Methodology to identify spatial patterns in coffee (*Coffea arabica* L.) production

Metodología para identificar patrones espaciales en la producción de café (*Coffea arabica* L.)

Camilo Alberto Calle Velásquez[1*], Iván Darío Aristizábal Torres[2], Francisco Javier Rodríguez Cortés[3], and Edilson León Moreno Cárdenas[2]

ABSTRACT

Coffee farming, a lifeline for numerous families in the mountainous regions of Latin America, faces challenges due to climate change and production variability, which complicate the use of forecast models at the territorial level. In response to these challenges, territorial inference has gained relevance, especially with the advancement of Geographic Information Systems (GIS), which provide useful tools for territorial analysis. Although spatial models are increasingly applied in GIS, coffee farming, like many agricultural subsectors, is hindered by a lack of information and spatial methodologies. This work proposes a methodology to identify spatial patterns of homogeneous production areas. Data from 140 farms, representing 3,900 members of the coffee grower cooperative of Andes, dispersed over 200,000 ha, were analyzed between 2019 and 2021. The variables used to measure productivity included the number of fruits per tree, the average fruit weight, planting density, and the conversion rate of cherry coffee to dry parchment coffee. A simple linear regression model was employed, and spatial dependency analyses were performed using the global and local Moran's Index to identify clusters of territorial subdivisions. The data were processed in R language, and the GeoDa™ program was used to obtain the spatial weight matrix. Territorial units with similar characteristics for high-quality mountain coffee production were identified through spatial dependency indicators. The methodology can contribute to estimating coffee production in large territories, improving the reliability of information and allowing for more informed decision-making to optimize coffee farming in mountainous areas.

Key words: mountain coffee production, clustering, Moran's index, spatial dependency, territorial planification.

RESUMEN

La caficultura, una fuente de sustento para numerosas familias en las regiones montañosas de América Latina, enfrenta desafíos debido al cambio climático y la variabilidad de la producción, lo que complica el uso de modelos de pronóstico a nivel territorial. A pesar de estos desafíos, la inferencia territorial ha ganado relevancia, especialmente con el avance de los Sistemas de Información Geográfica (SIG), que ofrecen herramientas útiles para el análisis territorial. Aunque los modelos espaciales se aplican cada vez más en SIG, la caficultura, al igual que muchos subsectores agrícolas, se ve limitada por la falta de información y metodologías espaciales. Este trabajo propone una metodología para identificar patrones espaciales de áreas de productividad homogénea. Se analizaron datos de 140 fincas, representando a 3900 miembros de la cooperativa de caficultores de Andes, distribuidos en 200.000 ha, entre 2019 y 2021. Las variables utilizadas para medir la productividad incluyeron el número de frutos por árbol, el peso promedio de los frutos, la densidad de plantación y la tasa de conversión de café "cereza" a café pergamino seco. Se empleó un modelo de regresión lineal simple y se realizaron análisis de dependencia espacial utilizando el Índice de Moran global y local para identificar agrupamientos de subdivisiones territoriales. Los datos se procesaron en el lenguaje R, y se utilizó el programa GeoDa™ para obtener la matriz de pesos espaciales. Mediante indicadores de dependencia espacial, se identificaron unidades territoriales con características similares para la producción de café en zonas montañosas. La metodología puede contribuir a estimar la producción de café en grandes territorios, mejorando la confiabilidad de la información y permitiendo una toma de decisiones más informada para optimizar la caficultura en áreas montañosas.

Palabras clave: producción de café de montaña, agrupamiento, índice de Moran, dependencia espacial, planificación territorial.

Introduction

The coffee industry plays a vital role in the economies of Colombia and Latin America, supporting millions of livelihoods, particularly in mountainous regions. However, coffee farming faces increasing challenges due to climate change and production variability, which complicate the effectiveness of traditional forecast models at the territorial

Received for publication: November 6, 2024. Accepted for publication: December 28, 2024. Doi: 10.15446/agron.colomb.v42n3.117455

[1] Universidad Nacional de Colombia, Facultad de Ciencias Agrarias, Medellín (Colombia).
[2] Universidad Nacional de Colombia, Facultad de Ciencias Agrarias, Departamento de Ingeniería Agrícola y Alimentos, Medellín (Colombia).
[3] Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Medellín (Colombia).
[*] Corresponding author: ccallev@unal.edu.co

level. These challenges highlight the need for more precise and adaptable methodologies to estimate coffee production, ensuring better resource allocation and supply chain stability (Mendoza *et al.,* 2013).

Estimating coffee production is essential for strategic planning across the supply chain, particularly for cooperatives and stakeholders involved in futures contracts, who rely on accurate yield predictions to mitigate financial risks (Choperena Bedoya & Couleau, 2021). Traditionally, coffee yield estimations have depended on economic and climatic variables. However, these factors alone provide limited precision in addressing the complexities of coffee production, especially in diverse topographies and heterogeneous farming conditions (Aparecido *et al*., 2017; Gil Serna, 2012; Moraes-Oliveira *et al*., 2017). Since the 1950s, forecasting models have been developed to assist in decision-making for fertilization planning, labor scheduling, and input purchases, helping producers optimize resource allocation. Over time, more than 50 factors influencing coffee production have been identified, spanning soil properties, climate conditions, social dynamics, agronomic practices, and crop management strategies (Montoya Restrepo *et al*., 2009).

Despite these advancements, forecasting models still face significant limitations. The high variability in phenological characteristics across coffee varieties (Ramírez Builes, 2014) and the reliance on flowering stages for prediction have yielded modest results, with determination coefficients around 0.4 (Rendón-Sáenz *et al*., 2008). Additionally, many models developed from experimental farm data lack field validation, reducing their applicability to traditional farms, which often operate under different environmental and management conditions.

To address these limitations, territorial inference has emerged as a promising approach for improving yield estimation. The increasing adoption of Geographic Information Systems (GIS) has facilitated more comprehensive spatial analyses, enabling the identification of productivity patterns over large areas. However, coffee farming, like many other agricultural subsectors, still faces a lack of spatial methodologies for production estimation.

To improve the accuracy of yield estimates, recent studies have explored artificial vision techniques for fruit counting on branches, achieving high precision (R^2>0.9) (Ramos *et al*., 2017). However, this method faces practical challenges in the field, as variations in the number of axes, branches, and branch types affect implementation. Artificial vision has also been applied to satellite and drone imagery to assess factors like plant density, leaf health, and coffee fruit ("cherry") development (Abreu Júnior *et al*., 2022; Martello *et al*., 2022; Tanaka *et al*., 2015). This approach is particularly promising in the context of climate change, where unpredictable conditions limit the reliability of traditional models. Although high costs are a limitation (Benos *et al*., 2021; Chlingaryan *et al*., 2018; Kouadio *et al*., 2018), artificial vision remains valuable for tasks such as fruit counting, an essential parameter in assessing coffee productivity (Adane & Bewket, 2021; Beksisa *et al*., 2018).

