SCIENTIA AGROPECUARIA

Scientia Agropecuaria

Web page: http://revistas.unitru.edu.pe/index.php/scientiaagrop

Facultad de Ciencias Agropecuarias

Universidad Nacional de Trujillo

RESEARCH ARTICLE

Machine learning analysis of subsistence determinants in Peruvian agricultural systems: A multi-regional approach using random forest and gradient boosting

Alejandro Coloma-Paxi[1]*; José Antonio Otoya-Barrenechea[1]; Edward Añaguari-Palomino[1]; Andrés Ladislao Cornejo-Pinto[2]

[1] Proyecto Mejoramiento del Sistema de Información Estadística Agraria y del Servicio de Información Agraria para el Desarrollo Rural del Perú (PIADER), Unidad Ejecutora de Gestión de Proyectos Sectoriales (UEGPS), Ministerio de Desarrollo Agrario y Riego del Perú, Lima 15047, Peru.

[2] Escuela Profesional de Ingeniería Agroindustrial, Universidad Nacional del Altiplano, Av. Floral 1153, Puno, Peru.

* Corresponding author: acolomap05@gmail.com (A. Coloma-Paxi).

Received: 10 December 2025. Accepted: 23 August 2026. Published: 7 September 2026.

Abstract

Subsistence agriculture persists across Peru with markedly heterogeneous patterns among ecological regions, yet multi-regional analyses using predictive methods remain scarce. This study identified determinants of household subsistence using random forest and gradient boosting classifiers. Microdata from the 2024 Peruvian National Agricultural Survey yielded 34,827 agricultural units after quality control. Subsistence was operationalised as the share of self-consumption in total production value and categorised as low (0% - 20%), medium (20% - 60%) and high (> 60%). On a held-out test set of 10,448 units both algorithms converged (random forest 82.9%, gradient boosting 82.6% accuracy), yet a nested sensitivity analysis showed that this performance depends substantially on predictors that enter the definition of the dependent variable. Once income-derived and market-participation variables were removed, accuracy settled at 65.7% against a 57.9% majority-class baseline, and harvested area and input expenditure emerged as the leading structural determinants. Three findings run against prevailing narratives. First, productive richness does not accompany commercialisation: Andean units combined the highest richness (7.9 versus 5.8 species on the Coast) with the highest subsistence (39.1% versus 15.0%). Second, direct-to-consumer selling was associated with higher subsistence (34.2%) than collector intermediation (14.3%), consistent with intermediaries operating as a pragmatic commercialisation route rather than a barrier. Third, accuracy was lowest precisely in the Andes (77.0%), where subsistence concentrates, and weakest for the transitional medium class (60.7%). Reporting predictive performance without separating definitional from structural predictors overstates what such models explain, a distinction rarely made in this literature.

Keywords: subsistence agriculture; machine learning; random forest; smallholder farmers; agricultural commercialisation.

DOI: https://doi.org/10.17268/sci.agropecu.2026.47

Cite this article:
Coloma-Paxi, A., Otoya-Barrenechea, A. A., Añaguari-Palomino, E., & Cornejo-Pinto, A. L. (2026). Machine learning analysis of subsistence determinants in Peruvian agricultural systems: A multi-regional approach using random forest and gradient boosting. *Scientia Agropecuaria*, *17*(3), 685-700.

1. Introduction

Subsistence agriculture remains a defining characteristic of smallholder farming systems in developing countries. Approximately 571 million farms operate worldwide, of which 500 million are smallholdings below two hectares that represent 85% of all farms, and these produce predominantly for household consumption rather than market exchange (FAO, 2025). In Peru, subsistence-oriented production persists despite decades of rural development interventions, particularly in Andean and Amazonian regions, where market exclusion, institutional fragmentation and ecological degradation perpetuate household food provisioning strategies (Vásquez Neyra et al., 2025).

Peru exemplifies this complexity. The country's agricultural sector spans three distinct ecological regions, namely the Coast, the Andes and the Amazon, each characterised by contrasting agroecological conditions, infrastructure endowments and levels of market integration (MIDAGRI, 2021). Coastal agriculture benefits from advanced irrigation infrastructure and digital technology adoption in regions such as Lima, Ica and Lambayeque, which enables commercially oriented production with productivity gains exceeding 50% (Calderan Gregolin et al., 2023). Andean smallholders, by contrast, face mountainous terrain, limited road access and minifundia land fragmentation, all of which perpetuate semi-subsistence systems (Trivelli & Berdegué, 2019). Studies of Andean smallholders in Peru have identified multiple livelihood capital domains affecting food access (Leah et al., 2013), although the relative weight of economic versus diversification strategies remains contested. Amazonian agriculture combines extensive land availability with low productivity and isolated markets, which creates distinct subsistence dynamics (Blanco et al., 2024). Adaptation to environmental variability in the central Andean coast has moreover shaped subsistence intensification over millennia, which suggests that these systems are long-standing adaptive responses rather than residual backwardness (Wilson et al., 2024).

How subsistence itself defines conditions what any analysis can conclude. Zantsi et al. (2026) show for South Africa that labelling smallholders as subsistence producers misrepresents their economic position, because most cultivate to supplement household food and income rather than as an exclusive livelihood, and that this misclassification has weakened agrarian policy. Operational definitions based on the destination of production, as adopted here, therefore require explicit scrutiny of how the resulting measure relates to the variables used to explain it.

Conventional analyses of subsistence determinants have primarily employed descriptive statistics or regression-based models such as probit and tobit specifications (Mgomezulu et al., 2024; Ogutu et al., 2020). These models impose a linear index of the predictors before the link or censoring transformation, so any interaction or threshold must be specified in advance rather than discovered. Subsistence decisions, however, plausibly involve complex non-linear interactions: economic thresholds that trigger diversification decisions, complementarities between productive diversity and market access, and region-specific trade-offs between household characteristics and welfare outcomes (Jimoh et al., 2024). Machine learning methods, particularly ensemble algorithms such as random forest and gradient boosting, capture such threshold effects and non-linear patterns without a priori functional form assumptions (Breiman, 2001; Hastie et al., 2009). These methods have proved effective in agricultural prediction tasks across diverse contexts (Asamoah et al., 2024; van Klompenburg et al., 2020), and comparative evidence on rural household surveys shows that they outperform conventional econometric specifications in predictive terms (Zhao et al., 2025). Machine learning is also increasingly combined with structural economic models to identify vulnerability in developing economies (Mukashov et al., 2026). Applications to subsistence determinants nevertheless remain limited, and reported accuracies are seldom accompanied by an assessment of how much of the predictive signal is definitional rather than substantive.

Marketing channel linkages such as contract farming have been theorised as critical pathways that enable agricultural commercialisation and rebalance livelihoods between subsistence and commodity production (Cole, 2022). Recent analyses of market linkages emphasise persistent barriers, including information asymmetry and high transaction costs (Ma et al., 2024), which particularly affect smallholders in developing regions. Conventional wisdom assumes that direct market access, through farmers' markets or direct-to-consumer sales, reduces subsistence by eliminating intermediaries. Transaction costs and minimum volume requirements may nevertheless favour collector intermediation for small-scale producers. Barriers to direct market participation, which include limited access to market information, credit constraints and transportation costs (Donkor et al., 2021), make collector relationships economically rational despite concerns about rent extraction. Distinguishing between channel usage patterns and their subsistence outcomes requires data-driven analysis.

The present study addresses three knowledge gaps. First, no prior research has applied machine learning methods to systematically identify subsistence determinants across Peru's ecological regions, which limits understanding of region-specific drivers. Second, the relative importance of economic capacity versus productive diversification strategies remains theoretically contested, and recent evidence suggests heterogeneous effects across income groups and livelihood portfolios (Jebessa et al., 2026; Nzira & Chegere, 2025; Tesfaye & Tirivayi, 2020). Third, the relationship between marketing channels and subsistence lacks robust quantification in Peruvian smallholder contexts, despite the policy emphasis on market linkage programmes (MIDAGRI, 2021). Understanding the determinants of subsistence levels is critical for designing effective agricultural policies, yet empirical evidence on multi-dimensional drivers remains fragmented, particularly with regard to region-specific patterns within ecologically heterogeneous countries.

This study therefore analysed the determinants of household subsistence levels in Peruvian agricultural systems using random forest and gradient boosting classifiers. The analysis first characterised and compared agricultural systems across the Coast, the Andes and the Amazon in terms of productive composition, economic capacity and marketing channel patterns. It then evaluated the comparative performance of both algorithms in classifying subsistence levels and examined relative variable importance through impurity-based measures, permutation importance and SHAP values. Finally, it quantified the contribution of socioproductive, economic and marketing channel factors, differentiated by ecological region, and separated predictors that are definitionally linked to the dependent variable from those that are structurally independent of it.

2. Methodology

The analysis is cross-sectional and predictive rather than causal. Its object is to establish how far the household and farm attributes recorded in a national agricultural survey discriminate between production orientations, and to separate the share of that discrimination that follows from the construction of the dependent variable itself from the share attributable to structural endowments. The unit of analysis is the agricultural unit, defined as a set of land under unified management. Four stages structure the procedure. The first assembles a unit-level dataset by integrating eight thematic modules of the survey and applying quality filters. The second operationalises subsistence as an ordinal outcome and constructs 19 predictors grouped in four conceptual domains. The third fits two ensemble classifiers on a stratified training partition and evaluates them exclusively on a held-out test partition. The fourth identifies determinants through three complementary importance measures and re-estimates the models under three nested specifications of decreasing definitional dependence on the outcome, which is what licenses the substantive claims reported in section 3. Figure 1 sets out this sequence together with the sample size retained at each filtering step.

