conditions, institutional factors such as extension and irrigation access, and knowledge spillovers facilitated by geographic proximity (Kondo et al., 2026). 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 (Forkuor et al., 2013; Akpoti et al., 2022; Osei-Gyabaah et al., 2023) but do not analyze yield variation across operating farms. Climate change projection studies employ spatial models to forecast future cocoa suitability under warming scenarios (Bunn et al., 2019; Läderach et al., 2013; Schroth et al., 2016) but do not explain current productivity patterns. Most studies (Aneani and Ofori-Frimpong, 2013; Abdulai et al., 2020; Asante et al., 2022; Wessel and Quist-Wessel, 2015; Wongnaa and Babu, 2020) examining cocoa yield variation have focused primarily on conventional determinants with limited attention given to the application of point pattern analysis and spatial analysis techniques that are essential for geographically targeted interventions. Spatial econometric methods have seen limited application in African agricultural research more broadly, with existing work concentrated on cross-national analyses of agricultural-environmental outcomes using spatial Durbin models (Sunge and Espoir, 2026) or harmonized panel evidence on smallholder productivity trends (Wollburg et al., 2024), rather than farm-level yield determinants within a single crop or region.

Discrete choice models have been widely applied to smallholder cocoa decision making in Ghana, including fertilizer adoption

¹ MyJoyOnline. *Ghana's 2024 Cocoa Export Revenue Crashes to \$1.7 Billion –Lowest in 15 Years*. 2025. Available at: <https://www.myjoyonline.com/ghanas-2024-cocoa-export-revenue-crashes-to-1-7-billion-lowest-in-15-years/> [Accessed 15 October 2025].

(Anang, 2016), organic production adoption (Djokoto et al., 2016), agroforestry tree density choice through an ordered probit (Amerino et al., 2024), and organic pesticide adoption through probit analysis with a log likelihood-link (Asigbaase et al., 2025). None of these studies, however, relax the assumption that farm level outcomes are independently distributed across space. This omission carries a methodological cost, since ignoring spatial autocorrelation among geographically proximate decision units may produce biased parameter estimates in binary outcome models (Pinkse and Slade, 1998), a problem that motivated the development of spatial probit estimators such as the Bayesian formulation of probit model with spatial dependencies Smith and LeSage (2004). Applications of spatial probit models have since appeared in contingent valuation research (Loomis and Mueller, 2013) and examining spatially heterogeneous preferences for coastal ecosystem restoration (Kabaya, 2022), where accounting for spatial dependence shifted estimated economic values substantially relative to a standard, non spatial probit. To date, no published study of cocoa productivity or adoption behavior in Ghana has incorporated a spatially autoregressive probit component, despite a strong *a priori* case for spatial dependence in farm level participation decisions driven by shared agroecological conditions, localized extension delivery, and farmer to farmer information diffusion.

2.3 The inverse farm-size productivity relationship

Agricultural economics literature consistently documents that smaller farms achieve higher yields per hectare than larger operations, a phenomenon referred to as “inverse relationship” (IR) (Debrah and Adanu, 2022; Barrett et al., 2010; Foster and Rosenzweig, 2022). Several mechanisms have been proposed to explain this inverse relationship, these include labor intensity, where smallholders apply more family labor per hectare (Debrah and Adanu, 2022), management quality differences as smaller farms receive more intensive oversight (Donkor et al., 2023; Kongor et al., 2018), and land quality selection patterns where small farms may occupy better soils. (Barrett et al., 2010) investigated the IR using data collected in Madagascar, while (Debrah and Adanu, 2022) found an L-shaped relationship in Ghanaian agriculture. However, it remains unclear whether this pattern holds for cocoa, a perennial tree crop with distinct management requirements compared to annual crops, and whether the relationship exhibits spatial variations.

3 Materials and methods

3.1 Research design and study area

The study was conducted in Ashanti Region of Ghana, the country’s leading cocoa-producing area. Figure 1 shows the distribution of cocoa farms geographically across Ghana’s Ashanti region.

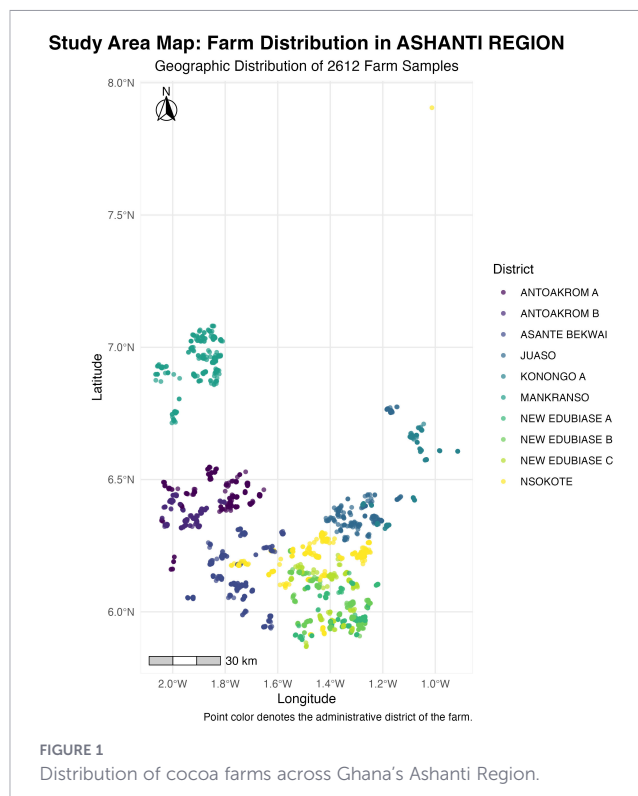


FIGURE 1
Distribution of cocoa farms across Ghana’s Ashanti Region.

A total of 2612 cocoa farms were randomly sampled 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 (Equation 1), which measures the degree to which farms with similar productivity levels are geographically proximate (Anselin, 1995, 1988). 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 (Bivand et al., 2013).

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (1)$$

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 autocorrelation, often assuming normality or using permutation tests.

3.2.2 Point pattern analysis

Using Ripley's K-function (Equation 2), we assessed whether farm locations clustered at multiple spatial scales (Ripley, 1977). 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 (Diggle, 2013). Additionally, we calculated Average Nearest Neighbor (ANN) statistics to test for localized clustering, comparing observed mean nearest-neighbor distances to expected distances under CSR (Equation 3).

Ripley's K function:

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

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 (G) Function:

$$G(r) = \frac{\sum_{i=1}^n I(d_i \leq r)}{n} \quad (3)$$

3.2.3 Spatial clustering

We identified distinct productivity zones using hierarchical spatial clustering of farm coordinates and yield. We applied Ward's linkage method with Euclidean distance to group farms into groups based on both geographic proximity and yield similarity. The optimal number of clusters was determined using the elbow method and silhouette scores.

3.3 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 used a two-hurdle spatial regression framework to unravel these processes (Cragg, 1971). 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.

3.4 Hurdle 1: spatial autoregressive probit for production participation

3.4.1 Model specification

We define a binary outcome variable (Equation 4):

$$Y_i = \begin{cases} 1 & \text{if farm } i \text{ produced cocoa (yield} > 0 \text{ kg/ha)} \\ 0 & \text{if farm } i \text{ produced no cocoa (yield} = 0 \text{ kg/ha)} \end{cases} \quad (4)$$

The spatial autoregressive probit (SAR Probit) model (Equations 5, 6) specifies:

$$y_i^* = \rho \mathbf{W} \mathbf{y}^* + \mathbf{X}_i \beta + u_i \quad (5)$$

$$y_i = 1(y_i^* > 0) \quad (6)$$

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
- β 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.

3.4.2 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:

$$\sum_{j=1}^n w_{ij} = 1 \quad \forall i \quad (7)$$

This standardization (Equation 7) allows ρ to be interpreted as the strength of spatial dependence, with values closer to 1 indicating stronger spatial correlation.