Advances in artificial vision continue to support the development of forecasting models adapted to image data (Cheng *et al*., 2017; Eugenio *et al*., 2020; Khaki & Wang, 2019; Qiao *et al*., 2021; Tsai & Chen, 2017; Wang *et al*., 2017; Zhang *et al*., 2019). Furthermore, spatial analysis techniques have introduced new approaches to territorial-level modeling, enabling the identification of homogeneous areas, which can improve resource management in agricultural landscapes. In a context of resource scarcity, spatial technologies allow for optimized land, water, and input use, promoting sustainability and efficiency (Anselin *et al*., 2004; Bivand *et al*., 2013). This integration of spatial techniques enables informed decision-making, improves productivity, and mitigates environmental impacts, addressing the pressing challenges of food security and sustainable production.

This research proposes a novel methodology to identify homogeneous productivity zones by analyzing data from 140 farms representing 3,900 cooperative members, covering over 200,000 ha in the coffee-growing region of the Andes in Colombia.

The proposed approach integrates spatial dependency analyses using Moran's Index to identify clusters of productivity within territorial subdivisions. A simple linear regression model is employed to analyze key productivity indicators, including fruit count per tree, average fruit weight, planting density, and cherry-to-parchment conversion rate. The data are processed in the R language, and the spatial weight matrix is generated using GeoDa™, ensuring robust geostatistical analysis.

This study contributes to the ongoing efforts to improve coffee production estimation at the territorial level, offering a methodology that enhances information reliability, decision-making processes, and strategic planning. By leveraging spatial statistics and GIS, this approach provides a scalable framework for identifying high-quality coffee production areas and ultimately optimizing coffee farming in mountainous landscapes.

This study proposes a methodology for identifying spatial patterns in coffee production using historical data (2019–2021) from selected farms in Southwest Antioquia, Colombia. The research focuses on analyzing variables related to production, along with images of all harvested fruits per tree. The primary goal is to identify specific areas within the region where productivity can be predicted based on the number of coffee fruits per tree. Additionally, the study examines the relationship between fruit count and the average weight of a sample of 30 fruits from the same tree, establishing a foundation for refined production forecasts at the local level.

Materials and methods

This study develops a methodology to identify homogeneous areas within rural zones. In Colombia, the smallest territorial unit is called a "vereda" (small village), which this study will refer to as the Minimum territorial subdivision (MTS). The study associates certain relevant variables to this MTS to predict coffee production. A simple forecasting model was implemented to estimate the number of fruits per tree at specific times, using three years of harvest data from the Southwestern Antioquia subregion, one of Colombia's most productive coffee areas.

Localization

The study area included the municipalities of Andes, Ciudad Bolívar, Betania, Hispania, and Jardín, all located in the Southwestern Antioquia subregion, with a potential of 3,900 producers (Fig. 1).

Sampling and data collection

Farms were selected from the Cooperandes Producers Association database, with sampling conducted from January to March for secondary harvests and from August to September for main harvests in 2019, 2020, and 2021. Farms were chosen randomly with stratified sampling by agroecological zone, and data collection on each farm included:

- Producer contact: Farms with at least three productive lots of varying ages and with no ongoing fumigations were selected;
- Lot selection: Three diverse lots per farm were prioritized, excluding those near renovation areas or with harvest issues;
- Tree selection: From each lot, three trees representing different productivity levels were chosen, avoiding those near paths or water sources;
- Fruit harvesting: All fruits from selected trees were labeled with producer and farm information;
- Sample imaging: Field personnel photographed fruit samples with geolocated smartphones;
- Fruit counting: Fruit images were analyzed using the CounThings© app (Countthings, 2021) with a mung bean template for better performance. The count data were stored in a CSV file;
- Production weighing: The total harvest per tree was recorded in grams as this is how data are stored in the cooperative database;
- Individual weight measurement: Thirty fruits per tree were individually weighed with a precision balance (model LS220A) at the cooperative's lab.

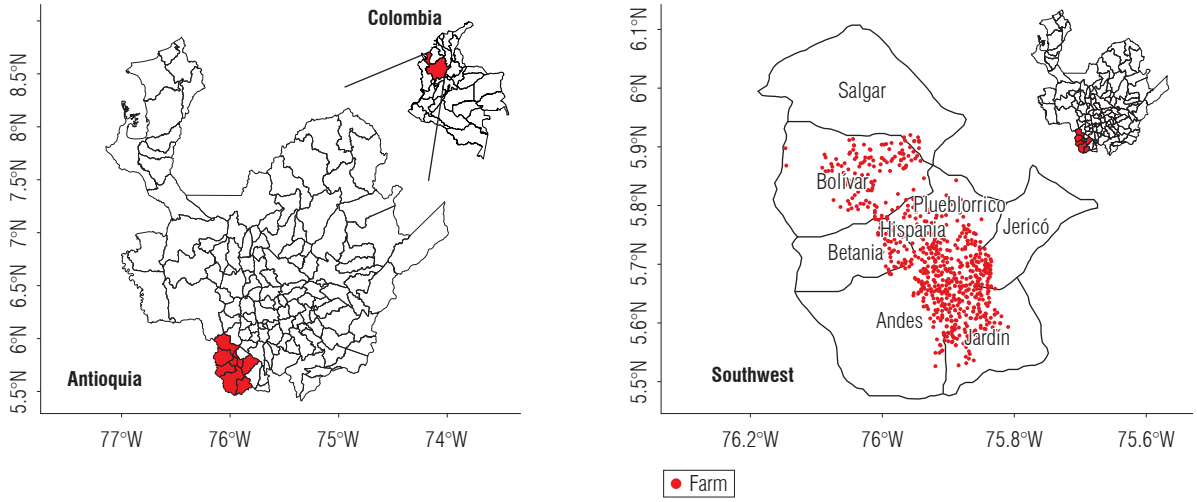

FIGURE 1. Distribution and location of the universe of producers within the territory: A) area of influence, B) territorial distribution of the farms.

Formation and selection of the dataset

Cooperandes provided 2019–2021 production data, including variables such as average fruit weight per tree, fruit count, planting density, and conversion rates for cherry coffee to dry parchment coffee, along with identifiers for lot, farm, and municipality. After data cleaning, 512 valid records from 161 farms remained. Using the National Administrative Department of Statistics (DANE) database, polygons of territorial subdivisions (MTS) were mapped for the municipalities of Andes, Ciudad Bolívar, Betania, Hispania, and Jardín. Farm coordinates were aligned with official maps from the Geographic Institute Agustín Codazzi (IGAC, 2021), resulting in 140 georeferenced farms. Each farm was assigned an MTS identifier, and the mean of each variable was calculated per MTS. Data analysis was performed in R (Bivand *et al*., 2013).

Proposed production model

A linear model (Eq. 1) obtained by regression was used to estimate the number of fruits per tree. It was determined that a sample of 30 fruits could represent the total population of fruits on each tree, regardless of the variability in axes, branches, or nodes that the trees may have.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

where Y_i is the number of fruits per tree, β_0 is the interception, β_1 is the slope, X_i is the average weight of 30 coffee fruits (g), and ε_i the error term (Montgomery & Peck, 2006).