2.1. Study area and data source

Primary data were obtained from Peru's 2024 National Agricultural Survey (ENA 2024), conducted by the National Institute of Statistics and Informatics (INEI) between May and October 2024. The survey employs a single-stage probabilistic area sampling design with complex stratification by province, ecological zone, agricultural intensity and plot size (MIDAGRI, 2021). The survey universe comprised agricultural units nationwide, defined as land areas under unified management for agricultural or livestock production. Complete microdata is publicly available through the INEI microdata portal as survey 973, together with the questionnaires, data dictionaries and the official technical datasheet that documents the sampling design and coverage (INEI, 2024). The 21 modules of the 2024 round are listed at https://proyectos.inei.gob.pe/microdatos/consulta.asp?cmbencuesta=ENCUESTA+NACIONAL+AGROPECUARIA&cmbanno=2024&cmbTrimestre=62 and the technical datasheet at https://proyectos.inei.gob.pe/iinei/srienaho/Descarga/DocumentosMetodologicos/2024-62/03_FICHA_TECNICA_ENA_2024.pdf

Purposive filtering selected agricultural units with complete data across the production, economic and sociodemographic modules, which yielded 34,827 units from an initial 40,237 cases, a retention of 86.5%. Exclusions comprised the top 2% of income outliers by region (n = 806), applied to reduce the influence of extreme values on model training, and cases with missing data for the subsistence calculation (n = 4,604), mostly attributable to zero production value or incomplete consumption records.

Data integration combined unit-level and detail-level files across eight thematic modules covering general characteristics, land tenure, production plans, commercialization, production costs and household composition. Records were merged using the survey identification keys for year, department, province, district, segment, producer and agricultural unit. Geographic information on departmental boundaries was obtained from the National Geographic Institute (IGN, 2023).

The study area encompassed three ecological regions: the Coast (n = 8,516; 24.5%), the Andes (n = 20,084; 57.7%) and the Amazon (n = 6,227; 17.9%), spanning 24 departments. All 34,827 units enter the statistical analyses, but the cartographic displays cover 33,759 of them: Callao (n = 3) is omitted because of its insufficient sample size, and 1,065 units (3.1%) enumerated with the ENA enterprise questionnaire carry no department code and therefore cannot be mapped. The latter are concentrated on the Coast, where they represent 9.4% of the regional sample against 0.7% in the Andes and 2.0% in the Amazon, so the departmental map slightly under-represents coastal units.

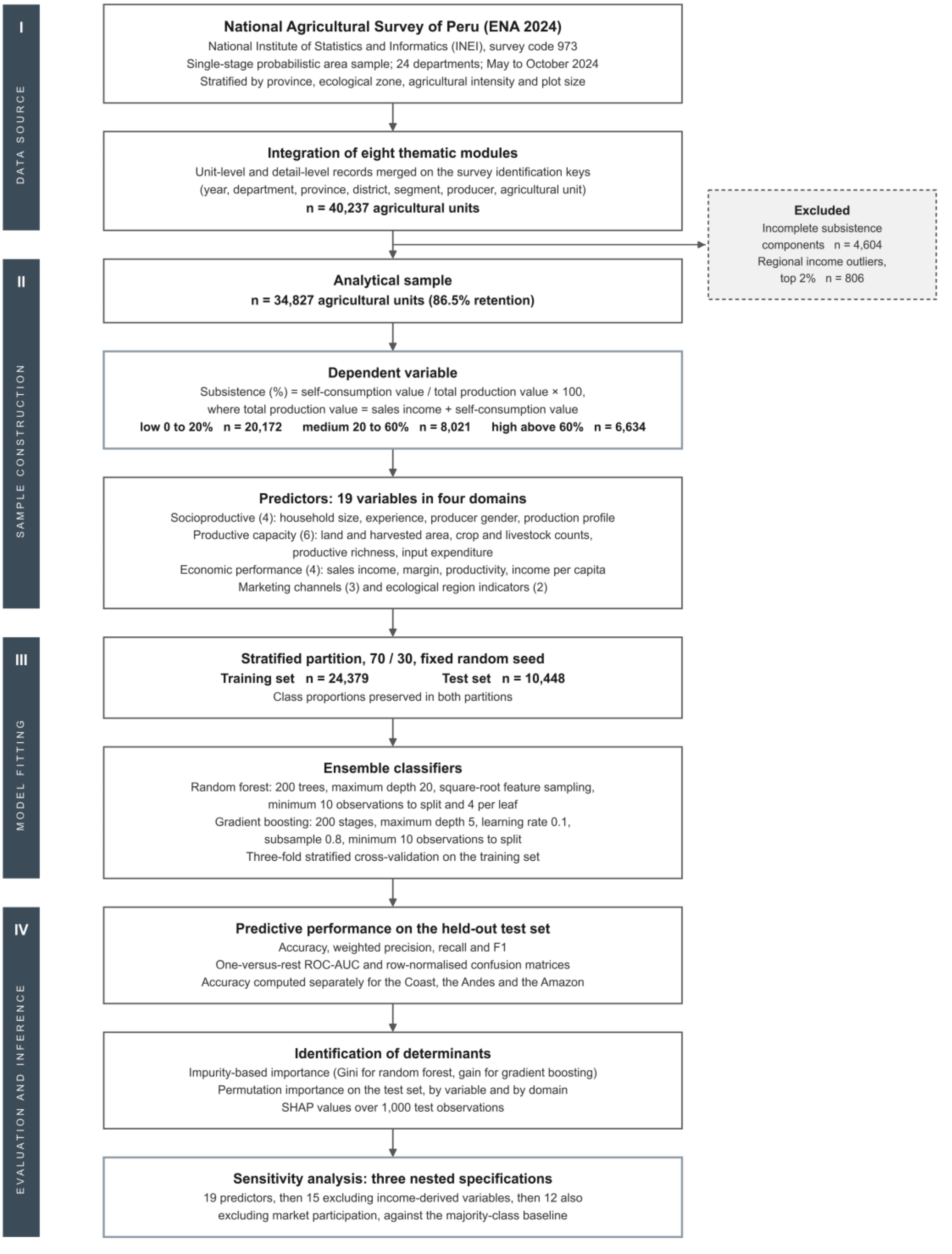

Figure 1. Analytical workflow, organised in four stages: assembly of the unit-level dataset from the ENA 2024 microdata, construction of the dependent variable and the predictors, model fitting on a stratified training partition, and evaluation on the held-out test partition. Solid boxes state the operation performed at each step and the sample size retained; dashed boxes record the data source and the methodological decisions that condition the following step.

2.2. Variable construction and operationalisation

The dependent variable, subsistence level, was operationalised as the percentage of self-consumption relative to total production value:

$$Subsistence_i\ (\%) = [Self\text{-}consumption\ value_i\ /\ Total\ production\ value_i] \times 100$$

where total production value equals sales income plus self-consumption value for both agricultural and livestock components. This metric aligns with the operational definition of subsistence orientation used by FAO (2025) and permits continuous measurement between 0% and 100% that is suitable for classification after categorisation. Subsistence levels were categorised following operational definitions adapted from the FAO farm orientation typology (Lowder et al., 2016): low subsistence (0-20%, market-oriented production), medium subsistence (20% - 60%, semi-subsistence systems) and high subsistence (above 60%, predominantly subsistence-oriented production). These thresholds align with Peruvian agricultural policy frameworks that distinguish commercial, transitional and subsistence producers (MIDAGRI, 2021).

Independent variables comprised four conceptual domains: socioproductive characteristics, productive capacity, economic performance and marketing channels. Table 1 details their operationalization.

Several methodological refinements were implemented. Following econometric practice, productivity was operationalised as income per harvested hectare rather than per total land area (Fuglie, 2018), because total area includes fallow and non-productive zones. Agricultural margin excluded livestock components, to avoid conflating distinct production systems with different cost structures. The term productive richness was adopted in preference to diversification, to avoid implying a strategic intentionality that cross-sectional data cannot capture (Nzira & Chegere, 2025).

2.3. Data processing and quality control

Missing values in count variables, such as the number of crops or livestock species, were imputed as zero under the assumption that the absence of a response indicated non-participation. Continuous variables with less than 5% missingness, namely surface areas, household size, production costs and farming experience, were imputed using regional medians in order to preserve ecological specificity (Little & Rubin, 2019). Four consistency checks were run on the analytical sample. First, 3,794 units (10.9%) record a subsistence value of exactly 100%; by construction all of them report zero sales income, and they were retained because a complete absence of commercial sales is a substantively meaningful state rather than a coding error. Second, harvested area exceeds total land area in 382 units (1.1%); these were retained without correction and are flagged here as a residual inconsistency of the source data, since harvested area may aggregate two campaigns on the same plot within the reference period. Third, no unit reports marketing channels together with zero sales income, so no case required adjudication on that ground. Fourth, 14 units (0.04%) record productivity above PEN 1,000,000 per harvested hectare; these were retained after the outlier trimming described in section 2.1, since they correspond to intensive horticulture and high-value crops on very small harvested areas.