3.4.3 Estimation method

The SAR Probit model was estimated via Markov Chain Monte Carlo using Bayesian methods. MCMC is necessary because maximum likelihood estimation is computationally intractable for spatial probit models due to the high-dimensional integrals involved (LeSage, 2000). The MCMC estimation proceeded with 20,000 post burn-in draws following a burn-in period of 2,000 iterations, with a thinning factor of 5 applied to reduce autocorrelation between successive draws. To assess convergence, three independent chains were run from different starting seeds, and the Brooks-Gelman-Rubin potential scale reduction statistic was computed for each parameter using the coda package (Plummer et al., 2006; Brooks and Andrew, 1998; Gelman and Rubin, 1992). All potential scale reduction factors equaled 1.00, confirming convergence to a common posterior distribution. 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.

3.5 Hurdle 2: spatial error model for conditional productivity

3.5.1 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 two candidate functional forms:

1. Original scale: $Y_i = \mathbf{X}_i\beta + \varepsilon_i$
2. Log-linear: $\log(Y_i) = \mathbf{X}_i\beta + \varepsilon_i$

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.

3.5.2 OLS baseline and lagrange multiplier tests

We first estimated an OLS baseline model (Equation 8) on the positive-yield sample:

$$\begin{aligned} \log(\text{yield}_i) = & \beta_0 + \beta_1(\text{farm size})_i + \beta_2(\text{surface mapped})_i \\ & + \beta_3(\text{nb farms 2122})_i \\ & + \beta_4(\text{received remediation})_i + \beta_5(\text{covered 2122})_i \\ & + \beta_6(\text{farmer coaching})_i \\ & + \beta_7(\text{fdp previously})_i + \beta_8(\text{map farm})_i \\ & + \beta_9(\text{gender})_i \\ & + \beta_{10}(\text{household size})_i + \beta_{11}(\text{nb farms})_i \\ & + \beta_{12} \log(\text{yield per tree})_i \\ & + \beta_{13}(\text{pods per tree})_i + \beta_{14} \log \\ & (\text{tree density} + 1)_i + \varepsilon_i \end{aligned} \quad (8)$$

Residuals from the OLS model were tested for spatial autocorrelation using Lagrange Multiplier (LM) tests implemented in `lm.LMtests()` from the `spdep` package (Bivand et al., 2013). 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 (Anselin et al., 1996).

3.5.3 Spatial model specifications

3.5.3.1 Spatial lag model

$$\begin{aligned} \log(\text{yield}_i) = & \rho \sum_j w_{ij} \log(\text{yield}_j) + \beta_0 + \beta_1(\text{farm size})_i \\ & + \beta_2(\text{surface mapped})_i + \beta_3(\text{nb farms 2122})_i \\ & + \beta_4(\text{received remediation})_i + \beta_5(\text{covered 2122})_i \\ & + \beta_6(\text{farmer coaching})_i + \beta_7(\text{fdp previously})_i \\ & + \beta_8(\text{map farm})_i + \beta_9(\text{gender})_i + \beta_{10}(\text{household size})_i \\ & + \beta_{11}(\text{nb farms})_i + \beta_{12} \log(\text{yield per tree})_i \\ & + \beta_{13}(\text{pods per tree})_i + \beta_{14} \log(\text{tree density} + 1)_i + \varepsilon_i, \end{aligned} \quad (9)$$

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

3.5.3.2 Spatial error model

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

$$u_i = \lambda \sum_j w_{ij} u_j + \varepsilon_i, \quad (11)$$

where u_i (Equation 11) 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 (Bivand and Piras, 2015). 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.

3.5.4 Addressing the multiplicative relationship between yield components

Following standard practices in agricultural economics and crop science, the model utilizes an agronomic yield component decomposition by including both tree density and yield per tree as separate predictors. Although yield per hectare can be broken down mathematically into these two components, simultaneously incorporating them does not collapse the log-linear specification into a deterministic identity. This model structure is justified by two primary methodological and empirical factors. First, the two components are derived from independent data generation processes subject to non-systematic measurement error, which prevents an exact mechanical overlap. Tree density is calculated using objective enumerator field counts and GPS boundary mapping, whereas the yield per tree metric is constructed from farmer recalled harvest data. Second, tree density acts as an indicator for broader farm management intensity and intra-canopy competition rather than a simple geometric scaling factor. Consequently, the log-transformed specification allows the model to estimate genuine structural and behavioral relationships without running into perfect multicollinearity. The remaining variation leaves sufficient degrees of freedom to identify the statistical significance of key exogenous variables, including farm size, surface mapped, and prior FDP participation, to isolate the distinct pathways of agricultural productivity.

3.5.5 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 (Equation 12):

$$E[Y|X] = E[\exp(\log Y|X)] \neq \exp(E[\log Y|X]) = \exp(\mathbf{X}\beta) \quad (12)$$

To obtain unbiased predictions, we employed Duan's smearing estimator (Equation 13) (Duan, 1983):

$$\hat{Y}_i = \exp(\mathbf{X}_i \hat{\beta}) \times \phi \quad (13)$$

where the smearing factor (Equation 14) is:

$$\phi = \frac{1}{n} \sum_{i=1}^n \exp(\hat{u}_i) \quad (14)$$

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 (Manning and Mullahy, 2001).

3.5.6 Model diagnostics

The spatial dependence was assessed by analyzing the SEM residuals for spatial autocorrelation using Moran's I test to check if the spatial dependence was properly considered. The Breusch-Pagan test of residuals versus fitted values was used to check for heteroskedasticity and Cook's distance and leverages plots were used to find influential observations. Furthermore, geographic mapping of residuals was done to investigate spatial residual patterns and to search for any residual spatial structure remaining.

3.6 Data collection and variable construction

3.6.1 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 socio-economic 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.

3.6.2 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.

3.6.3 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 (Equation 15) was calculated as:

$$\text{Yield per hectare} = \frac{\text{Total production (kg)}}{\text{GPS-measured farm area (ha)}} \quad (15)$$

3.7 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 agri-cultural 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

4 Results and discussion

4.1 District yield statistics

Cocoa yield distributions across the 10 sampled districts reveal pronounced spatial heterogeneity in production performance. Four districts, Juaso (282 farms), Konongo A (266 farms), Mankranso (221 farms), and New Edubiase C (213 farms), recorded zero mean and median yields with zero standard deviation, indicating complete farm-level non-production across these areas during the survey season. A chi-square test of homogeneity of proportions across districts confirms this heterogeneity across districts is not attributable to chance ($\chi^2 = 1198.6$, $df = 9$, $p < 2.2 \times 10^{-16}$), although with a sample this large, almost any real difference between districts would produce a significant result, so the test itself is not what makes this finding notable. We cannot fully distinguish the cause with a single cross-section, but several non-exclusive explanations are plausible. A coordinated agronomic shock, concentrated CSSVD infection, drought stress, or an aging tree cohort, could produce district-wide collapse, consistent with the spatial clustering in Section 4.3. A second possibility specific to this region is land-use conversion from illegal small-scale mining (galamsey), which is well documented in Ashanti cocoa districts including Amansie West, home to several sampled communities such as Antoakrom, and can displace cultivation entirely through pit excavation and topsoil loss (Osman et al., 2022).

Among the six producing districts, New Edubiase B (214 farms) exhibited the highest mean yield at 116.95 kg/ha (median = 129.43 kg/ha; SD = 80.24), followed closely by New Edubiase A (221 farms; mean = 100.38 kg/ha; median = 85.62 kg/ha; SD = 80.22). The remaining producing districts performed substantially lower: Asante Bekwai (293 farms; mean = 58.14 kg/ha; median = 8.26 kg/ha; SD = 71.53), Antoakrom A (378 farms; mean = 35.31 kg/ha; median = 5.80 kg/ha; SD = 51.78), Nsokote (154 farms; mean = 34.06 kg/ha; median = 23.45 kg/ha; SD = 36.82), and Antoakrom B (370 farms; mean = 26.09 kg/ha; median = 20.43 kg/ha; SD = 26.95). The wide gap between zero-yield districts and even the lowest-performing producing districts, combined with the large standard deviations in Asante Bekwai and Antoakrom A, points to a threshold management effect rather than a smooth gradient in productivity across the study area.