An analysis of the proposed model was conducted in the municipalities where the samples were obtained, seeking to determine if the detected areas had better determination coefficients than the models at the municipal level. Table 1 indicates the parameters and statistical significance of the proposed model at the municipal scale, showing that only the municipalities of Andes, Betania, and Ciudad Bolívar had significant models (P<0.1). However, these models explain little variability of the phenomenon (adjusted R^2 between 0.14 and 0.3), and they do not adapt to the higher (Jardín) or lower (Hispania) zones of the territory, making their practical application challenging in conditions of high climatic variability. The normality assumptions were verified using the Shapiro-Wilk test to analyze the models.

This analysis aims to identify spatial autocorrelation in a dataset, revealing patterns of clustering or geographic dispersion among its values. Spatial autocorrelation suggests that the values of the variables have some type of spatial association, meaning that nearby observations are more likely to be associated compared to observations farther apart for the same variable. The Moran's index models this relationship, allowing the identification of the magnitude of the degree of spatial autocorrelation between the areas that define the study region. Through standard statistical tests, a statistically significant level for spatial correlation was calculated, which allows the determination of specific areas within the territory (Moran, 1950).

This is achieved by constructing a matrix of neighborhood or spatial weights matrix, for which missing data must first be imputed. For this study, the nearest neighbor method was used, which consisted of interpolation based on the arithmetic mean of the neighbors, specifically the closest or first-order neighbors. In this process, the following software programs were used for data processing, cleaning, and visualization: The R language (R Core Team, 2020) and the open-access GeoDa™ software (GeoDa Foundation, 2020).

The calculation of Moran's Index, or Moran's spatial autocorrelation coefficient (Eq. 2), was based on the correlation between a variable's values and the values of its neighboring observations (Moran, 1950).

$$MI = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(z_i-\bar{z})(z_j-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2} \quad (2)$$

where MI is Moran's Index, wij is the Spatial Weights Matrix (Neighborhood Matrix), zi and zj are the observed values of the variable in areas i and j, $\bar{z}$ is the expected value of the variable, and n is the number of areas (MTSs).

With the Moran test, the observed Moran's Index was compared against the expected distribution under the

TABLE 1. Analysis of models at the municipal level.

Municipality	Multiple correlation coefficient	Determination coefficient R^2	R^2 adjusted	Typical error	Interception	Slope	F	Critical value of F
Andes	0.37	0.14	0.12	801.95	6,041.67	- 3,503.38	9.63	0.00
Jardín	0.14	0.02	-0.04	925.44	3,747.28	- 1,465.72	0.35	0.56
Hispania	0.53	0.28	0.20	783.76	5,118.43	- 2,041.45	3.43	0.10
Betania	0.35	0.12	0.09	1,175.8	5,762.76	- 2,713.96	3.68	0.07
Ciudad Bolívar	0.55	0.30	0.26	908.71	8,832.80	- 5,796.82	6.84	0.02

null hypothesis of no spatial autocorrelation. When the observed Moran's coefficient is significantly different from zero, the null hypothesis is rejected, indicating that spatial autocorrelation exists in the data (Moran, 1950).

LISA (Local Indicators of Spatial Association) is a dimensionless index representing the number of standard deviations by which an area deviates from the mean of its surrounding.

The calculation of LISAs, like Moran's Index, represented a weighted measure of correlations based on a neighborhood criterion (Anselin, 1995), for which the following equation was used:

$$MI = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(z_i-\bar{z})(z_j-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2} \quad (3)$$

where IMi is the Moran Index of area i, Ji is the neighbors set of area i, Zj is the mean of observations of areas neighboring area i, n is the number of areas in area J (surrounding MTSs), and W_{ij} is the matrix of spatial weights for group J of rural MTSs (neighborhood matrix).

The result of the local spatial dependence test was depicted in a map of local spatial significance to visualize and understand the spatial structure of the data. In the graphs, the values of the variables – fruit weight, total fruits, planting density, and conversion – were placed on the x-axis for each geographical location, and the calculated values of spatial autocorrelation of the data at each location (or IMi) were positioned on the y-axis. Using these graphs, each MTS was classified according to its autocorrelation pattern, which, for this study, was conducted using LISAs. The classification included categories such as high-high (high spatial concentration of high values), low-high (low spatial concentration of high values), high-low (high spatial concentration of low values), and low-low (low spatial concentration of low values). Other authors have employed this classification for spatial analysis of climatic variables (temperature, rainfall, and humidity) in Colombia (de Corso Sicilia *et al*., 2017).

The local spatial dependence test was used to detect specific areas with clustering or association patterns among the same variable within neighboring areas. Additionally, to test the hypothesis that these spatial autocorrelation relationships may indicate other relationships among those variables, a model was proposed to explain the number of fruits per coffee tree. The coefficient of determination of this model for the identified MTS groups was proposed as the level of predictability represented by this model.

In the study, it was established that an R^2 greater than 0.7 indicated forecastable areas. Areas with R^2 values between 0.5 and 0.7 were considered modelable, while those with R^2 values below 0.5 or without local or global spatial dependence were considered undetected areas.

The research hypothesis proposed that at least one MTS group exhibits both global and local spatial dependence for the variables Fruit weight and total fruits per tree, and that these variables are significantly correlated and can form a regression model (Eq. 1).

Results

Exploratory spatial analysis of the variables

The following table shows the descriptive results of the variables by municipality.

Figure 2 illustrates the frequency distribution of the variables of fruit weight, total fruits per tree, cherry coffee to dry parchment conversion, and planting density. The Shapiro-Wilk test was applied to these variables, showing the W values and their respective "*P*" values for each. For the variable fruits per tree, the W value is relatively high, suggesting a slight deviation from normality (*P*-value = 0.06); however, upon examining the distribution graph, its tendency towards a normal distribution is evident, similar to the other variables.

TABLE 2. Overview of the variables used in the study.

Municipality	Average fruit weight (g)		Fruits per tree (quantity)		Conversion (kg cherry coffee/kg dry parchment coffee)		Planting density (trees ha^{-1})	
	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation
Andes	1.03	0.17	2,579	1,415	5.36	0.62	6,036	1,390
Betania	1.04	0.21	2,829	1,264	5.19	0.81	6,619	1,454
Ciudad Bolívar	1.04	0.17	2,906	1,380	5.28	0.52	5,571	995
Hispania	0.98	0.18	3,131	1,093	5.41	0.75	6,838	1,808
Jardín	1.18	0.29	2,186	1,408	5.23	0.65	5,581	1,036