Table 1
Variable operationalization

Category	Variable	Description	Unit
Dependent variable	Subsistence level	Self-consumption as a share of total production value	Low (0-20%), medium (20-60%), high (> 60%)
Socioproductive characteristics	Household size	Number of household members	Count
	Farming experience	Years in agricultural production	Years
	Producer gender	Gender of the agricultural producer	Male / female / legal entity
	Production profile	Type of production system	Mixed / livestock only
Productive capacity	Total land area	Total farm area	Hectares
	Harvested area	Area under active cultivation	Hectares
	Number of crops	Count of distinct crop species cultivated	Count
	Number of livestock species	Count of distinct livestock species raised	Count
	Productive richness	Sum of crop and livestock species	Count
	Agricultural production cost	Total input costs (seed, fertiliser, pesticide, labour)	PEN
Economic performance	Total sales income	Agricultural plus livestock revenue	PEN
	Agricultural margin	Agricultural sales minus agricultural costs	PEN
	Productivity	Total income per harvested hectare	PEN per hectare
	Income per capita	Total income divided by household size	PEN per capita
Marketing channels	Agricultural channels	Number of agricultural marketing channels used	Count (0-6)
	Livestock channels	Number of livestock marketing channels used	Count (0-5)
	Total marketing channels	Sum of agricultural and livestock channels	Count
Ecological region	Andes indicator	Unit located in the Andes	Binary (Coast is the reference)
	Amazon indicator	Unit located in the Amazon	Binary (Coast is the reference)

PEN = Peruvian soles. The table lists the 19 variables that enter the models. Channel types comprise collector, wholesaler, retailer, cooperative, agro-industry and direct consumer. Producer gender was taken from the household roster record corresponding to the producer, identified by the parenthood code, rather than from a household-level aggregate. The same procedure was used to recover producer age, which is reported later as a sample descriptor but is not used as a predictor.

Regional dummy variables were constructed using one-hot encoding with a drop-first strategy to avoid multicollinearity, which created binary indicators for the Andes and the Amazon with the Coast as reference category. Producer gender and production profile were label-encoded as categorical variables suitable for tree-based algorithms.

2.4. Machine Learning Models

Two ensemble tree-based algorithms were employed for their complementary strengths: random forest (Breiman, 2001) and gradient boosting (Friedman, 2001). Random forest constructs multiple decorrelated decision trees through bootstrap aggregation and random feature sampling, which captures non-linear relationships while maintaining a low overfitting risk (Hastie et al., 2009). Gradient boosting sequentially builds trees that correct the residual errors of previous iterations, which achieves high predictive accuracy but requires careful hyperparameter selection to avoid overfitting (James et al., 2013).

Model specifications are detailed in Table 2. Hyperparameters were selected based on established defaults and preliminary grid search experiments, prioritising computational efficiency over exhaustive tuning given the large sample size (Probst et al., 2019).

Table 2
Machine learning model specifications

Parameter	Random forest	Gradient boosting
n_estimators	200	200
max_depth	20	5
learning_rate	-	0.1
min_samples_split	10	10
min_samples_leaf	4	-
subsample	-	0.8
max_features	sqrt	-
random_state	42	42

A dash indicates that the parameter does not apply to the algorithm. Random forest uses bootstrap aggregation with square-root feature sampling; gradient boosting employs stochastic gradient descent with learning rate shrinkage.

The dataset was partitioned into training (70%, n = 24,379) and test (30%, n = 10,448) subsets using stratified random splitting with a fixed random seed, which preserved class proportions across subsistence categories. Models were fitted on the training subset, and all performance metrics reported in this study were computed on the held-out test subset. Three-fold stratified cross-validation on the training subset assessed model stability.

Variable importance was quantified through three complementary approaches. Impurity-based importance, computed as the mean decrease in Gini impurity for random forest and as gain for gradient boosting, represents the average contribution of each variable across all trees (Louppe et al., 2013). Because impurity-based measures are known to distribute importance arbitrarily among correlated predictors, permutation importance was additionally computed on the test subset with ten repetitions. Finally, SHAP values (Lundberg & Lee, 2017) were obtained for a random subsample of 1,000 test observations, which provides feature contribution estimates rooted in cooperative game theory.

2.5. Structural dependence between the dependent variable and the predictors

The dependent variable is a ratio whose denominator contains sales income, and sales income also enters the predictor set as total income together with three of its transformations: income per capita, agricultural margin and productivity. The dependent variable can be reproduced exactly from total income and self-consumption value for all 34,827 units, so these variables are related to the outcome by construction rather than through an independent empirical channel. An analogous dependence affects the marketing channel counts: units reporting no channel record no sales, and 83.8% of the 4,525 such units have a subsistence value of exactly 100%, whereas none of the 30,302 units reporting at least one channel reaches that value.

Reporting a single accuracy figure would therefore conflate definitional with substantive prediction. Following the practice adopted in comparable work on household survey microdata, where variables used to construct the label are excluded from the predictor set (Coelho da Silva et al., 2025), three nested specifications were estimated on the identical partition. The full specification retains all 19 predictors. The second excludes the four income-derived variables. The third additionally excludes the three marketing channel counts, and thus retains only household attributes, land and input endowments, productive composition and region, none of which enters the construction of the dependent variable. The third specification is treated as the basis for substantive claims about determinants; the full specification is retained for comparability with the predictive literature.

2.6. Model evaluation

Classification performance was assessed using standard multi-class metrics: accuracy, weighted precision, weighted recall, weighted F1-score and the area under the receiver operating characteristic curve, computed through a one-versus-rest strategy for the multi-class extension. Confusion matrices normalised by row visualised class-specific prediction patterns. Cross-validation mean accuracy and standard deviation quantified generalisation stability. Because the three subsistence classes are unbalanced, the accuracy of a majority-class classifier is reported as a reference baseline.

Regional performance heterogeneity was evaluated by computing accuracy separately for the Coast, Andes and Amazon subsets of the test partition, which tested whether model effectiveness varied across ecological contexts. The marketing channel analysis focused on producers with positive sales income and an identifiable channel (n = 30,302). The principal channel was defined as the first channel reported in hierarchical order, with agricultural channels prioritised over livestock channels, since the binary channel data record only use and non-use and therefore preclude identifying the highest-value channel.

All analyses were conducted in Python 3.12 using scikit-learn 1.6.1, pandas 2.2.2, NumPy 2.0.2, matplotlib 3.10.0, seaborn 0.13.2, geopandas 1.1.1 for cartographic analysis and shap 0.50.0 for SHAP values. Analysis scripts that reproduce every table and figure from the public ENA 2024 microdata are available from the corresponding author on reasonable request.

3. Results and discussion

3.1. Sample characterisation and regional disparities

Table 3 presents descriptive statistics by ecological region and reveals pronounced socioeconomic disparities. Andean producers constituted 57.7% of the sample (n = 20,084), followed by the Coast (24.5%, n = 8,516) and the Amazon (17.9%, n = 6,227). Producers averaged 55.9 years of age, and Amazonian producers were the youngest (52.7 years). Women headed 25.2% of the units, with the highest share in the Andes (28.5%) and the lowest in the Amazon (19.4%). Household size was largest in the Amazon (2.94 members) and smallest on the Coast (2.52 members), which is consistent with rural-urban migration patterns and coastal labour market opportunities.

Regional economic disparities were extreme. Coastal farmers earned 31.9 times more income than their Andean counterparts (PEN 188,458 versus PEN 5,912) and retained 63.9 times higher agricultural margins (PEN 131,387 versus PEN 2,056). Productivity was 2.7 times higher on the Coast (PEN 20,688 per hectare versus PEN 7,664 per hectare), which reflects irrigation infrastructure, technology adoption and high-value crop specialisation. These disparities express structural inequalities in market access and infrastructure that require national-level investment beyond farm-level interventions.

Andean systems exhibited 35% greater productive richness (7.87 species versus 5.83 on the Coast) yet earned one thirty-second of the income and retained one sixty-fourth of the margin. Multi-species farming therefore appears to represent a risk-reduction strategy under subsistence constraints rather than a commercialisation pathway, which aligns with evidence from East Africa (Tesfaye & Tirivayi, 2020) and South Asia (Akber et al., 2022) where diversification emerges as a survival strategy under market failure. Evidence from Ethiopian forest communities points in the same direction, since livelihood strategies that raise food security combine off-farm activities with agriculture rather than intensifying on-farm diversity alone (Jebessa et al., 2026). Subsistence levels averaged 28.5% nationally but varied markedly by region. The Andes (39.1%) exhibited 2.6 times higher subsistence than the Coast (15.0%) and 3.0 times higher than the Amazon (13.0%). The low subsistence of the Amazon despite low income (PEN 38,382) probably reflects extensive livestock production with high sales volumes but low per-unit profitability. Amazonian agriculture faces distinct challenges, including disease pressure and institutional support gaps, as documented for cacao production systems in Ucayali (Blanco et al., 2024).