4.2 Descriptive statistics

The descriptive statistics presented in [Table 1](#) 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.

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, a pattern attributed to the compounding effects of CSSVD infection, aging trees, erratic weather, and input deficiencies that made Ghana's 2023/2024 season its worst in 15 years ([United States Department of Agriculture Foreign Agricultural Service, 2025](#)). 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)

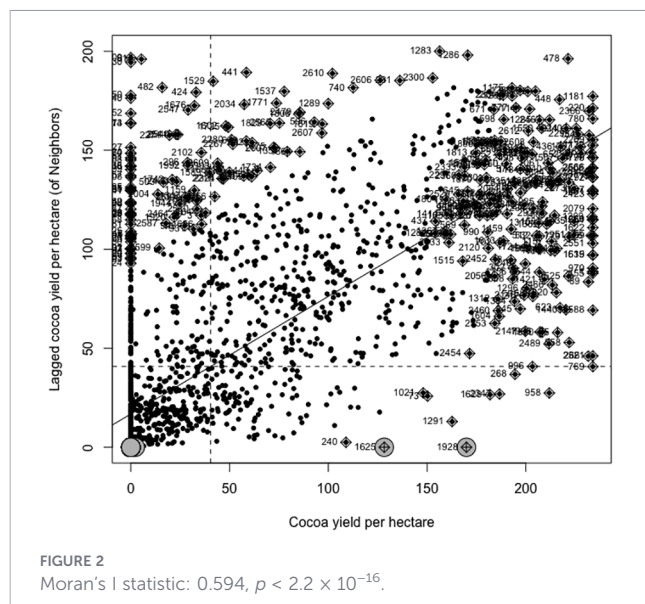
The Moran's I test under randomization was conducted to formally assess the degree of spatial autocorrelation in property values. The results are summarized in [Figure 2](#).

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 and spatial relationship has also been documented on various levels in other

TABLE 1 Descriptive statistics.

Sample characteristics	Value					
Number of observations (N)	2,612					
Number of districts	10					
Number of communities covered under the 10 districts	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

Statistics are computed over the full sample ($N = 2,612$), including 1,434 farms (54.9%) with zero recorded yield. Statistics restricted to the 1,178 producing farms are reported in Section 4.7.



crop yields (Bua et al., 2020) and studies in Ghana (Naazie et al., 2024) and Kenya (Sussy et al., 2019), 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

Figure 3 presents the results of the point pattern analysis using Ripley's K-function and the nearest neighbor G-function. Ripley's K-function analysis confirmed non-random clustering across all spatial scales. 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, with the increasing divergence with distance reflecting progressively stronger aggregation at regional scales. The extreme magnitude of clustering likely reflects structural land-use organization rather than random concentration of farms. Cocoa farming in West Africa has historically expanded through forest clearing and settlement along established land corridors (Ruf and Schroth, 2004), and the resulting inherited farm locations, combined with limited availability of suitable land, produce the geographically concentrated patterns observed in our data. 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.

The nearest neighbor G-function further revealed pronounced short-distance clustering, with 90% of farms located within 0.4 km of another farm compared to only 5% expected under CSR (Kolmogorov-Smirnov test: $D = 0.85$, $p < 0.001$). This tight spatial aggregation has important 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 (Ofori et al., 2022; Ameyaw et al., 2023).

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. 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 2. 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.

These clusters point toward differentiated targeting rather than a uniform program. Clusters 0 and 1 (Mankranso/Antoakrom A; Konongo A/Juaso; combined $n = 563$, mean yield near zero) may benefit from a diagnostic assessment first, disease incidence, tree age, land-use status, before input subsidies are committed. Cluster 4 (Nsokote/New Edubiase C, $n = 765$, the largest cluster) offers the greatest aggregate leverage for density-focused replanting given its size. Cluster 7 (Asante Bekwai/New Edubiase C, $n = 77$, the smallest cluster) may be better suited to individualized farm visits than a broader rollout. Clusters 3 and 2 (moderate- and higher-producing) are reasonable candidates for land-documentation support and management coaching to facilitate credit access. Cluster 5 (New Edubiase A/B, highest mean yield) could serve as a reference site for farmer-field-school demonstrations, testing whether its practices transfer to lower-performing clusters.

4.6 Farm size vs. productivity (yield per surface area)

Figure 5 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

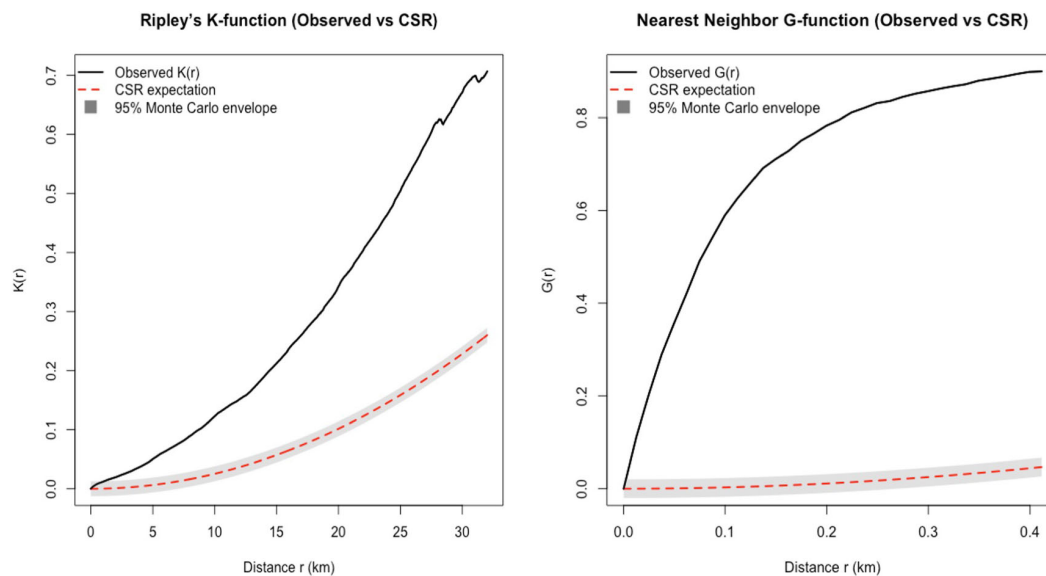


FIGURE 3

Ripley's K-function (left) and Nearest Neighbor G function (right) analysis of spatial clustering.

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. This inverse farm size-productivity relationship is well-documented in agricultural economics literature and has been observed across various crops and regions. Debrah and Adanu (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 (Debrah and Adanu, 2022). Several theoretical explanations have been proposed for this phenomenon, Foster and Rosenzweig (2022) (Foster and Rosenzweig, 2022) 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



FIGURE 4

Spatial clusters of farm productivity.

TABLE 2 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

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

documented by Barrett et al. (2010) across multiple African countries (Barrett et al., 2010). In the context of cocoa farming in Ghana, where an estimated 90% of cocoa is grown by smallholders with farms less than 2 hectares (Wessel and Quist-Wessel, 2015). 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 (Donkor et al., 2023). Similar findings were reported by Kongor et al. (2018), who identified management intensity as a key constraint for larger cocoa farms in Ghana (Kongor et al., 2018).