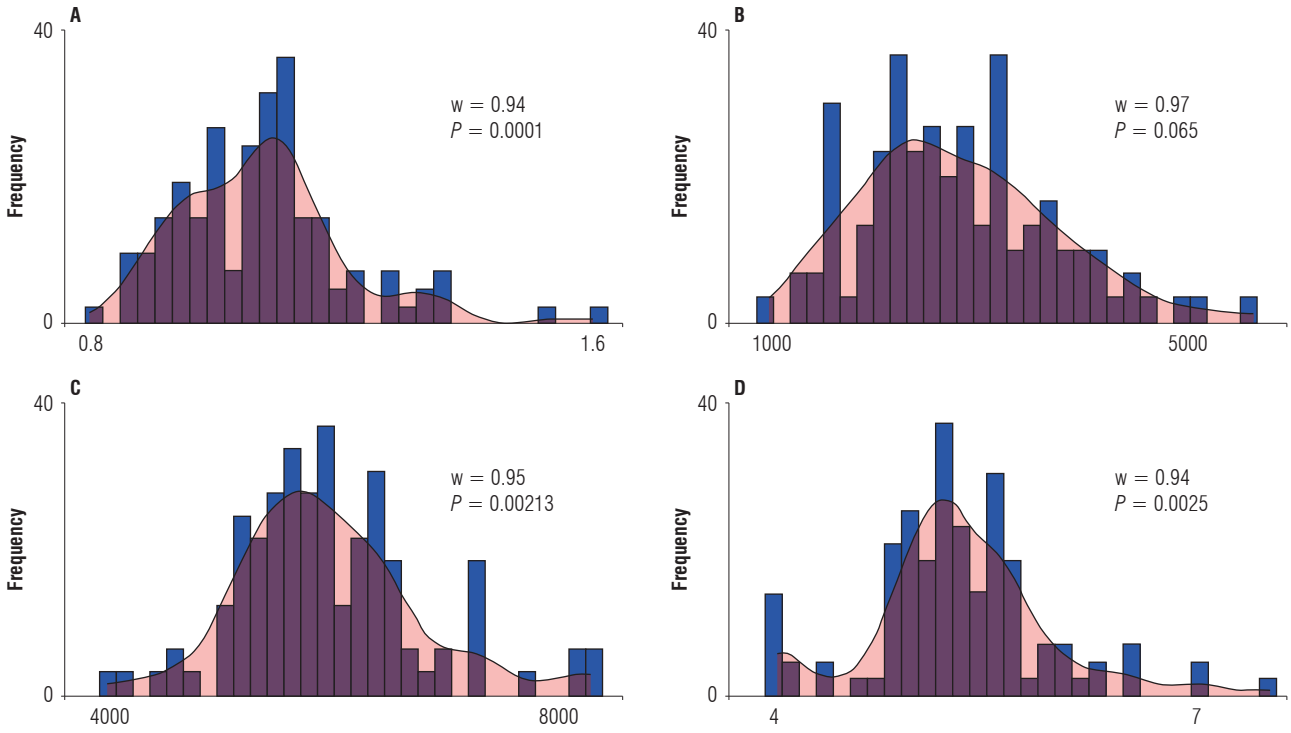

FIGURE 2. Frequency distribution of the variables used in the study: A) fruit weight, B) number of fruits per tree, C) planting density, D) conversion rate of cherry coffee to dry parchment coffee.

Some variables may cluster according to their values; however, this visual analysis is subjective and requires more robust techniques, such as LISAs.

Table 3 presents the calculated correlation between the variables. It is important to note that a negative correlation (-0.537) was detected between conversion and the number of fruits per tree. A negative correlation indicates that more efficient systems (lower conversions, *i.e.*, more technified) have a greater quantity of fruits. Despite its biological logic, this relationship also suggests the need for a different model specification. Therefore, in this study, the conversion variable was not used to refer to the harvest forecast, given the initially proposed model (Eq. 1).

TABLE 3. Correlation between variables.

Variable	Conversion	Planting density	Weight per fruit	Fruits per tree
Conversion	X	-0.085	0.277	-0.537
Planting density		X	-0.045	0.150
Weight per fruit			X	0.386
Fruits per tree				X

The neighborhood criterion used was the queen and first-order, employing the GeoDa™ program (Chasco Yrigoyen, 2006). To create the connectivity map, the average of the surrounding areas was assigned to the zones with missing data to avoid disconnected areas. The MTSs with missing data and the connectivity map are shown in Figure 4.

Spatial dependence indices

Table 4 presents the Moran's I values and their global test for each of the studied variables. The spatial dependence index is relatively low, despite all variables showing global spatial dependence. The variable with the highest spatial dependence was Fruit Weight. The global index indicated that at least one pair of MTSs exhibited spatial dependence. Subsequently, local spatial dependence was calculated to detect homogeneous areas or groups of MTSs.

TABLE 4. Global spatial dependence tests based on the Moran's Index (MI).

Item	MI	Value of *P*	Statistical significance
Planting density	0.3276	1.30E-10	***
Conversion	0.2183	9.60E-06	***
Fruits per tree	0.1760	0.0002849	***
Weight per fruit	0.3341	5.32E-11	***