Table 3
Sample characterisation by ecological region

Variable	Coast	Andes	Amazon	Overall
Sample size (n)	8,516	20,084	6,227	34,827
Producer age (years)	56.9	56.5	52.7	55.9
Female producers (%)	21.0	28.5	19.4	25.2
Household size	2.52	2.84	2.94	2.78
Total area (ha)	17.03	7.31	32.37	14.16
Productive richness	5.83	7.87	6.49	7.13
Subsistence (%)	15.02	39.09	13.02	28.54
Total income (PEN)	188,458	5,912	38,382	56,355
Productivity (PEN per ha)	20,688	7,664	8,255	10,846
Agricultural margin (PEN)	131,387	2,056	28,094	38,336

Producer age and gender are available for the 33,762 units enumerated with the ENA household questionnaire. They are undefined for 1,065 units (3.1%) enumerated with the separate ENA enterprise questionnaire, whose producer is a legal entity and which therefore has no household roster; these units are excluded from the age and gender columns but retained in every other statistic. All monetary values are expressed in Peruvian soles (PEN).

Figure 2 displays the geographic distribution of subsistence levels. The choropleth map reveals a spatial concentration in the central and southern Andes: Apurímac (48.2%), Ayacucho (41.2%), Puno (40.7%) and Ancash (40.0%) are the only departments above 40% mean subsistence, and all four are Andean. At the opposite end lie Tumbes (4.5%), Ucayali (5.3%), San Martín (9.8%) and Lima (13.1%). The contrast is clearer at the level of the natural region, where the Andes average 39.1% against 15.0% on the Coast and 13.0% in the Amazon, than at department level, since several coastal and Amazonian departments exceed 25%, among them Piura (32.5%), La Libertad (30.1%) and Madre de Dios (27.4%). The ranking of the 15 departments with the largest producer samples shows that Ancash (n = 2,893), Cajamarca (n = 2,554) and Puno (n = 2,414) lead, all of them Andean, which confirms that subsistence agriculture is predominantly an Andean phenomenon in Peru.

Regional distributions (Figure 3) revealed distinct patterns. Andean subsistence showed a wide interquartile range and a median of approximately 27%, which indicates high within-region heterogeneity, whereas the Coast and the Amazon exhibited compressed distributions with medians below 5%. Total income distributions on a logarithmic scale demonstrated the coastal income advantage over the Andes, close to an eightfold difference at the median. Land area distributions confirmed Amazonian land extensivity against Andean minifundia, yet the larger Amazonian farms did not translate into higher subsistence, which evidences that land size alone does not determine subsistence when market access and productivity are taken into account.

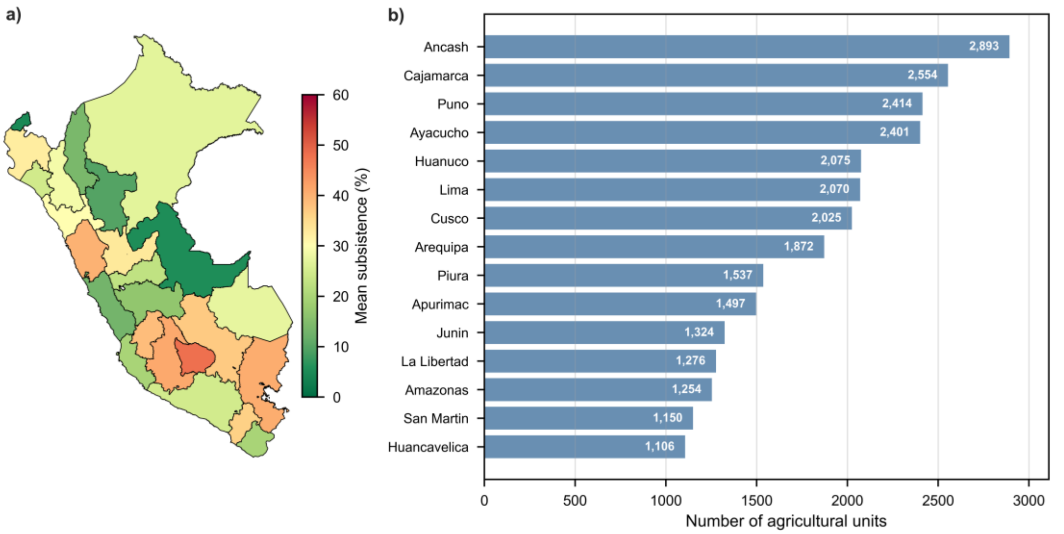

Figure 2. Geographic distribution of subsistence in Peru, based on the 33,759 agricultural units with a department code. a) Mean subsistence percentage by department; b) the 15 departments with the largest number of surveyed agricultural units. Callao (n = 3) and the 1,065 units enumerated with the enterprise questionnaire, which carry no department code, are excluded from the map but retained in all statistical analyses.

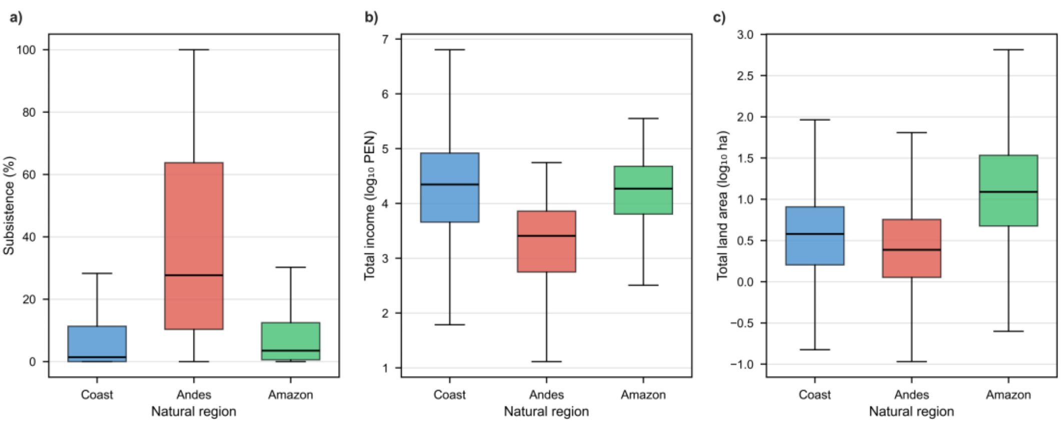

Figure 3. Regional comparison of key variables across the Coast, the Andes and the Amazon. Boxes span the interquartile range, the horizontal line is the median and whiskers extend to 1.5 times the interquartile range, with outliers omitted for legibility. a) Subsistence percentage; b) total income on a base-10 logarithmic scale; c) total land area on a base-10 logarithmic scale.

3.2. Model performance

Table 4 presents comparative model performance on the held-out test set alongside reference values from recent studies that classify household status from survey microdata. Random forest and gradient boosting achieved almost identical accuracy (82.88% versus 82.63%) and F1-scores (0.827 versus 0.825), which indicates convergence on a common predictive solution. Cross-validation confirmed generalisation stability, with random forest averaging 82.86% (SD 0.25%) and gradient boosting 82.49% (SD 0.25%) across folds.

Placing these values in context matters. The accuracy obtained here sits close to the 83% and F1-score of 0.81 reported by Gholami et al. (2022) for Malawian household food security, and below the 86.6-88.5% obtained by Nkurunziza et al. (2025) for household consumption classes in Rwanda with a comparable national survey. It exceeds the 71.6-72.6% that Solano-Jiménez et al. (2026) obtained with logistic specifications in rural Uganda and the 65.7% that Zhao et al. (2025) obtained for rural China. The closest Latin American reference, Mazenda (2026), reports an area under the curve of 0.758 for food insecurity among rural Colombian households, well below the values obtained here; the sensitivity analysis presented in section 3.3 indicates that this gap is largely attributable to the definitional predictors retained in the full specification rather than to superior model performance. Comparisons with studies reporting coefficients of determination, such as the random forest regression of maize yield in Ghana with a model efficiency coefficient of 0.81 (Asamoah et al., 2024), are informative about the family of methods but not about classification performance, since the prediction target differs. Earlier work on smallholder commercialisation in developing countries has relied predominantly on causal-inference designs, including generalised propensity score and dose-response models (Ogutu et al., 2020), which are optimised for unbiased treatment-effect estimation rather than for out-of-sample classification, so their goodness-of-fit statistics are not an appropriate benchmark for the present results.

Figure 4 decomposes performance by class. Discrimination was strongest for the high subsistence class (AUC 0.982 for both algorithms), intermediate for low subsistence (0.948 and 0.949) and weakest for the medium class (0.893 and 0.891). The confusion matrices show the same ordering. Random forest classified 90.9% of low subsistence units and 85.2% of high subsistence units correctly, but only 60.7% of medium subsistence units, of which 31.4% were assigned to the low class and 7.9% to the high class. Gradient boosting behaved almost identically, with 90.2%, 85.0% and 61.5% respectively.