- 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

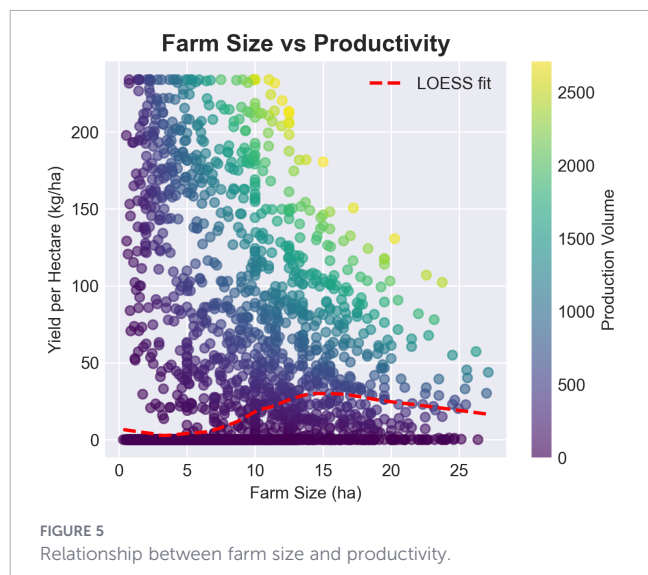
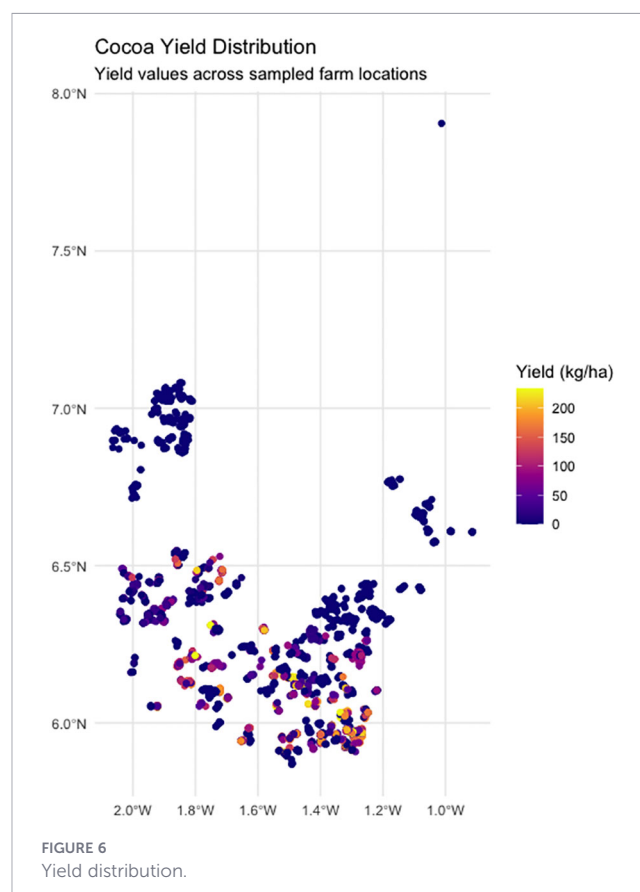
- Mean yield: 89.681 kg/ha (SD = 67.844)
- Median yield: 73.716 kg/ha
- Range: 0.326 to 233.916 kg/ha

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.

4.7 Yield distribution and sample characteristics

Figure 6 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)

Hurdle 1 breakdown:



4.8 Hurdle 1: spatial autoregressive probit for participation

This spatial autoregressive probit model 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.

Program interventions emerge as critical determinants: farms covered in the 2021–22 program show higher probability of being productive ($\beta = 2.117$, $p < 0.001$), as do those receiving farmer coaching ($\beta = 0.305$, $p = 0.004$), previous FDP participation ($\beta = 0.530$, $p = 0.002$), and farm mapping services ($\beta = 0.294$, $p = 0.006$), suggesting an association between program engagement and productive status. These findings are directionally consistent with research reporting positive associations between extension services and cocoa productivity in Ghana, with NGO extension programs increasing yields by 14.3% and farmer business school participation improving both productivity and food security (Attipoe et al., 2021; Owiredo et al., 2022). 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 ($\beta = 0.013$, $p = 0.001$). The negative coefficient on number of farms ($\beta = -0.414$, $p < 0.001$) 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 ($\beta = -0.040$, $p = 0.001$) hints that non-productive farms may compensate for low quality with high pod quantity, reflecting quality versus quantity tradeoffs. Larger farm sizes ($\beta = 0.050$, $p < 0.001$) are significantly associated with higher 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 ($\beta = 0.121$, $p < 0.001$) positively impacts productivity. Gender shows no significant effect ($\beta = 0.159$, $p = 0.129$). 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 ($\beta = 0.070$, $p = 0.005$) positively affects productivity, suggesting that proper farm documentation and potentially more secure land tenure encourage productivity improvements. However, receiving remediation shows no significant effect ($\beta = 0.132$, $p = 0.617$), 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. The high spatial correlation suggests that spatial targeting strategies addressing entire clusters, rather than isolated farms, may be more effective, as geographic concentration of non-productivity likely reflects self reinforcing spatial patterns of poor performance.

4.9 Hurdle 2: model selection

Table 3 compares the performance of the OLS, SLM, and SEM models (Equations 8–10) 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.

4.10 Spatial error model for conditional productivity

The Spatial Error Model (Table 4) 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

TABLE 3 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

TABLE 4 Hurdle 2: Spatial error model for conditional yield (producing farms only).

Variable	Coefficient	Std. Error	z-value	p-value
(Intercept)	-0.4897	0.04836	-10.1264	< 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) Spatial Parameter	1.2902	0.0126	102.2414	< 2.2e -16 ***
Lambda (λ)	0.690	0.023	30.619	< 2.2e -16 ***
Num. obs.	1,178			
Parameters estimated	17			
Log Likelihood	711.901			
ML Residual (sigma)	0.1266			
LR statistic	408.13			
Wald statistics	937.54, ($p < 2.22e -16$)			

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Several variables significant in the participation model (Hurdle 1, Table 5), including farm coverage in 2021-22, farmer coaching, previous FDP participation, and farm mapping, are not significant here. These variables appear to influence whether a farm produces any cocoa at all, but not how much a farm yields once it is already producing.

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 (Anselin, 1988; LeSage and Pace, 2009).

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 superelastic 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 (Ahenkorah et al., 1987) who found lower yield 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) (Ministry of Finance, Ghana, 2025), 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 evidence from rehabilitation programs in Ghana and Nigeria (Nkamleu et al., 2010). 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 (Doaré et al., 2020) 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 (Barrett et al., 2010; Carletto et al., 2013). 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

TABLE 5 Hurdle 1: MCMC spatial autoregressive probit model summary.

Execution time		1250.16 secs			
N draws	20000 (3 chains)	N omit (burn-in)	2000		
N observations	2,612	K covariates	15		
No. of 0 Y values	1434	No. of 1 Y values	1178		
Min rho	-1.000	Max rho	1.000		
	Estimate	Std. dev	p-level	t-value	Pr(> z)
(Intercept)	-4.423472	0.433339	0.000000	-10.208	< 2e-16 ***
Farm size (ha)	0.049736	0.009379	0.000000	5.303	1.24e-07 ***
Surface mapped	0.069719	0.024867	0.002150	2.804	0.005089 **
No. farms 2021-22	0.322540	0.048202	0.000000	6.691	2.69e-11 ***
Received remediation	0.131694	0.263281	0.313550	0.500	0.616974
Farm covered 2021-22	2.117021	0.260130	0.000000	8.138	6.66e-16 ***
Farmer coaching	0.311248	0.106245	0.001350	2.930	0.003424 **
FDP previously	0.530467	0.165595	0.000650	3.203	0.001374 **
Farm mapping	0.294868	0.105545	0.002350	2.794	0.005248 **
Gender (male)	0.158942	0.104656	0.062750	1.519	0.128956
Household size	0.121153	0.023646	0.000000	5.124	3.21e-07 ***
No. of farms	-0.413572	0.086886	0.000000	-4.760	2.04e-06 ***
Yield per tree	0.023584	0.001708	0.000000	13.805	< 2e-16 ***
Pods per tree	-0.039862	0.010784	0.000050	-3.696	0.000223 ***
Tree density	0.012890	0.003450	0.000550	3.737	0.000191 ***
P	0.697913	0.009163	0.000000	76.163	< 2e-16 ***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