*** highly significant.

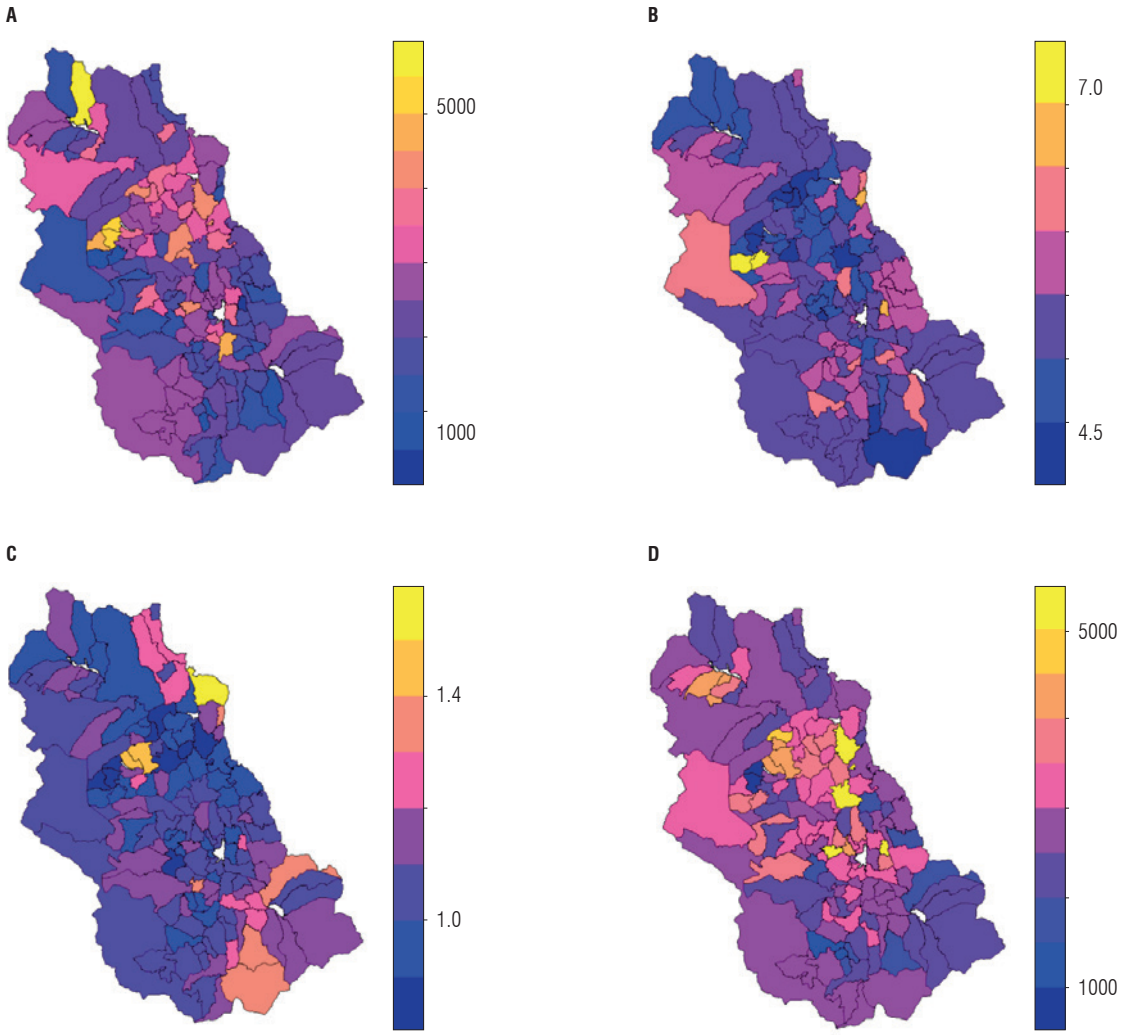

FIGURE 3. Spatial distribution of the study variables: A) number of fruits per tree, B) conversion rate of cherry coffee to dry parchment coffee, C) fruit weight, D) planting density.

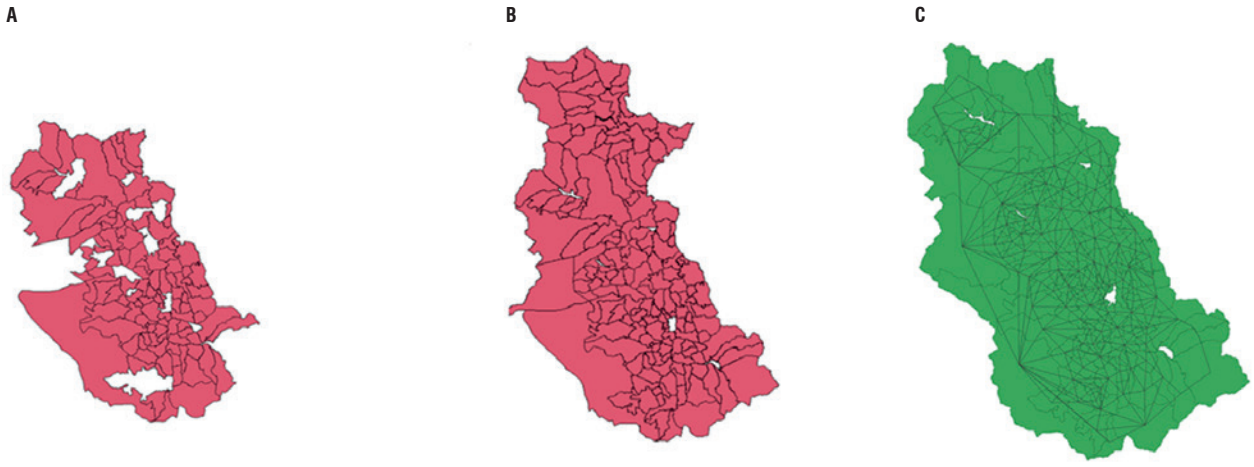

FIGURE 4. Spatial data and connectivity matrix: A) missing data, B) compete data, C) connectivity map.

Local spatial dependence index (IMi)

Figure 5 shows the estimated values of the LISAs, identifying areas with homogeneous characteristics.

Using local spatial autocorrelation analysis (LISA), MTSs presenting local dependence for the analyzed variables were identified, resulting in a total of 7 MTSs with significant spatial dependence for fruit weight and fruits per tree. Table 5 presents the LISA values for the identified MTSs. Notably, out of these 7 MTSs, two also exhibited local dependence for density and conversion variables. This suggests that in these areas, all variables considered in this study could be