Table 4
Model performance in the present study and reference values from recent studies classifying household status from survey microdata

Study and context	Model	n	Target	Accuracy	Other metrics
Present study, Peru (ENA 2024)	Random forest	34,827	Subsistence, 3 classes	0.8288	F1 0.827; AUC 0.89-0.98; CV 0.8286 ± 0.0025
Present study, Peru (ENA 2024)	Gradient boosting	34,827	Subsistence, 3 classes	0.8263	F1 0.825; AUC 0.89-0.98; CV 0.8249 ± 0.0025
Present study, structural predictors only	Random forest	34,827	Subsistence, 3 classes	0.6573	F1 0.646; AUC 0.76-0.86
Nkurunziza et al. (2024), Rwanda (EICV5)	Multiple-kernel SVM, XGBoost, multinomial logit	14,580	Consumption class	0.866-0.885	Best of 12 classifiers
Gholami et al. (2022), Malawi	Random forest	1,886 households	Food security, 2 classes	0.83	F1 0.81
Solano-Jiménez et al. (2026), Uganda	Logistic regression	173	Food security, 2 classes	0.716-0.726	AUC 0.754-0.762; Nagelkerke R^2 0.286-0.292
Zhao et al. (2025), rural China	Random forest, LASSO	1,080	Food insecurity	0.657	Outperformed econometric specifications
Coelho da Silva et al. (2025), Brazil (Ceará)	Stacked ensemble, GBM, XGBoost	33,595	Food insecurity, 4 classes	0.553	F1 0.559; 0.756 when binarised
Mazenda (2026), rural Colombia	Ensemble and logistic regression	2,771	Food insecurity, 2 classes	n.r.	AUC 0.758
Asamoah et al. (2024), Ghana	Random forest (regression)	3,136 plots	Maize yield	n.a.	MEC 0.81; 0.53-0.63 for agronomic efficiency

Metrics for the present study were computed on the 30% held-out test set (n = 10,448). CV = three-fold stratified cross-validation on the training set. Reference studies differ in target definition, number of classes and sample size, so their metrics are indicative of the range reported in this literature rather than directly comparable. MEC = model efficiency coefficient.

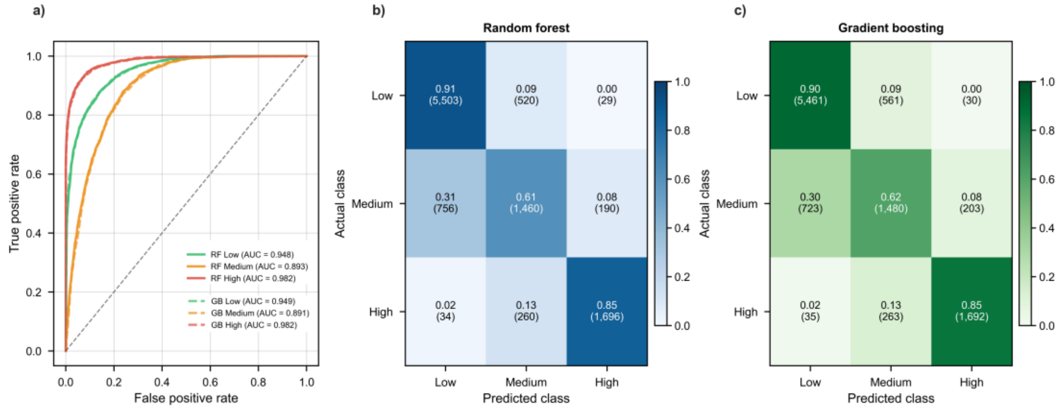

Figure 4. Classification performance on the held-out test set (n = 10,448). a) One-versus-rest ROC curves by subsistence class, with solid lines for random forest and dashed lines for gradient boosting; b) row-normalised confusion matrix for random forest; c) row-normalised confusion matrix for gradient boosting. In both matrices the absolute number of agricultural units is shown in parentheses, and proportions below 0.005 are displayed as 0.00.

The medium category is therefore the weakest point of the classification, not its strongest. This is methodologically interpretable: the 20% - 60% band spans a wide range of production orientations, and units near either threshold share characteristics with the adjacent class. The rarity of confusion between the low and high classes, below 2% in both directions, confirms that the models separate fundamentally different production orientations even where boundary cases exist. The practical implication is that these models identify clearly commercial and clearly subsistence-oriented units reliably, but should not be used to target transitional producers, who are precisely the population that commercialization programs seek to reach.

3.3. Determinants of subsistence and the limits of predictive importance

Table 5 reports variable importance for the full specification. Total income and income per capita dominate, and together account for 43.8% of impurity-based importance in random forest and 73.2% in gradient boosting. This apparent dominance, however, largely restates the construction of the dependent variable, since sales income enters its denominator, as established in section 2.5. The two variables are moreover strongly collinear (Spearman correlation 0.96), so impurity-based measures divide importance between them in a way that depends on the training sample: the estimation reported here ranks total income first, whereas their combined importance remains stable across random partitions. Interpreting either income variable individually as the leading determinant is therefore unwarranted. Block-wise permutation of the test set makes the point plainly: permuting the whole economic performance domain costs 35.5 percentage points of accuracy for random forest and 35.6 for gradient boosting, against 5.3 and 6.1 for productive capacity, 0.2 and 0.3 for marketing channels, 0.1 and 0.3 for region, and none for socioproductive characteristics. Permuting total income on its own already costs 15.9 percentage points for random forest and 38.9 for gradient boosting; that the single-variable figure exceeds the block figure for gradient boosting is not a contradiction but a consequence of collinearity, since permuting the block jointly leaves the model without any of the correlated income signals and it falls back on the remaining domains, whereas permuting one variable in isolation destroys the signal the model relies on most while the correlated survivors continue to mislead it.

Table 6 reports the nested sensitivity analysis. Removing the four income-derived variables lowered accuracy from 82.9% to 74.8%, and additionally removing the three marketing channel counts lowered it to 65.7%, against a majority-class baseline of 57.9%. Of the 25.0 percentage points by which the full model exceeds the baseline, 17.2 are therefore attributable to variables that are definitionally tied to the outcome, and 7.8 to structural endowments. It is on this last specification that the substantive conclusions of this study rest.

In the structurally independent specification, harvested area (importance 0.225 for random forest and 0.264 for gradient boosting) and agricultural production cost (0.210 and 0.198) were the leading determinants, followed by total land area (0.124 and 0.100) and the Andes indicator (0.104 and 0.189). Productive richness ranked well below these (0.064 and 0.058). Productive scale and the level of input expenditure, rather than the number of species produced, therefore distinguish subsistence-oriented from market-oriented units. This result is consistent with the descriptive pattern of section 3.1

and provides the substantive counterpart to it: richness correlates positively with subsistence (Spearman 0.26), so productive diversity accompanies subsistence rather than commercialisation.
These finding challenges paradigms that treat diversification as a universal commercialisation pathway. Recent evidence shows that diversification effects depend on crop portfolio composition, with cash crops promoting commercialisation while staples show negative associations (Nzira & Chegere, 2025). The Peruvian evidence presented here indicates that, in the current context, promoting species diversity without addressing productive scale and input access is unlikely to shift units towards the market. Figure 5 visualises importance patterns for the full specification. The comparison of impurity-based importance confirms the concentration on income variables and reveals algorithm-specific emphases, since gradient boosting allocates a far larger share to total income than random forest. Mean absolute SHAP values corroborate this ordering. The dependence plot for income per capita shows a decreasing and saturating pattern rather than a gradual one. SHAP contributions to the high subsistence class are positive below approximately PEN 400 per capita, cross zero between PEN 400 and PEN 650, and then remain essentially flat at about -0.045 across three further orders of magnitude of income. The inflection therefore occurs at a few hundred soles per capita and not in the upper tail of the income distribution, where fewer than 2% of units are located. It should moreover be read as a descriptive feature of the definitional relationship between sales income and the subsistence ratio rather than as a behavioural threshold, a distinction that the present analysis draws explicitly.

Table 5
Variable importance in the full specification

Variable	Random forest	Gradient boosting	Mean
Total income (PEN)[a]	0.2513	0.6196	0.4355
Income per capita (PEN)[a]	0.1863	0.1124	0.1493
Agricultural margin (PEN)[a]	0.1217	0.0365	0.0791
Productivity (PEN per ha)[a]	0.0914	0.0365	0.0640
Productive richness	0.0416	0.0645	0.0531
Harvested area (ha)	0.0461	0.0203	0.0332
Agricultural production cost	0.0377	0.0235	0.0306
Total number of channels[b]	0.0474	0.0045	0.0259
Total area (ha)	0.0286	0.0211	0.0248
Number of livestock species	0.0255	0.0211	0.0233
Experience (years)	0.0205	0.0112	0.0159
Number of crops	0.0225	0.0081	0.0153
Andes region	0.0217	0.0085	0.0151
Number of agricultural channels[b]	0.0255	0.0035	0.0145
Number of livestock channels[b]	0.0141	0.0027	0.0084
Household members	0.0103	0.0037	0.0070
Producer gender	0.0041	0.0013	0.0027
Amazon region	0.0029	0.0008	0.0018
Production profile	0.0008	0.0002	0.0005

Importance values sum to 1.0 within each model. Mean importance is the simple average of the two algorithms. [a] Variables that are definitionally linked to the dependent variable, because sales income enters the denominator of the subsistence ratio; they are reported for completeness rather than as substantive determinants. [b] Market-participation variables, which are also partly definitional, since units reporting no channel record no sales.