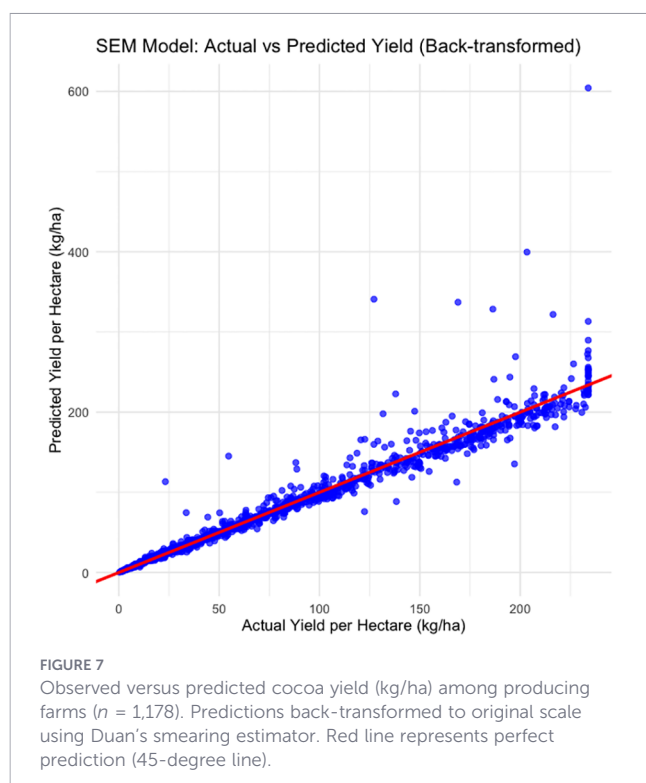
Log-linear not employed due to excessive 0 values in the full sample. Brooks-Gelman-Rubin potential scale reduction factors equaled 1.00 for all parameters across three independent chains. ***p<0.001; **p<0.01; *p<0.05.

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 (Kilic et al., 2015).

4.11 Model performance and predictive accuracy

Figure 7 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. Performance metrics on the original scale demonstrate exceptional model fit: $R^2 = 0.9207$ indicates that 92.07% 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.

4.12 Diagnostic tests

4.12.1 Multicollinearity assessment

VIF analysis confirms minimal multicollinearity across all predictors, with all values remaining well below the conventional threshold of 5. The two highest VIF values belong to tree density (VIF = 2.03) and farm size (VIF = 1.72), followed by yield per tree (VIF = 1.54), surface mapped (VIF = 1.46), and pods per tree (VIF = 1.45). The remaining predictors exhibit near-unity values, indicating excellent independence: farmer coaching (VIF = 1.41), number of farms 2021-22 (VIF = 1.35), number of farms (VIF = 1.21), farm mapping (VIF = 1.18), FDP previously (VIF = 1.16), household size (VIF = 1.09), gender (VIF = 1.05), received remediation (VIF = 1.04), and farm covered 2021-22 (VIF = 1.03). The absence of high inter-predictor correlations justifies the simultaneous inclusion of all covariates in both hurdle models, ensuring that coefficient estimates remain stable and interpretable.

4.12.2 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.

4.12.3 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.

4.13 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.

The ten most influential observations had standardized residuals ranging from -12.48 to 3.82 , with farm sizes between 1.35 and 13.47 ha and actual yields between 23.17 and 233.92 kg/ha. Observation 95 exhibited the most extreme residual (standardized residual = -12.48), corresponding to a 13.47 ha farm reporting an actual yield of 23.17 kg/ha against a predicted yield of 113.34 kg/ha — a prediction error of 90.17 kg/ha — despite recording zero tree density. The next most influential observations were Obs. 1100 (standardized residual = -7.73 ; 6.15 ha; actual yield 127.11 kg/ha; tree density 0.00 trees/ha), Obs. 789 (-7.65 ; 7.00 ha; 54.71 kg/ha; 0.00 trees/ha), Obs. 521 (-7.44 ; 5.90 ha; 233.92 kg/ha; 0.00 trees/ha), Obs.

1105 (-6.26 ; 4.62 ha; 33.57 kg/ha; 0.00 trees/ha), Obs. 1003 (-5.40 ; 4.15 ha; 168.99 kg/ha; 0.00 trees/ha), Obs. 394 (-5.28 ; 4.60 ha; 203.36 kg/ha; 0.00 trees/ha), Obs. 484 (-4.42 ; 4.10 ha;

186.37 kg/ha; 0.00 trees/ha), Obs. 524 (3.82; 1.35 ha; 122.43 kg/ha; 0.00 trees/ha), and Obs. 1090 (−3.73; 5.50 ha; 137.93 kg/ha; 44.28 trees/ha). The zero predicted errors combined with zero recorded tree density for 9 of the top 10 observations suggest incomplete farm mapping or data entry errors rather than genuine agronomic conditions, indicating systematic rather than random measurement problems in tree enumeration.

4.13.1 Sensitivity analysis

To assess model robustness, we re-estimated the spatial error model excluding observation 95. Table 6 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.

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. 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.

4.13.2 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

TABLE 6 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	−0.4897	−0.4151	0.0746	−15.2
Farm size (ha)	−0.0287	−0.0294	−0.0007	2.4
Surface mapped	0.0043	0.0038	−0.0005	−11.1
No. of farms 2021–22	−0.0136	−0.0141	−0.0005	3.7
Received remediation	−0.0011	−0.0011	0.0000	−3.2
Farm coverage 2021–22	0.0446	−0.0087	−0.0534	−119.6***
Farmer coaching	−0.0261	−0.0236	0.0025	−9.5
FDP previously	−0.0842	−0.0991	−0.0149	17.7
Farm mapping	−0.0051	−0.0061	−0.0010	19.9
Gender(male)	−0.0041	−0.0061	−0.0020	48.9
Household size	0.0029	0.0025	−0.0004	−14.4
No. of farms	0.0147	0.0160	0.0013	9.1
log (Yield per tree)	0.9745	0.9768	0.0023	0.2
Pods per tree	−0.0023	−0.0030	−0.0007	29.4
log (Tree density + 1)	1.2902	1.2783	−0.0119	−0.9
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

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

adequately captures unobserved location-specific factors generating extreme residuals, as evidenced by the model's high explanatory power ($R^2 = 0.9207$ 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.

4.14 Integrated analysis: comparing spatial determinants across the two-stage hurdle model

4.14.1 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 ($n = 2,612$), 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).

4.14.2 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.

4.14.3 Divergent effects across hurdles

Farm size exhibits contrasting effects across the two stages. Larger farms are significantly more likely to produce cocoa (Hurdle 1: $\beta = 0.0497$, $p < 0.001$), 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: $\beta = 0.0697$, $p = 0.0051$) 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: $\beta = 0.3225$, $p < 0.001$) 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.

4.15 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: $\beta = 0.1317$, $p = 0.6170$) 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: $\beta = 2.1170$, $p < 0.001$) but shows no significant effect on yield (Hurdle 2: $\beta = 0.0446$, $p = 0.0975$). This pattern indicates that program coverage in 2021–22 may have 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. That is, a positive and statistically significant for production likelihood (Hurdle 1: $\beta = 0.3112$, $p = 0.0034$) but negative and insignificant for yield (Hurdle 2: $\beta = -0.0261$, $p = 0.1532$), which may suggest 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: $\beta = 0.5305$, $p = 0.0014$) 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. Several non-exclusive mechanisms could produce this pattern and targeting bias is one possibility. Thus, if FDP historically prioritized underperforming farms for enrollment, farms with previous FDP exposure represent a selected group that was already disadvantaged before enrollment, so the negative coefficient may reflect persistent baseline conditions rather than a harmful program effect. There's also a mechanism of temporal decay, where previous FDP participation is defined over a 1–5 year

window, so input-based gains (pruning, fertilizer application) may have faded by the 2022/2023 season absent follow-up reinforcement, while farmers no longer receiving active support may also reduce complementary investment if they assumed earlier gains were durable. A third possibility involves program exit itself: farms with previous but not current FDP status are, by construction, farms that left the program, and exit may correlate with unobserved factors, such as an aging farmer, a land dispute, or relocation, that are independently associated with lower yield. Distinguishing these mechanisms would require participation histories linked to pre-enrollment yield data, which this cross-sectional design cannot provide. Farm mapping shows positive and marginally significant effects on production decisions (Hurdle 1: $\beta = 0.2949$, $p = 0.0052$) but negative and insignificant effects on yield (Hurdle 2: $\beta = -0.0051$, $p = 0.5627$). An implication is that mapping interventions help facilitate entry into production but do not contribute to enhanced productivity once farms are producing.