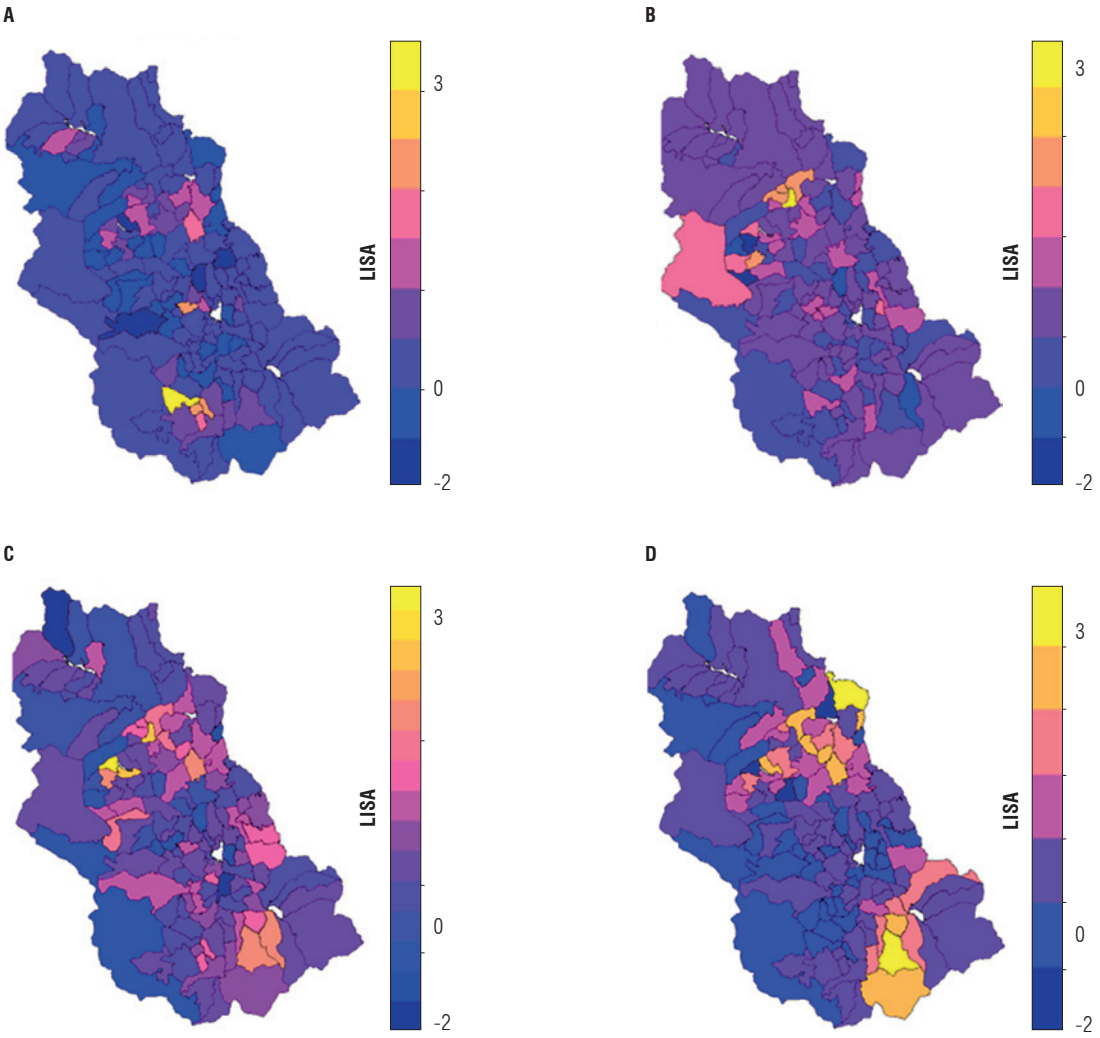

FIGURE 5. Local Moran indices: A) planting density, B) conversion rate of cherry coffee to dry parchment coffee, C) number of fruits per tree, D) fruit weight.

TABLE 5. Local evidence of spatial dependence of the selected areas.

Municipality	MTSs	*P* value Fruits per tree	*P* value Weight of fruit	*P* value Planting density	*P* value Conversion
Jardín	Macanas	0.02*	0.03*	0.02*	0.73
Jardín	Gibraltar	0.04*	0.00*	0.24	0.14
Jardín	Verdún	0.05*	0.00*	0.14	0.53
Hispania	La Armenia	0.02*	0.00*	0.05*	0.03*
Hispania	La Seca	0.03*	0.00*	0.01*	0.20
Betania	Las Animas	0.00*	0.00*	0.00*	0.04*
Betania	El Tablazo	0.03*	0.01*	0.20	0.00*

*$P<0.05$ according to IMi test. MTS – Minimum territorial subdivision.

used in a forecast model for total parchment coffee for this smaller zone. However, this study focused on the analysis of fruit-related variables and total fruits, suggesting a forecast that could extend to cherry coffee.

The degree of spatial dependence between space and the analyzed variables can be observed in the Moran plots (Fig. 6). The Moran plots show that the variables "fruit weight" and "planting density" exhibit higher spatial dependence, while " number of fruits per tree" and "conversion" show lower spatial dependence. Additionally, it is important to highlight that all variables present positive spatial dependence.

Analysis of the detected areas

The results of the linear regression models and their statistical significance for the group of detected MTSs, the group of detected MTSs plus their surroundings, and the group of all MTSs are shown in Table 6. It was demonstrated that, in the detected areas, a production model that explains 79% of the number of fruits per tree from their mean weight performs better, and that the detected area and its surroundings explain 62% of the variability of the phenomenon.

In Figure 7A, the five municipalities included in the study are shown. In Figure 7B, the non-detected areas are shown

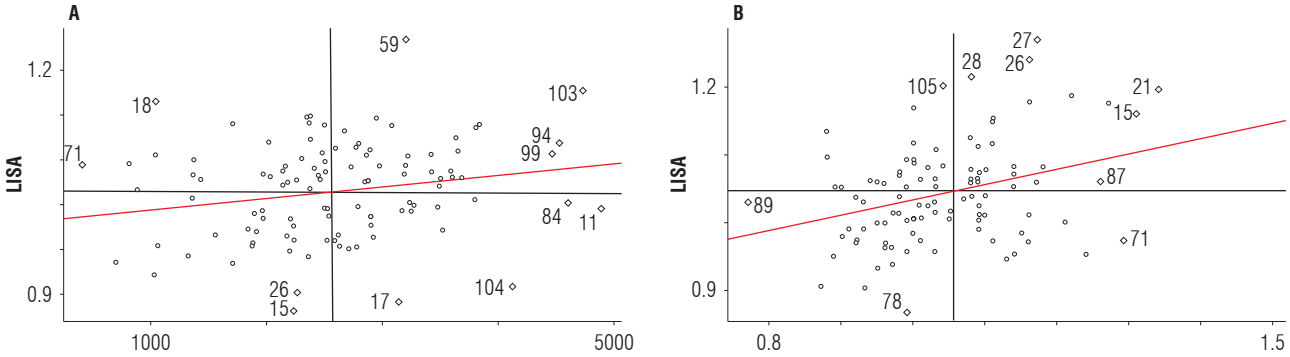

FIGURE 6. Moran graphics: A) number of fruits per tree, B) fruit weight.

TABLE 6. Parameters of the production models according to the selected MTSs.

Model	Slope	Interception	MTSs	Total area (ha)	R^2
Detected area	-4,423	7,376	7	6,482	0.79
Detected area and surrounding	-5,372	8,344	30	20,361	0.62
Study zone	-3,180	5,906	140	109,626	0.15

MTS – Minimum territorial subdivision.

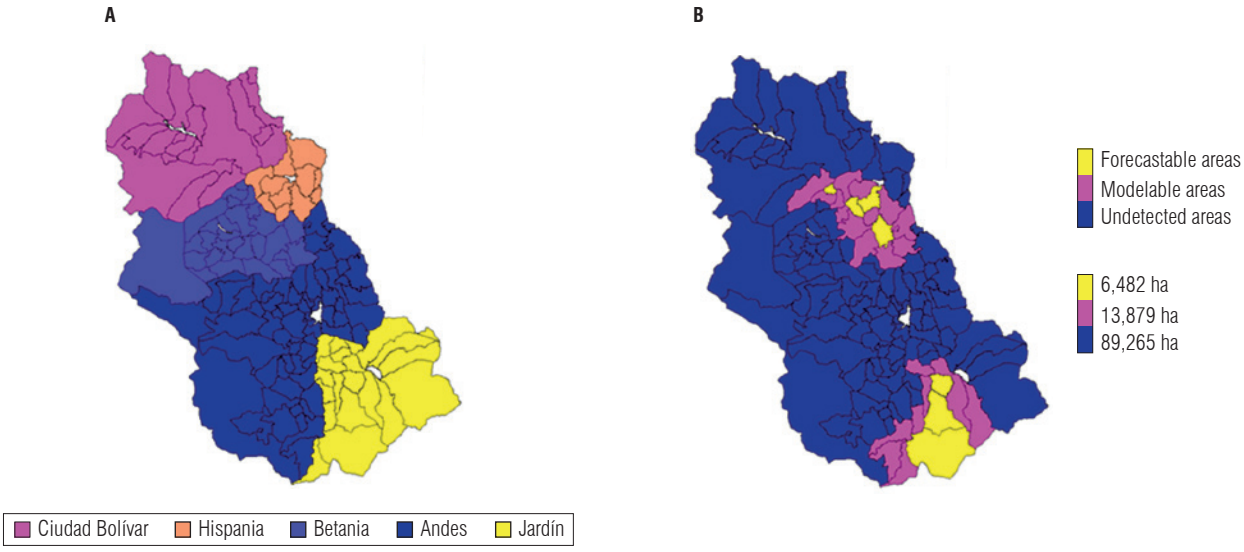

FIGURE 7. Location of the detected areas or MTSs: A) municipalities, B) detected areas.

in blue, the predictable areas in yellow (6,400 ha), and the modelable areas, which are the surrounding zones, in magenta (13,800 ha). Comparing the modelable area with a small municipality, such as Hispania, it can be observed that the area is slightly larger than the municipality. This finding implies that a space covering slightly over 20,000 ha has been identified where it is possible to apply a production model under the conditions of the present study.