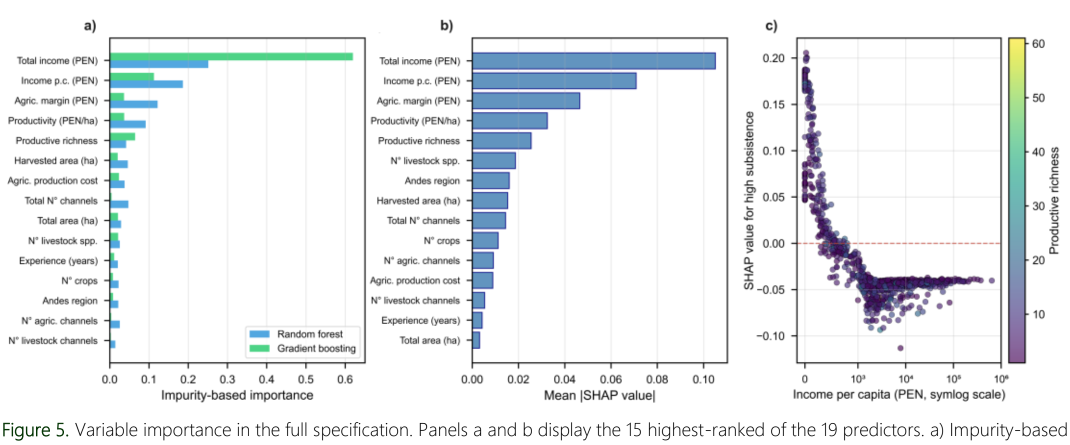

Figure 5. Variable importance in the full specification. Panels a and b display the 15 highest-ranked of the 19 predictors. a) Impurity-based importance for random forest and gradient boosting; b) mean absolute SHAP value of the random forest over 1,000 test observations, averaged across the three subsistence classes; c) SHAP dependence plot of the random forest for income per capita on the high subsistence class, with points coloured by productive richness.

Table 6
Nested sensitivity analysis on the held-out test set

Specification	Predictors	Algorithm	Accuracy	F1	AUC low	AUC medium	AUC high
Full	19	Random forest	0.8288	0.8266	0.948	0.893	0.982
Full	19	Gradient boosting	0.8263	0.8249	0.949	0.891	0.982
Excluding income-derived predictors	15	Random forest	0.7481	0.7434	0.897	0.816	0.937
Excluding income-derived predictors	15	Gradient boosting	0.7416	0.7395	0.896	0.812	0.938
Excluding income and market participation	12	Random forest	0.6573	0.6458	0.858	0.759	0.819
Excluding income and market participation	12	Gradient boosting	0.6561	0.6480	0.858	0.756	0.817
Majority-class baseline	-	-	0.5792	-	-	-	-

All specifications were estimated on the identical stratified partition (training n = 24,379; test n = 10,448). The majority-class baseline on the test set is 0.5792. AUC values are reported per class using a one-versus-rest strategy.

3.4. Regional heterogeneity in determinants and performance

Figure 6 examines region-specific patterns. Rank correlations were used in preference to product-moment correlations because the income distribution is extremely right-skewed, particularly on the Coast, where a small number of very large operations otherwise dominates the linear coefficient. The difference is not cosmetic: on the Coast, the product-moment correlation between total income and subsistence is 0.13 whereas the rank correlation is 0.68, and for harvested area the two measures give 0.04 and 0.53 respectively.

Total income was the strongest correlate of subsistence in all three regions (0.68 on the Coast, 0.78 in the Andes and 0.59 in the Amazon), which is the expected consequence of the definitional relationship set out in section 2.5. The informative heterogeneity lies in what follows income. On the Coast, the next strongest correlates are measures of productive scale and intensity, namely agricultural production cost (0.55), harvested area (0.53) and the number of livestock species (0.45). In the Andes, by contrast, the number of agricultural marketing channels ranks second (0.38), ahead of harvested area (0.33) and production cost (0.25). In the Amazon, livestock species (0.41), harvested area (0.37) and productive richness (0.36) lead, whereas marketing channels are essentially uncorrelated with subsistence (0.02). Market access, in other words, discriminates subsistence levels in the Andes but not in the Amazon, where the binding constraint appears to be the composition and scale of production rather than the route to market.

Model accuracy varies substantially by region. The Coast achieved the highest accuracy (91.7% for both algorithms), followed by the Amazon (89.5% and 89.2%), while the Andes recorded the lowest values (77.0% and 76.6%). This gradient is the inverse of what policy relevance would require, since the Andes concentrates subsistence agriculture and simultaneously proves hardest to classify. Within-region heterogeneity explains the pattern: Andean subsistence spans the full 0-100% range within a single ecological zone, which blurs class boundaries. The comparison of observed and predicted class distributions for the random forest shows that the model under-predicts the medium class in the Andes, where it is most prevalent, and over-assigns units to the low class; gradient boosting behaves in the same way, as its confusion matrix in Figure 4c indicates.

This heterogeneity carries differentiated policy implications. The Andes is the only region where marketing channel access ranks among the leading correlates of subsistence once income is set aside, which supports concentrating on market linkage programmes there rather than applying them uniformly. Collective action approaches that connect Andean producers with market actors have demonstrated success through networks that facilitate access to local markets and increase household income (Vallejo-Rojas et al., 2022), and the present evidence indicates that this is the region where such approaches have most traction. On the Coast, where subsistence averages 15% and channel access is already widespread, the relevant margin is productive scale and input intensity, so policy should target the constraints that limit harvested area and input use rather than market articulation. The Amazon presents a distinct challenge, since extensive landholdings averaging 32 hectares coexist with low productivity of PEN 8,255 per hectare and with channel counts that carry no information about subsistence; policies promoting intensification through improved pastures and silvopastoral systems could unlock commercialisation without expanding the agricultural frontier into primary forest. Across all three regions, the structurally independent specification points to the same lever, namely the scale of harvested production and the capacity to purchase inputs, which is consistent with the conclusion that subsistence responds to productive capacity rather than to species diversity.

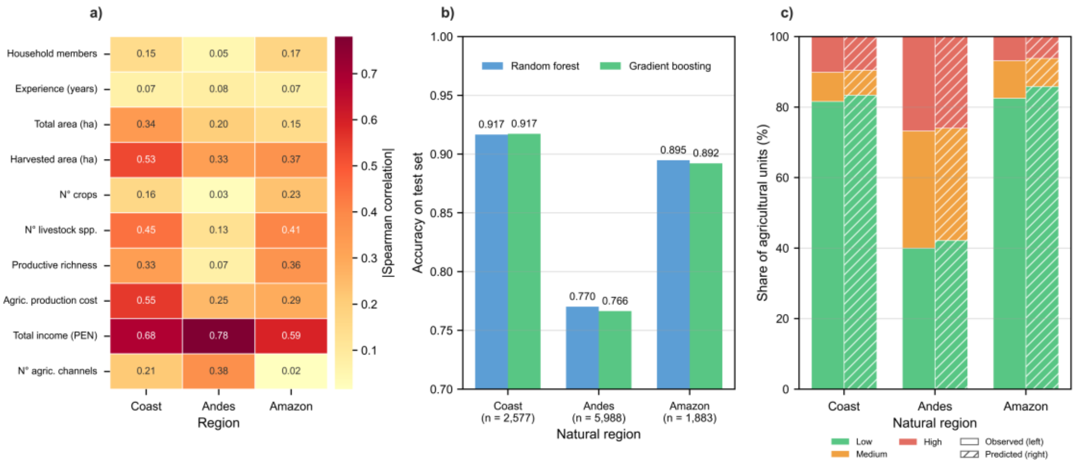

Figure 6. Regional heterogeneity in determinants and performance. a) Absolute Spearman correlations between subsistence percentage and the ten continuous predictors of the socioproductive and productive capacity domains, by region; b) accuracy of both algorithms by region on the test set, with the vertical axis truncated at 0.70 to make the differences visible; c) observed distribution of subsistence classes by region compared with the distribution predicted by the random forest.

3.5. Marketing channel patterns and transaction cost implications

Figure 7 analyses marketing channel patterns and their relationship with subsistence among the 30,302 producers with an identifiable channel. Channel usage by region reveals regional specialisation. Agricultural collectors dominated across all regions but with varying intensity: the Coast (50.0%), the Amazon (45.7%) and the Andes (31.8%). Wholesaler access was highest on the Coast (25.0%) and in the Amazon (25.5%) and lowest in the Andes (17.4%), which suggests infrastructural barriers in mountainous terrain. Cooperative sales were marginal nationwide (0.5% of producers with an identifiable channel) and agro-industry sales were only slightly more frequent (1.9%), which contradicts the policy emphasis on associative and contract-based models. Retailer and final consumer channels were concentrated in the Andes (13.2% and 12.9%) and much rarer on the Coast (7.2% and 4.0%) and in the Amazon (11.4% and 3.3%), a pattern consistent with the greater weight of local fairs and short chains in highland areas.

A counterintuitive pattern emerges from the channel means. The final consumer channel is associated with the highest subsistence (34.2%) and low income (PEN 12,434), whereas agricultural collectors show much lower subsistence (14.3%) with far higher income (PEN 56,861). Wholesalers occupy the commercial extreme (11.6% subsistence and PEN 84,092), and retailers and livestock collectors fall in between (26.3% and 26.9%). Direct selling is therefore associated with more subsistence orientation, not less, which contradicts the assumption that eliminating intermediaries advances commercialisation.