4.16 Demographic factors: limited direct effects

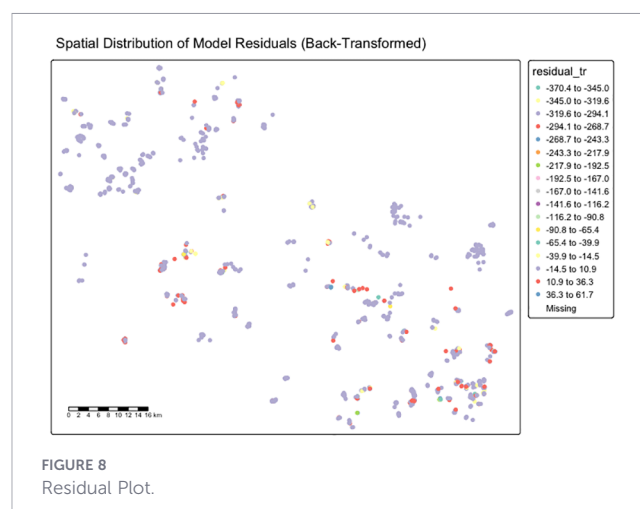
Gender and household size exhibit minimal influence on production outcomes. Gender is insignificant in both stages (Hurdle 1: $\beta = 0.1589$, $p = 0.1290$; Hurdle 2: $\beta = -0.0041$, $p = 0.6000$), as is household size in the second stage (Hurdle 2: $\beta = 0.0029$, $p = 0.2107$), despite being highly significant in the first stage (Hurdle 1: $\beta = 0.1212$, $p < 0.001$). The number of farms managed by a household shows negative and highly significant effects on production participation (Hurdle 1: $\beta = -0.4136$, $p < 0.001$) 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.

4.16.1 Spatial residual plot

The spatial distribution of the SEM residuals (Figure 8) 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.

4.17 Correlation matrix

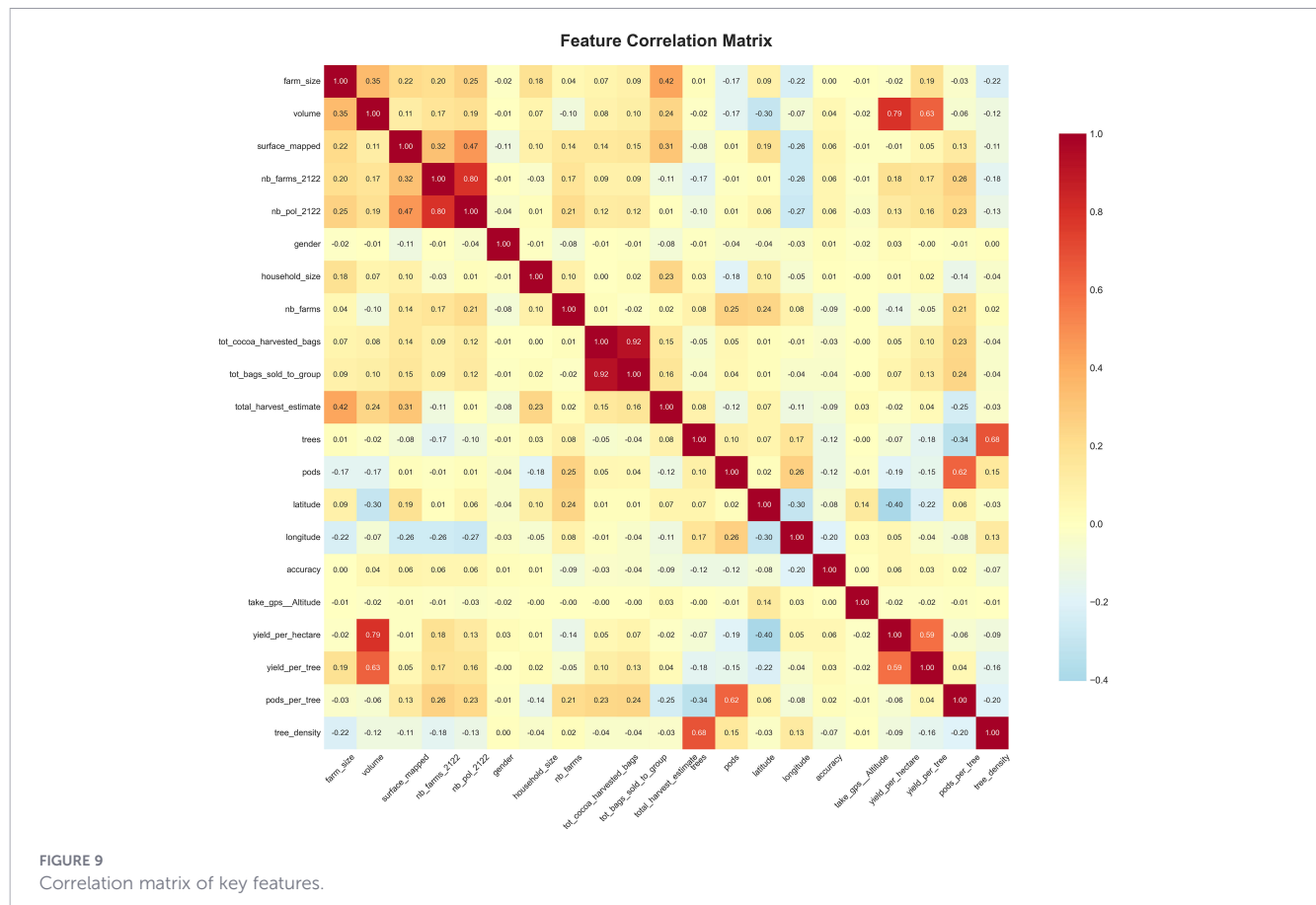
The correlation matrix (Figure 9) provides a quantitative view of how strongly pairs of variables are linearly related, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). First, the correlation between farm size and yield per hectare is near zero ($r = -0.02$), 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.

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 (Debrah and Adanu, 2022; Barrett et al., 2010; Foster and Rosenzweig, 2022), 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, which decomposes yield into its tree density and per-tree productivity components alongside farm, household, and program covariates, explains 92.07% of yield variation, with strong spatial autocorrelation ($\lambda = 0.690$) and our spatial probit model with $\rho = 0.698$, 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 tracking the same farms across multiple seasons. This would allow farm fixed effects to absorb time-invariant unobserved heterogeneity, such as soil quality or inherited land access, and would distinguish transitory zero yields, a single bad season, from structural non-production, permanent exit from cultivation, a distinction this study cannot make with one observation per farm. Repeated tree counts over time would also let future work test whether the super-elastic tree density effect reflects a genuine intensification opportunity or an artifact of comparing farms at different points in a replanting cycle. Panel data on program

enrollment, combined with quasi-experimental designs exploiting variation in rollout timing, would permit stronger causal identification of FDP and remediation effects than the observational comparisons presented here.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study was exempt from full ethical review as it is based entirely on anonymized secondary data obtained through a formal data-sharing agreement with an international cocoa sustainability program operating in Ghana's Ashanti Region. All primary data were collected by the partner organization under standard informed consent procedures before anonymization. The authors conducted no direct human subject recruitment, intervention, or experimentation. The anonymized nature of the dataset ensures that no individual farmer or household can be identified from the data or results presented in this manuscript. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