Table 7 presents the most homogeneous MTSs according to spatial criteria and the relationship between the study variables.

The results suggest that a classifier of MTSs can be created according to their level of predictability. For the data in this study, the proposed model allows predicting production in an area of over 6,000 ha (R^2 = 0.79) and reducing uncertainty (R^2 = 0.61) to slightly over 20,000 ha, which constituted the most homogeneous set of MTSs and their surroundings. It should be noted that in the detected area, there was a wide variety of cultivars (Castillo, Colombia, Cenicafé 1, Caturra, Catimor, among others), in different stages of development and under different cultivation systems. Despite this, the methodology identified spatial patterns that make it possible to predict production using fruit weight as the only variable.

It is important to note that, from a territorial sampling perspective, the present approach required weighing only 30 fruits per tree, from three trees per lot, in three lots per farm, and three farms per MTS, to estimate coffee production in fruits per tree over an area of a little over 20,000 ha. This method can be easily implemented in the territory and represents a significant contribution to the coffee sector in mountainous areas, as it allows covering larger areas with minor adaptations to the parameters, resulting in practical and easy-to-apply models in rural areas.

Future projections and implications for regional coffee farming

The results of this study establish a solid foundation for improving agricultural planning in mountain coffee farming. Identifying areas with higher predictability allows for optimizing resource allocation, improving harvest planning, and reducing uncertainty in commercialization. However, to maximize impact and scalability, it is essential to incorporate advanced tools in computer vision and artificial intelligence.

Use of artificial intelligence and computer vision in yield prediction

The spatial analysis of coffee production can be enhanced with image detection and classification models, such as YOLO (You Only Look Once) (Bazame *et al*., 2021) and Mask R-CNN (Chen *et al*., 2019)which would allow them to allocate optimal labor and equipment, as well as other resources for harvesting, transportation, and marketing. Accurate estimation of the number of strawberry flowers and their distribution in a strawberry field is, therefore, imperative for predicting the coming strawberry yield. Usually, the number of flowers and their distribution are estimated manually, which is time-consuming, labor-intensive, and subjective. In this paper, we develop an automatic strawberry flower detection system for yield prediction with minimal labor and time costs. The system used a small unmanned aerial vehicle (UAV, to identify patterns in crop development. The integration of drone and satellite imagery with segmentation algorithms would enable:

- Estimating fruit load per tree through automated fruit detection in high-resolution images;
- Identifying phenotypic and developmental variability in large-scale plantations, correlating visual characteristics with productivity;

TABLE 7. Description of the detected MTSs.

Municipality	MTS	Area (ha)	Conversion (kg cherry coffee/kg dry parchment coffee)	Planting density (trees ha^{-1})	Average fruit weight (g)	Fruits per tree
Jardín	Macanas	3187	4.00	5935	1.31	2245
Jardín	Gibraltar	1310	5.40	4713	1.34	1045
Jardín	Verdún	522	5.38	5548	1.27	1552
Hispania	La Armenia	296	4.74	6573	0.93	3303
Hispania	La Seca	472	5.60	6987	0.95	2507
Betania	Las Animas	576	4.73	6497	0.93	3665
Betania	El Tablazo	119	4.36	6470	0.92	3485

MTS – Minimum territorial subdivision.

- Determining fruit maturity based on color changes, optimizing harvest timing;
- Monitoring crop health, detecting early signs of water stress, nutrient deficiencies, or pests.

Scalability and model expansion

Combining spatial data with computer vision models would extend yield prediction to larger territories, generating detailed information without the need for direct field measurements. To achieve this, key steps include:

- Expanding the collection of images and geospatial data, incorporating historical records and new sensors;
- Developing an automatic image labeling system, facilitating the creation of databases for training more accurate models;
- Integrating this methodology into agricultural monitoring platforms, allowing producers to access real-time predictions.

Conclusions

The proposed methodology allowed for identifying a regression model associated with coffee plant productivity, which estimates the number of its fruits based on a sample of 30 fruits, facilitating on-field monitoring of lot and farm productivity. The proposed methodology combines linear regression models and spatial analysis and could serve as a complementary tool for estimating coffee production in addition to traditional methods, contributing to the precision and reliability of coffee harvest forecasts in mountainous areas. The results of this study enable the identification of specific areas within the territory where coffee production can be jointly estimated. This result is valuable for farmers and cooperatives as it allows them to focus their efforts and resources on the most productive areas and better plan outreach activities led by cooperatives within the territory. The application of this methodology could facilitate more informed decision-making in mountainous coffee farming, enabling more sustainable and profitable practices in coffee production, benefiting both farmers and the sector. The incorporation of computer vision and machine learning in yield prediction would strengthen precision coffee farming, enabling more efficient and sustainable decision-making. This approach would position the region as a leader in agricultural innovation applied to mountain coffee farming.

Acknowledgments

This work was funded by the project with code BPIN 2021000100003, High-Level Human Capital Formation – Universidad Nacional de Colombia Corte II Nacional, and by Universidad Nacional projects (Hermes code 612113: Coffee Harvest Forecasting at the Farm Level Using Computer Vision Techniques and Hermes code 52937: Center of Excellence (CE) for Innovation in Mechanization and Energy Options for Family Agriculture). Francisco J. Rodríguez-Cortés has been partially supported by Universidad Nacional de Colombia, HERMES projects, Grant/Award Number: 61213.

Conflict of interest statement

The authors declare no conflicts of interests related to the publication of this article.

Author contributions

CACV: conceptualization, methodology, software, data curation, writing – original draft preparation, visualization, research, writing – review & editing. IDAT: writing – review & editing. FJRC: methodology, writing – review & editing. ELMC: writing – original draft, writing – review & editing. All authors reviewed the final version of the manuscript.