This pattern aligns with transaction cost economics. Direct-to-consumer sales correlate with higher subsistence most plausibly because producers selling directly at local markets lack access to collector networks that demand consistent volumes, or because final consumer sales occur at isolated rural fairs where subsistence farmers sell small surpluses rather than at urban farmers' markets. Collector intermediation, which concentrates volume and sustains regular transactions, is in turn associated with markedly lower subsistence. Small-scale producers face documented costs in accessing direct marketing channels independently, including market information acquisition, credit constraints and transportation barriers (Donkor et al., 2021), which is consistent with intermediaries operating as a pragmatic commercialisation pathway. All of these associations are cross-sectional and none should be read as evidence that channel choice causes subsistence levels, since both may respond to unobserved differences in scale and market proximity; the ENA 2024 records channel use as a binary indicator and does not collect transaction prices, so the relative profitability of each channel cannot be assessed with these data.

The distribution of subsistence categories by principal channel confirms these patterns. Agricultural collectors and wholesalers served predominantly low subsistence producers, whereas the final consumer channel exhibited the most subsistence-skewed profile. Cooperative and agro-industry channels, not displayed because they account for 0.5% and 1.9% of producers respectively, showed even lower mean subsistence (5.1% and 2.5%) alongside the highest mean incomes of the sample, which suggests that formal market linkages require a threshold level of commercialisation rather than serving as subsistence escape mechanisms.

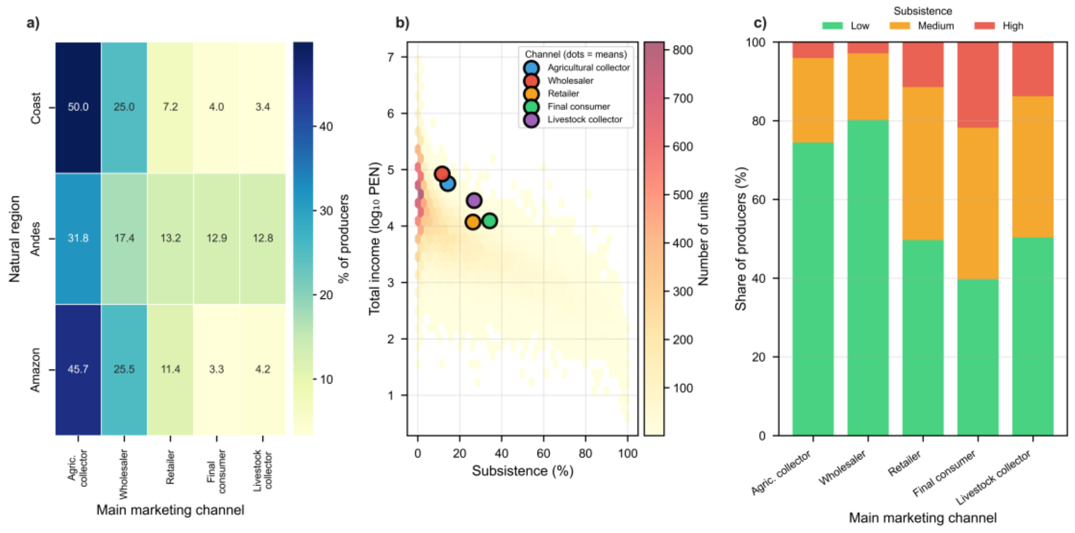

Figure 7. Marketing channels and subsistence among producers with an identifiable channel (n = 30,302). a) Channel use by region; b) joint density of subsistence and total income, with large points marking the mean of each channel; c) distribution of subsistence classes by principal channel.

The policy implications are clear but uncomfortable for prevailing development paradigms: facilitating intermediary relationships may be more effective than promoting direct sales for subsistence producers. This includes regulating collector behaviour to prevent monopsonistic exploitation through transparent pricing and contract enforcement, supporting collector-farmer associations for collective bargaining, and subsidising collection infrastructure such as rural assembly points and cold storage rather than urban farmers' markets. Direct market access programmes should target already-commercial producers seeking margin optimisation, not subsistence farmers who lack the basic preconditions for commercialisation.

4. Conclusions

This study classified subsistence levels in 34,827 Peruvian agricultural units from the 2024 National Agricultural Survey and reached 82.9% accuracy with random forest and 82.6% with gradient boosting on a held-out test set. A nested sensitivity analysis established that most of this performance is definitional rather than substantive: once income-derived and market-participation variables were removed, accuracy settled at 65.7% against a 57.9% majority-class baseline, and harvested area and input expenditure emerged as the leading structural determinants. Productive richness does not accompany commercialisation in the Peruvian context, since Andean units combine the highest richness with 2.6 times the subsistence of the Coast and one sixty-fourth of its agricultural margin, which contradicts paradigms that equate diversification with market integration. Marketing channel patterns likewise challenge disintermediation assumptions, because direct-to-consumer sales were associated with 34.2% subsistence against 14.3% for collector intermediation, which suggests that transaction costs outweigh intermediary rents for smallholders. Classification proved least accurate in the Andes (77.0%), the region where subsistence concentrates, and weakest for the transitional medium class (60.7%), which is precisely the population that commercialisation programmes target. Regional heterogeneity demands differentiated strategies. The Andes is the only region where marketing channel access ranks among the leading correlates of subsistence once income is set aside, which argues for concentrating market linkage programmes there rather than deploying them uniformly. On the Coast, where channel access is already widespread, the binding margin is productive scale and input intensity, and in the Amazon channel counts carry no information about subsistence, so sustainable intensification of extensive systems is the relevant lever. Among the variables the survey makes available, subsistence in Peru is therefore governed by productive scale and market infrastructure; whether producer knowledge and access to extension add explanatory power could not be tested here, because the ENA 2024 does not record them at the unit level.

Three gaps remain open and define a research agenda to which these results contribute. First, the cross-sectional design identifies associations but cannot establish causality; linking successive ENA rounds into household panels would permit difference-in-differences and event-study designs that trace transition pathways between subsistence and commercial orientation, and the structurally independent specification reported here provides the predictor set on which such designs should be built. Second, the poor classification of transitional producers indicates that the determinants separating semi-subsistence from commercial units are not captured by the survey; targeted instruments measuring perceived income needs, risk preferences and market distance, should be tested as complements to the ENA questionnaire. Third, the analysis contains no agroclimatic dimension, although climate variability is a documented driver of smallholder production decisions, coupling the ENA microdata with gridded climate series and altitude would allow the models presented here to be re-estimated with environmental predictors and would test whether the regional indicators used in this study are proxies for agroclimatic conditions. Beyond research, the classification developed here can support the geographic targeting of MIDAGRI extension and market linkage programmes, provided it is applied to identify clearly subsistence-oriented units rather than transitional ones, and provided that the distinction between definitional and structural predictors documented in this study is preserved in any operational use.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Ministerio de Desarrollo Agrario y Riego (MIDAGRI) through the Unidad Ejecutora Gestión de Proyectos Sectoriales (UEGPS) under Service Order No. 0002671, within the framework of the Programa de Investigación Aplicada para el Desarrollo Rural (PIADER). The authors also thank the National Institute of Statistics and Informatics (INEI) for producing and publicly releasing the ENA 2024 microdata.

Authors contribution

A. Coloma-Paxi: Conceptualization, Methodology, Formal analysis, Data curation, Software, Validation, Visualization, Writing - original draft, Writing - review and editing. J. A. Otoya-Barrenechea: Conceptualization, Methodology, Resources, Supervision, Writing - review and editing, Project administration. E. Añaguari-Palomino: Data curation, Formal analysis, Software, Validation, Visualization. A. L. Cornejo-Pinto: Conceptualization, Methodology, Resources, Supervision, Writing - review and editing, Funding acquisition.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Data availability statement

The microdata analysed in this study are publicly available from the National Institute of Statistics and Informatics of Peru through its microdata portal at https://proyectos.inei.gob.pe/microdatos/, under survey code 973, National Agricultural Survey, year 2024, period 62. The analysis scripts that reproduce every table and figure of this article from those microdata are available from the corresponding author on reasonable request.

Ethical considerations

This study used secondary data from Peru's 2024 National Agricultural Survey (ENA 2024) conducted by the National Institute of Statistics and Informatics (INEI). The original survey protocol complied with national statistical confidentiality legislation and with international standards for research involving human participants, including informed consent procedures. All records were anonymised before release, and no personal identifiable information was accessed or reported in this study.