WB: Validation, Conceptualization, Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Writing – original draft, Data curation, Visualization. EO: Software, Investigation, Writing – review & editing, Project administration, Supervision. NF: Writing – review & editing, Supervision. MO: Writing – review & editing, Software, Conceptualization, Validation. EB: Investigation, Conceptualization, Writing – original draft, Software. KO: Writing – original draft, Conceptualization, Investigation, Software. AS: Writing – original draft, Investigation, Software, Conceptualization. EA: Writing – review & editing, Formal analysis, Data curation, Methodology. SA: Project administration, Supervision, Writing – review & editing. WP: Project administration, Supervision,

Writing – review & editing. HD: Formal analysis, Writing – review & editing, Data curation.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

The authors thank the cocoa sustainability program for providing access to the farm-level data and the enumerators for their diligent work during data collection.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abdulai, I., Hoffmann, M. P., Jassogne, L., Asare, R., Graefe, S., Tao, H.-H., et al. (2020). Variations in yield gaps of smallholder cocoa systems and the main determining factors along a climate gradient in Ghana. *Agric. Syst.* 181, 102812. doi: 10.1016/j.agry.2020.102812
- Ahenkorah, Y., Halm, B. J., Appiah, M. R., Akrofi, G. S., and Yirekyi, J. E. K. (1987). Twenty years' results from a shade and fertilizer trial on Amazon cocoa (*Theobroma cacao*) in Ghana. *Exp. Agric.* 23, 31–39. doi: 10.1017/S0014479700001101
- Akpoti, K., Higginbottom, T. P., Foster, T., Adhikari, R., and Zwart, S. J. (2022). Mapping land suitability for informal, small-scale irrigation development using spatial modelling and machine learning in the upper east region, Ghana. *Sci. Total Environ.* 803, 149959. doi: 10.1016/j.scitotenv.2021.149959
- Amerino, J., Apedo, C. K., and Anang, B. T. (2024). Factors influencing adoption of cocoa agroforestry in Ghana: Analysis based on tree density choice. *Sustain. Environ.* 10, 2296162. doi: 10.1080/27658511.2023.2296162
- Ameyaw, G. A., Domfeh, O., and Gyamera, E. (2023). Epidemiology and diagnostics of cacao swollen shoot disease in Ghana: Past research achievements and knowledge gaps to guide future research. *Viruses* 16, 43. doi: 10.3390/v16010043

- Anang, B. T. (2016). A probit analysis of the determinants of fertilizer adoption by cocoa farmers in Ghana. *Asian J. Agric. Extension Econ Sociology* 8, 1–8. doi: 10.9734/AJAES/2016/21540
- Aneani, F., and Ofori-Frimpong, K. (2013). An analysis of yield gap and some factors of cocoa (Theobroma cacao) yields in Ghana. *Sustain. Agric. Res.* 2, 117–127. doi: 10.5539/sar.v2n4p117
- Anselin, L. (1988). “Spatial econometrics: methods and models,” in *Studies in Operational Regional Science*, vol. 4. (Springer Netherlands, Dordrecht). doi: 10.1007/978-94-015-7799-1
- Anselin, L. (1995). Local indicators of spatial association—LISA. *Geogr. Anal.* 27, 93–115. doi: 10.1111/j.1538-4632.1995.tb00338.x
- Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Sci. Urban Econ* 26, 77–104. doi: 10.1016/0166-0462(95)02111-6
- Appiah, M. R., Ofori-Frimpong, K., and Afrifa, A. A. (2000). Evaluation of fertilizer application on some peasant cocoa farms in Ghana. *Ghana J. Agric. Sci.* 33, 183–190. doi: 10.4314/gjas.v33i2.1869
- Asante, P. A., Rahn, E., Anten, N. P. R., Zuidema, P. A., Morales, A., and Rozendaal, D. M. A. (2025). Climate change impacts on cocoa production in the major producing countries of west and central Africa by mid-century. *Agric. For. Meteorol.* 362, 110393. doi: 10.1016/j.agrformet.2025.110393
- Asante, P. A., Rahn, E., Zuidema, P. A., Rozendaal, D. M. A., van der Baan, M. E. G., Läderach, P., et al. (2022). The cocoa yield gap in Ghana: A quantification and an analysis of factors that could narrow the gap. *Agric. Syst.* 201, 103473. doi: 10.1016/j.agry.2022.103473
- Asigbaase, M., Abugre, S., Banowiiri, M., and Akutteh, J. (2025). Adoption dynamics of organic pesticides among cocoa producers in two ecological zones of Ghana. *Global Challenges* 9, e00045. doi: 10.1002/gch2.202500045
- Asitoakor, B. K., Asare, R., Ræbild, A., Ravn, H. P., Eziah, V. Y., Owusu, K., et al. (2022). Influences of climate variability on cocoa health and productivity in agroforestry systems in Ghana. *Agric. For. Meteorol.* 327, 109199. doi: 10.1016/j.agrformet.2022.109199
- Attipoe, S. G., Cao, J., Opoku-Kwanowaa, Y., and Ohene-Sefa, F. (2021). Assessing the impact of nongovernmental organization's extension programs on sustainable cocoa production and household income in Ghana. *J. Integr. Agric.* 20, 2820–2836. doi: 10.1016/S2095-3119(21)63607-9
- Barrett, C. B., Bellemare, M. F., and Hou, J. Y. (2010). Reconsidering conventional explanations of the inverse productivity–size relationship. *World Dev.* 38, 88–97. doi: 10.1016/j.worlddev.2009.06.002
- Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis With R, Second Edition* (NY: Springer).
- Bivand, R., and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *J. Stat. Software* 63, 1–36. doi: 10.18637/jss.v063.i18
- Brooks, S. P., and Andrew, G. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graphical Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787
- Bua, S., Mejahed, K., Maccarthy, D., Adogoba, D., Kissiedu, I., Atakora, W., et al. (2020). “Yield responses of maize to fertilizers in Ghana,” in *Ferari Research Report 2, International Fertilizer Development Center (Ifdc)*.
- Bunn, C., Läderach, P., Quay, A., Muilerman, S., Noponen, M. R. A., and Lundy, M. (2019). Recommendation domains to scale out climate change adaptation in cocoa production in Ghana. *Clim. Serv.* 16, 100123. doi: 10.1016/j.cliser.2019.100123
- Carletto, C., Savastano, S., and Zezza, A. (2013). Fact or artifact: The impact of measurement errors on the farm size–productivity relationship. *J. Dev. Econ* 103, 254–261. doi: 10.1016/j.jdevco.2013.03.004
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39, 829–844. doi: 10.2307/1909582
- Debrah, G., and Adanu, K. (2022). Does the inverse farm size–productivity hypothesis hold beyond five hectares? Evidence from Ghana. *J. Agric. Appl. Econ.* 54, 548–559. doi: 10.1017/aae.2022.20
- Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. 3rd edn* (Boca Raton, FL: Chapman and Hall/CRC). doi: 10.1201/b15326
- Djokoto, J. G., Owusu, V., and Awunyo-Vitor, D. (2016). Adoption of organic agriculture: Evidence from cocoa farming in Ghana. *Cogent Food. Agric.* 2, 1242181. doi: 10.1080/23311932.2016.1242181
- Doaré, F., Ribeyre, F., and Cilas, C. (2020). Genetic and environmental links between traits of cocoa beans and pods clarify the phenotyping processes to be implemented. *Sci. Rep.* 10, 9888. doi: 10.1038/s41598-020-66969-9
- Donkor, E., Amegbe, E. D., Ratering, T., and Hejkrlik, J. (2023). The effect of producer groups on the productivity and technical efficiency of smallholder cocoa farmers in Ghana. *PLoS One* 18, e0294716. doi: 10.1371/journal.pone.0294716
- Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *J. Am. Stat. Assoc.* 78, 605–610. doi: 10.1080/01621459.1983.10478017
- Forkuor, G., Pavelic, P., Asare, E., and Obuobie, E. (2013). Modelling potential areas of groundwater development for agriculture in northern Ghana using GIS/RS. *Hydrol. Sci. J.* 58, 437–451. doi: 10.1080/02626667.2012.754101
- Foster, A. D., and Rosenzweig, M. R. (2022). Are there too many farms in the world? Labor market transaction costs, machine capacities, and optimal farm size. *J. Political Economy* 130, 636–680. doi: 10.1086/717890
- Gelman, A., and Rubin, B. D. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136
- Ghana Cocoa Board (COCOBOD) (2024a). Increasing cocoa production in Ghana—the importance of the 4ps *Cocobod News*.
- Ghana Cocoa Board (COCOBOD) (2024b). Review of the producer price of cocoa for the 2024/2025 cocoa season *Cocobod News*.
- Kabaya, K. (2022). Examining spatially heterogeneous preferences for coastal ecosystem restoration with Bayesian spatial probit approaches. *Lett. Spatial Resour. Sci.* 15, 99–112. doi: 10.1007/s12076-022-00301-z
- Kilic, T., Palacios-López, A., and Goldstein, M. (2015). Caught in a productivity trap: A distributional perspective on gender differences in Malawian agriculture. *World Dev.* 70, 416–463. doi: 10.1016/j.worlddev.2014.06.017
- Kondo, E., Asem, F. E., Osei-Asare, Y., and Mensah-Bonsu, A. (2026). Spatial dependence and spillover effects in crop diversification practices: Farm household level evidence from Ghana. *Front. Sustain. Food Syst.* 10, 1759867. doi: 10.3389/fsufs.2026.1759867
- Kongor, J. E., De Stur, H., Van De Walle, D., et al. (2018). Constraints for future cocoa production in Ghana. *Agrofor. Syst.* 92, 1373–1385. doi: 10.1007/s10457-017-0082-9
- Läderach, P., Martínez-Valle, A., Schroth, G., and Castro, N. (2013). Predicting the future climatic suitability for cocoa farming of the world's leading producer countries, Ghana and Côte d'Ivoire. *Clim. Change* 119, 841–854. doi: 10.1007/s10584-013-0774-8
- LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geogr. Anal.* 32, 19–35. doi: 10.1111/j.1538-4632.2000.tb00413.x
- LeSage, J. P., and Pace, R. K. (2009). *Introduction to Spatial Econometrics* (Boca Raton, FL: Chapman and Hall/CRC). doi: 10.1201/9781420064254
- Loomis, J. B., and Mueller, J. M. (2013). A spatial probit modeling approach to account for spatial spillover effects in dichotomous choice contingent valuation surveys. *J. Agric. Appl. Econ.* 45, 53–63. doi: 10.1017/S1074070800004570
- Manning, W. G., and Mullahy, J. (2001). Estimating log models: To transform or not to transform? *J. Health Econ* 20, 461–488. doi: 10.1016/S0167-6296(01)00086-8
- Ministry of Finance, Ghana (2025). *Review of producer price for the 2025/26 cocoa season* (Ministry of Finance Ghana).
- Naazie, G. K., Agyemang, I., and Tampah-Naah, A. M. (2024). Spatial pattern and distribution of urban farms and climate change adaptation implications in urban Ghana. *Cogent Food. Agric.* 10, 2320210. doi: 10.1080/23311932.2024.2320210
- Nkamleu, G. B., Nyemack, J., and Gockowski, J. (2010). “Technology gap and efficiency in cocoa production in west and central Africa: implications for cocoa sector development,” in *Working Paper Series 241, African Development Bank*.
- Ofori, A., Padi, F. K., Ameyaw, G. A., Dadzie, A. M., Opoku-Agyeman, M., Domfeh, O., et al. (2022). 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, e0262461. doi: 10.1371/journal.pone.0262461
- Osei-Gyabaah, A. P., Antwi, M., Addo, S., and Osei, P. (2023). Land suitability analysis for cocoa (Theobroma cacao) production in the sunyani municipality, bono region, Ghana. *Smart Agric. Technol.* 5, 100262. doi: 10.1016/j.atech.2023.100262
- Osman, N., Afele Tennyson, J., Nimo, E., Gorleku Ofoe, D., Ofori Adomaa, L., and Abunyewa Adutwum, A. (2022). Assessing the impact of illegal small-scale mining (galamsey) on cocoa farming and farmer livelihood: A case study in the amansie west district of Ghana. *Pelita Perkebunan (A Coffee Cocoa Res. Journal)* 38, 70–82. doi: 10.22302/iccri.jur.pelita-perkebunan.v38i1.496
- Owiredo, P., Wongnaa, C. A., Acheampong, P. P., Addison, M., Adu, K. A., and Awunyo-Vitor, D. (2022). Farmer business school participation and its impact on cocoa productivity and food security in Ghana. *J. Agribusiness Developing Emerging Economies* 14, 637–654. doi: 10.1108/JADEE-05-2022-0102
- Parra-Paitan, C., Meyfroidt, P., Verburg, P. H., and Zu Ermgassen, E. K. H. J. (2024). Deforestation and climate risk hotspots in the global cocoa value chain. *Environ. Sci. Policy* 158, 103796. doi: 10.1016/j.envsci.2024.103796
- Pinkse, J., and Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete choice models. *J. Econometrics* 85, 125–154. doi: 10.1016/S0304-4076(97)00097-3
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News* 6, 7–11.
- Quaye, A. K., Doe, E. K., Attua, E. M., et al. (2021). Geospatial distribution of soil organic carbon and soil pH within the cocoa agroecological zones of Ghana. *Geoderma* 386, 114921. doi: 10.1016/j.geoderma.2020.114921