Literature cited

Abreu Júnior, C. A. M., Martins, G. D., Xavier, L. C. M., Vieira, B. S., Gallis, R. B. A., Fraga Junior, E. F., Martins, R. S., Paes, A. P. B., Mendonça, R. C. P., & Lima, J. V. N. (2022). Estimating coffee plant yield based on multispectral images and machine learning models. *Agronomy*, *12*(12), Article 3195. https://doi.org/10.3390/AGRONOMY12123195

Adane, A., & Bewket, W. (2021). Effects of quality coffee production on smallholders' adaptation to climate change in Yirgacheffe, Southern Ethiopia. *International Journal of Climate Change Strategies and Management*, *13*(4–5), 511–528. https://doi.org/10.1108/IJCCSM-01-2021-0002

Anselin, L. (1995). Local Indicators of Spatial Association – LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/J.1538-4632.1995.TB00338.X

Aparecido, L. E. O., Rolim, G. S., Lamparelli, R. A. C., Souza, P. S., & Santos, E. R. (2017). Agrometeorological models for forecasting coffee yield. *Agronomy Journal*, *109*(1), 249–258. https://doi.org/10.2134/agronj2016.03.0166

Bazame, H. C., Molin, J. P., Althoff, D., & Martello, M. (2021). Detection, classification, and mapping of coffee fruits during harvest with computer vision. *Computers and Electronics in Agriculture, 183*, Article 106066. https://doi.org/10.1016/J.COMPAG.2021.106066

Beksisa, L., Alamerew, S., Ayano, A., & Daba, G. (2018). Genotype environment interaction and yield stability of Arabica coffee (*Coffea arabica* L.) genotypes. *African Journal of Agricultural Research*, *13*(4), 210–219. https://doi.org/10.5897/ajar2017.12788

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11), Article 3758. https://doi.org/10.3390/S21113758

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-7618-4

Chasco Yrigoyen, C. (2006). Análisis estadístico de datos geográficos en geomarketing: el programa GeoDa. *Distribución y Consumo*, *16*(86), 34–47. https://dialnet.unirioja.es/servlet/articulo?codigo=1970463

Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., & He, Y. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. *Remote Sensing, 11*(13), Article 1584. https://doi.org/10.3390/rs11131584

Cheng, H., Damerow, L., Sun, Y., & Blanke, M. (2017). Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. *Journal of Imaging*, *3*(1), Article 6. https://doi.org/10.3390/jimaging3010006

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/J.COMPAG.2018.05.012

Choperena Bedoya, E., & Couleau, A. (2021). *El incumplimiento en la entrega de los contratos de futuros de café*. Informe Especial No. 45. EAFIT. https://www.eafit.edu.co/escuelas/economiayfinanzas/noticias-eventos/Documents/PDF_InformeEspecial_Diciembre_20211213.pdf

CountThings. (2021). CountThings. Dynamic Ventures, Inc. https://www.countthings.com

De Corso Sicilia, G. B., Pinilla Rivera, M., & Gallego Navarro, J. (2017). Métodos gráficos del análisis exploratorio de datos espaciales. *Cuadernos Latinoamericanos de Administración*, *13*(25), 92–104. https://www.redalyc.org/pdf/4096/409655122009.pdf

Eugenio, F. C., Grohs, M., Venancio, L. P., Schuh, M., Bottega, E. L., Ruoso, R., Schons, C., Mallmann, C. L., Badin, T. L., & Fernandes, P. (2020). Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. *Remote Sensing Applications: Society and Environment*, *20*, Article 100397. https://doi.org/10.1016/j.rsase.2020.100397

GeoDa Foundation. (2020). GeoDa. An introduction to spatial data science. https://geodacenter.github.io/

Gil Serna, J. G. (2012). *Estimación de un pronóstico de exportaciones de café suave colombiano: redes neuronales artificiales y ARDL* [Master thesis, Universidad EAFIT]. https://repository.eafit.edu.co/bitstream/handle/10784/11485/JuanGabriel_GilSerna_2016.pdf?sequence=2

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, Article 621. https://doi.org/10.3389/fpls.2019.00621

Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., Mushtaq, S., & Phuong Nguyen, V. (2018). Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Computers and Electronics in Agriculture*, *155*, 324–338. https://doi.org/10.1016/J.COMPAG.2018.10.014

Martello, M., Molin, J. P., Wei, M. C. F., Canal Filho, R., & Nicoletti, J. V. M. (2022). Coffee-yield estimation using high-resolution time-series satellite images and machine learning. *AgriEngineering*, *4*(4), 888–902. https://doi.org/10.3390/AGRIENGINEERING4040057

Mendoza, R, Fernández, E., & Kuhnekath, K. (2013). ¿Institución patrón-dependiente o indeterminación social? Genealogía crítica del sistema de habilitación en el café. *Ensayos sobre Economía Cafetera*, *26*(29), 145–161.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). John Wiley & Sons, Inc. https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis%2C+6th+Edition-p-9781119578727

Montoya Restrepo, E. C., Arcila Pulgarín, J., Jaramillo Robledo, A., Riaño Herrera, N. M., & Quiroga Zea, F. (2009). Modelo para simular la producción potencial del cultivo del café en Colombia. *Boletín Técnico Cenicafé,* (33), 5–52. https://biblioteca.cenicafe.org/handle/10778/588

Moraes-Oliveira, A. F., Aparecido, L. E. O., & Figueira, S. R. F. (2017). Economic and climatic models for estimating coffee supply. *Pesquisa Agropecuaria Brasileira*, *52*(12), 1158–1166. https://doi.org/10.1590/S0100-204X2017001200004

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrik*a, *37*(1/2), 17–23. https://doi.org/10.2307/2332142

Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Zhang, L., & Tian, Z. (2021). Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks. *International Journal of Applied Earth Observation and Geoinformation*, *102*, Article 102436. https://doi.org/10.1016/J.JAG.2021.102436

R Core Team (2020). The R project for statistical computing. https://www.r-project.org

Ramírez Builes, V. H. (2014). La fenología del café una herramienta para apoyar la toma de decisiones. *Avances Técnicos Cenicafé*, (114), 1–8. https://biblioteca.cenicafe.org/handle/10778/489

Ramos, P. J., Prieto, F. A., Montoya, E. C., & Oliveros, C. E. (2017). Automatic fruit count on coffee branches using computer vision. *Computers and Electronics in Agriculture, 137*, 9–22. https://doi.org/10.1016/j.compag.2017.03.010

Rendón-Sáenz, J. R., Arcila-Pulgarín, J., & Montoya-Restrepo, E. C. (2008). Estimación de la producción de café con base en los registros de floración. *Cenicafé*, *59*(3), 238–259. https://www.cenicafe.org/es/publications/arc059%2803%29238-259.pdf

Tanaka, S., Kawamura, K., Maki, M., Muramoto, Y., Yoshida, K., Akiyama, T., Cheng, T., Yang, Z., Inoue, Y., Zhu, Y., Cao, W., & Thenkabail, P. S. (2015). Spectral index for quantifying leaf area index of winter wheat by field hyperspectral measurements: A case study in Gifu prefecture, Central Japan. *Remote Sensing*, *7*(5), 5329–5346. https://doi.org/10.3390/RS70505329

Tsai, D. M., & Chen, W. L. (2017). Coffee plantation area recognition in satellite images using Fourier transform. *Computers and Electronics in Agriculture*, *135*, 115–127. https://doi.org/10.1016/j.compag.2016.12.020

Wang, Z., Walsh, K. B., & Verma, B. (2017). On-tree mango fruit size estimation using RGB-D images. *Sensors*, *17*(12), Article 2738. https://doi.org/10.3390/s17122738

Zhang, L., Zhang, Z., Luo, Y., Cao, J., & Tao, F. (2019). Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches. *Remote Sensing*, *12*(1), Article 21. https://doi.org/10.3390/RS12010021