ORCID

A. Coloma-Paxi https://orcid.org/0000-0002-2678-5104
J. A. Otoya-Barrenechea https://orcid.org/0009-0007-0702-6958
E. Añaguari-Palomino https://orcid.org/0009-0006-8236-7819
A. L. Cornejo-Pinto https://orcid.org/0009-0007-2094-6521

References

Abdullahi, M. A., Muse, A. H., Muse, Y. H., Hassan, M. A., & Nadarajah, S. (2026). Harvests under heat: Quantifying the long-run impacts of climate and fertilizers on Somalia cereal yields: An ARDL analysis. *Earth Systems and Environment*, *10*(5), 6599-6610. https://doi.org/10.1007/s41748-025-00972-w

Akber, M. A., Islam, M. A., Rahman, M. M., & Rahman, M. R. (2022). Crop diversification in southwest coastal Bangladesh: Insights into farming adaptation. *Agroecology and Sustainable Food Systems*, *46*(2), 316-324. https://doi.org/10.1080/21683565.2021.1989105

Aniye, H. W., Bekele, T., & Worku, W. (2025). Trends of climate variability and its association with farmers' perceptions of climate change in Gassera district, southeastern Ethiopia. *Scientific Reports*, *15*(1), Article 4399. https://doi.org/10.1038/s41598-025-89017-w

Asamoah, E., Heuvelink, G. B. M., Chairi, I., Bindraban, P. S., & Logah, V. (2024). Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana. *Heliyon*, *10*(17), e37065. https://doi.org/10.1016/j.heliyon.2024.e37065

Blanco, M., Mosquera, L. E., Crisostomo, D., Díaz Casana, C., Ritter, T., & Mockshell, J. (2024). Exploring cacao business models and agroecological transitions in Ucayali, Peru (Agroecological Transitions Program Report). Bioversity International and International Center for Tropical Agriculture.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Calderan Gregolin, A., Contreras, M., Salgado-Funes, E., Tejada, G., & Zabaleta, I. (2023). The use of efficient Smart Agro 4.0 irrigation systems for cotton in Peru and its public policy implications. *World Water Policy*, *9*(4), 746-755. https://doi.org/10.1002/wwp2.12157

Coelho da Silva, T. L., Furtado, L. S., Fernandes, G. S., Fernandes de Macêdo, J. A., Almada Cruz, L., Pires Magalhães, R., & Amorim Gomes, L. G. (2025). Machine learning methods and models to predict food insecurity levels for families in Ceará, Brazil, based on employment, housing and other social indicators. *Journal of the Brazilian Computer Society*, *31*(1), 203-218. https://doi.org/10.5753/jbcs.2025.3100

Cole, R. (2022). Cashing in or driving development? Cross-border traders and maize contract farming in northeast Laos. *Journal of Agrarian Change*, *22*(1), 139-161. https://doi.org/10.1111/joac.12460

Donkor, E. A., Garnevska, E., Siddique, M. I., & Donk or, E. (2021). Determinants of rice farmer participation in the direct marketing channel in Ghana. *Sustainability*, *13*(9), 5047. https://doi.org/10.3390/su13095047

Food and Agriculture Organization of the United Nations. (2025). The state of food and agriculture 2025: Addressing land degradation across landholding scales. FAO. Rome.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Fuglie, K. O. (2018). Is agricultural productivity slowing? *Global Food Security*, *17*, 73-83. https://doi.org/10.1016/j.gfs.2018.05.001

Gholami, S., Knippenberg, E., Campbell, J., Andriantsimba, D., Kamle, A., Parthasarathy, P., Sankar, R., Birge, C., & Lavista Ferres, J. (2022). Food security analysis and forecasting: A machine learning case study in southern Malawi. *Data & Policy*, *4*, e33. https://doi.org/10.1017/dap.2022.25

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Instituto Geográfico Nacional. (2023). Límites departamentales [Data set]. Plataforma Nacional de Datos Abiertos. https://www.datosabiertos.gob.pe/dataset/limites-departamentales

Instituto Nacional de Estadística e Informática. (2024). Encuesta Nacional Agropecuaria 2024: Ficha técnica. Dirección Nacional de Censos y Encuestas, INEI. https://proyectos.inei.gob.pe/iinei/srienaho/Descarga/DocumentosMetodologicos/2024-62/03_FICHA_TECNICA_ENA_2024.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jebessa, G. M., Engura, T. T., & Bekele, A. E. (2026). Exploring the nexus between forest resource use, livelihood strategies, and household food security in Ethiopia. Discover Food, 6(1), Article 285. https://doi.org/10.1007/s44187-026-01021-9

Jimoh, A. K., Olawuyi, S. O., Balogun, T. A., & Ganiyu, M. O. (2024). Causal effect of agricultural diversification on smallholder farmers' welfare status in Nigeria. Research on World Agricultural Economy, 5(4), 458-480. https://doi.org/10.36956/rwae.v5i4.1223

Leah, J., Pradel, W., Cole, D. C., Prain, G., Creed-Kanashiro, H., & Carrasco, M. V. (2013). Determinants of household food access among small farmers in the Andes: Examining the path. *Public Health Nutrition*, *16*(1), 136-145. https://doi.org/10.1017/S1368980012000183

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems*, *26*, 431-439.

Lowder, S. K., Skoet, J., & Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, *87*, 16-29. https://doi.org/10.1016/j.worlddev.2015.10.041

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765-4774.

Ma, W., Rahut, D. B., Sonobe, T., & Gong, B. (2024). Linking farmers to markets: Barriers, solutions, and policy options. *Economic Analysis and Policy*, *82*, 1102-1112. https://doi.org/10.1016/j.eap.2024.05.005

Mazenda, A. (2026). Predicting household food insecurity through a livelihood lens: A machine learning approach in rural Colombia. *Journal of International Development*, *38*(5), 827-840. https://doi.org/10.1002/jid.70092

Mgomezulu, W. R., Chitete, M. M. N., Maonga, B. B., Kachingwe, L., Phiri, H. H., Mambosasa, M., & Folias, L. (2024). Does shifting from subsistence to commercial farming improve household nutrition and poverty? Evidence from Malawi, Tanzania and Nigeria. *Research in Globalization*, *8*, 100201. https://doi.org/10.1016/j.resglo.2024.100201

Ministerio de Desarrollo Agrario y Riego. (2021). Metodología de la construcción del marco maestro muestral para la ENA-MIDAGRI y demás operaciones estadísticas en el periodo intercensal (1.ª ed.). Unidad Ejecutora Gestión de Proyectos Sectoriales, Proyecto PIADER. https://ena.midagri.gob.pe/media/attachments/2021/10/08/metodologia_construccion_mmm_ena-midagri.pdf

Mukashov, A., Robinson, S., Arndt, C., Thurlow, J., & Thomas, T. (2026). Systematic risk profiling: Assessing compounding economic risks in developing countries. *Economic Modelling*, *157*, 107511. https://doi.org/10.1016/j.econmod.2026.107511

Nkurunziza, F., Kabanda, R., & McSharry, P. (2025). Enhancing poverty classification in developing countries through machine learning: A case study of household consumption prediction in Rwanda. *Cogent Economics & Finance*, *13*(1), 2444374. https://doi.org/10.1080/23322039.2024.2444374

Nzira, J. E., & Chegere, M. J. (2025). Diversify and thrive: Panel data evidence on crop portfolio effects on farmer income and farm output market supply. *Journal of Agriculture and Food Research*, *24*, 102339. https://doi.org/10.1016/j.jafr.2025.102339

Ogutu, S. O., Gödecke, T., & Qaim, M. (2020). Agricultural commercialisation and nutrition in smallholder farm households. *Journal of Agricultural Economics*, *71*(2), 534-555. https://doi.org/10.1111/1477-9552.12359

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Solano-Jiménez, J., Abadía, R., Pinilla Palleja, R., & Martínez-Carrasco, L. (2026). Determinants of household food security in rural Uganda: Perceived needs as key predictors. *Agriculture & Food Security*, *15*(1), Article 14. https://doi.org/10.1186/s40066-025-00594-5

Tesfaye, W., & Tirivayi, N. (2020). Crop diversity, household welfare and consumption smoothing under risk: Evidence from rural Uganda. *World Development*, *125*, 104686. https://doi.org/10.1016/j.worlddev.2019.104686

Trivelli, C., & Berdegué, J. A. (2019). Rural transformation: Looking towards the future of Latin America and the Caribbean (2030 - Food, Agriculture and Rural Development in Latin America and the Caribbean No. 1). Food and Agriculture Organization of the United Nations. https://openknowledge.fao.org/handle/20.500.14283/ca5508en

Vallejo-Rojas, V., Rivera-Ferre, M. G., & Ravera, F. (2022). The agri-food system (re)configuration: The case study of an agroecological network in the Ecuadorian Andes. *Agriculture and Human Values*, *39*(4), 1301-1327. https://doi.org/10.1007/s10460-022-10318-1

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Vásquez Neyra, J. M., Gomes Haensel Schmitt, V., & Cequea, M. M. (2025). Sustainable circular strategies and short supply chains among smallholder producers: The case of Peru's Agroferias Campesinas. *Agricultural and Food Economics*, *13*(1), 12. https://doi.org/10.1186/s40100-025-00423-0

Wilson, K. M., McCool, W. C., & Brenner Coltrain, J. (2024). Climate and oceanic condition changes influence subsistence economic adaptation through intensification on the central Andean coasts. *Quaternary International*, *689-690*, 79-92. https://doi.org/10.1016/j.quaint.2023.07.002

Zantsi, S., Masola, L., & Shiba, W. (2026). Appropriate farmer classification is key to agricultural development: Rethinking subsistence agriculture in South Africa. *International Journal of Social Economics*, *53*(8), 1288-1298. https://doi.org/10.1108/IJSE-04-2025-0401

Zhao, L., Yang, M., Min, S., & Qing, P. (2025). Prediction of household food insecurity in rural China: An application of machine learning methods. *International Food and Agribusiness Management Review*, *28*(2), 296-316. https://doi.org/10.22434/ifamr1137