- Ripley, B. D. (1977). Modelling spatial patterns. *J. R. Stat. Soc. Ser. B. Stat. Method.* 39, 172–192. doi: 10.1111/j.2517-6161.1977.tb01615.x
- Ruf, F., and Schroth, G. (2004). “Chocolate forests and monocultures: An historical review of cocoa growing and its conflicting role in tropical deforestation and forest conservation,” in *Agroforestry and Biodiversity Conservation in Tropical Landscapes*, 107–134.
- Schroth, G., Läderach, P., Martinez-Valle, A. I., Bunn, C., and Jassogne, L. (2016). Vulnerability to climate change of cocoa in west Africa: Patterns, opportunities and limits to adaptation. *Sci. Total Environ.* 556, 231–241. doi: 10.1016/j.scitotenv.2016.03.024
- Smith, T. E., and LeSage, J. P. (2004). “A Bayesian probit model with spatial dependencies,” in *Spatial and Spatiotemporal Econometrics, Advances in Econometrics, Vol. 18* (Emerald Group Publishing Limited), 127–160. doi: 10.1016/S0731-9053(04)18004-3
- Sunge, R., and Espoir, D. K. (2026). Analyzing the environmental impact of agricultural production in africa through a comprehensive spatial econometric framework. *Environ. Impact Assess. Rev.* 117, 108183. doi: 10.1016/j.eiar.2025.108183
- Sussy, M., Hall, O., Francisca, A. B. M., Boke-Olén, N., Onyango, M. C., Oluoch-Kosura, W., et al. (2019). Micro-spatial analysis of maize yield gap variability and production factors on smallholder farms. *Agriculture* 9, 219. doi: 10.3390/agriculture9100219
- United States Department of Agriculture Foreign Agricultural Service (2025). *Ghana—Cocoa Sector Overview 2025* (Washington, DC: USDA). GAIN Report GH2025-0008.
- Wessel, M., and Quist-Wessel, P. M. F. (2015). Cocoa production in west Africa, a review and analysis of recent developments. *NJAS Wageningen J. Life Sci.* 74–75, 1–7. doi: 10.1016/j.njas.2015.09.001
- Wollburg, P., Bentze, T., Lu, Y., Udry, C., and Gollin, D. (2024). Crop yields fail to rise in smallholder farming systems in sub-Saharan Africa. *Proc. Natl. Acad. Sci.* 121, e2312519121. doi: 10.1073/pnas.2312519121
- Wongnaa, C. A., and Babu, S. (2020). Building resilience to shocks of climate change in Ghana's cocoa production and its effect on productivity and incomes. *Technol. Soc.* 62, 101288. doi: 10.1016/j.techsoc.2020.101288
- World Bank (2022). Ghana tree crop diversification project (P180060) *World Bank Documents*.