Agricultural Sciences, 2026, 17(7), 617-632
https://www.scirp.org/journal/as
ISSN Online: 2156-8561
ISSN Print: 2156-8553

Determinants of Fish Farming Production in N'Zérékoré, Guinea: A Weighted Logistic Regression and Random Forest Approach

Ibrahima Sory Mamikouny Camara[1], Bintou Traore[2], Aladji Babacar Niang[1], Bakary Traore[1]

[1]Department of Mathematics, Faculty of Science and Technique, University of N'Zérékoré, N'Zérékoré, Guinea
[2]High Institute of Technology (IST), Mamou, Guinea
Email: mamikouny2013@gmail.com, btraore850@gmail.com

How to cite this paper: Camara, I.S.M., Traore, B., Niang, A.B. and Traore, B. (2026) Determinants of Fish Farming Production in N'Zérékoré, Guinea: A Weighted Logistic Regression and Random Forest Approach. *Agricultural Sciences*, **17**, 617-632.
https://doi.org/10.4236/as.2026.177037

Received: June 15, 2026
Accepted: July 14, 2026
Published: July 17, 2026

Abstract

This study examines the determinants of annual fish production in the prefecture of N'Zérékoré (Guinea) using data from a representative sample of 212 fish farmers, corresponding to a weighted population of approximately 530 producers. A weighted logistic regression model and a survey-weighted random forest were jointly applied to 1) identify the main factors associated with high annual fish production and 2) assess out-of-sample predictive performance under a complex multi-stage sampling design. The results show that sexing practice, the number of fingerlings stocked per 100 units, and total pond surface area per 100 m^2 were significantly and positively associated with high annual fish production. Farmer experience showed a positive but non-significant association in the final model. The weighted logistic regression model achieved a survey-weighted AUC of 0.861, indicating excellent discriminatory ability, while the weighted random forest reached an accuracy of 0.817 and a survey-weighted AUC of 0.856 on the test set. Despite methodological differences, both approaches converge on the same key determinants, particularly stocking intensity, pond surface area, and sexing practice, confirming the robustness of the findings. By combining an interpretable statistical model with a flexible non-parametric method, this study provides a coherent and complementary analytical framework for analyzing weighted agricultural data. Beyond its contribution to understanding annual fish production in Guinea, the study proposes a reproducible methodological approach that can be applied to similar agricultural surveys in West Africa.

Keywords

Fish Farming, Production, Survey Weights, Weighted Logistic Regression, Random Forest, R

1. Introduction

Fish farming has become an essential strategy for strengthening food security, generating income, and promoting sustainable development in low-income countries. In a context marked by rapid population growth, overexploitation of fishery resources, and increased vulnerability of rural households, aquaculture offers a viable solution for diversifying the supply of animal protein and enhancing the resilience of local agricultural systems [1].

However, the sector remains severely constrained by the effects of climate change, including recurrent droughts and variable rainfall patterns, which disrupt aquatic ecosystems and affect fish reproduction [2] [3]. In addition to these environmental factors, several economic and institutional constraints hamper fish production: the orientation of fish exports towards international markets at the expense of local supply [4], issues related to land governance, limited access to credit—a well-documented obstacle in Sub-Saharan Africa [5]—and persistent challenges in technical support. Several studies also emphasize the importance of innovation and skills development for improving aquaculture performance. Among the strategies explored are the integration of agro-industrial by-products in fish feeding [6] [7], the use of water recirculating systems [8], optimized pond-draining management [9], and the ecological enhancement of ponds as biodiversity reservoirs [10].

In Côte d'Ivoire, a study conducted in the Gontougo region showed that 67% of fish farmers produce less than 500 kg of fish per year, while only 33% achieve between 500 and 2000 kg [11]. These low yields were attributed to limited access to inputs, low levels of technical training, and a mismatch between practices and environmental conditions. Similar findings were reported in Bangladesh, where [12] demonstrated that tilapia production performance is closely linked to technical efficiency, stocking density, pond size, and farmer experience.

In the Forest Region of Guinea, particularly in the prefecture of N'Zérékoré, fish farming has experienced significant growth over the past two decades, supported by various development programs. The number of fish farmers increased from approximately 350 in 2008 to around 1800 in 2021, distributed across 242 localities [13]. Despite this expansion, the sector remains fragile. Most producers practice rice-fish farming, whose annual yields are highly variable and difficult to estimate accurately. According to [14], the main obstacles are not related to land availability but rather to insufficient technical support, a gradual loss of local expertise, and weak knowledge-sharing networks.

A major challenge lies in the difficulty of reliably measuring and predicting annual fish production at the farm level. Available data are often heterogeneous, characterized by extreme values, collinearity among socio-economic variables, and substantial environmental variability. Moreover, official statistics typically focus on structural indicators such as the number of producers or the cultivated area, without incorporating the technical or socio-economic factors that explain differences in annual fish production.

From a methodological perspective, the statistical analysis of these data presents several challenges. The information arises from a multi-stage sampling design, requiring the use of survey weights to obtain unbiased and representative estimates. Furthermore, several variables exhibit skewed distributions, are sensitive to extreme values, and involve complex interactions that are difficult to model parametrically. In this context, combining a weighted logistic regression model with a survey-weighted random forest offers a robust analytical framework: the former provides interpretable effect estimates under a complex survey design, while the latter captures nonlinear relationships and interactions that are difficult to specify a priori. Few studies in West Africa have jointly applied these complementary methods to weighted agricultural data, despite their methodological relevance.

In this study, annual fish production is defined as the total quantity of fish produced per farm over one year. We combine two complementary supervised learning approaches, weighted logistic regression to identify the key explanatory factors, and a survey-weighted random forest to assess variable importance and improve predictive performance. To our knowledge, this study represents one of the first joint applications of these methods in Guinea for analyzing and forecasting fish production, with a dual objective: to advance empirical research in the West African context and to provide policymakers with more reliable tools for aquaculture management.

2. Materials and Methods

2.1. Population

The target population consisted of fish farmers operating in the N'Zérékoré prefecture who had completed at least two fishing seasons. The sampling frame was provided by the Guinean Federation of Fish and Rice Farmers (FPR-G) and included 11 sub-prefectures (SP), each subdivided into D_h districts.

The following notation is used: N_{hd} denotes the number of fish farmers in district d within sub-prefecture h,

$$N_h = \sum_{d=1}^{D_h} N_{hd}, \quad N = \sum_{h=1}^{H} N_h = 635, \tag{1}$$

where N_h is the number of fish farmers in sub-prefecture h, H is the total number of sub-prefectures, and N is the total number of fish farmers in the N'Zérékoré prefecture. The value $N = 635$ refers to the initial sampling frame, before field participation and analytical eligibility criteria were applied.

2.2. Sample Size

The sample size was determined in order to estimate a mean with an absolute precision set at $\delta = 50$ kg. This value was established during preliminary consultations with local technical experts, who considered it a relevant threshold for comparing annual fish production differences. The estimated standard deviation from the pilot survey conducted in six representative districts was $s = 387.07$ kg, and a 95% confidence level was used ($Z = 1.96$).

The formula for determining the sample size under simple random sampling without replacement [15] is:

$$n_{SRS} = \frac{Z^2 \times N \times s^2}{Z^2 \times s^2 + (N-1) \times \delta^2}. \quad (2)$$

This gives $n_{SRS} = 169.16 \approx 169$.

To account for the clustered sampling design, the design effect was applied:

$$DEFF = 1 + (m-1)\rho, \quad (3)$$

where m is the average cluster size and ρ the intra-cluster correlation. In this study, $m = 5$ and $\rho = 0.043$, producing $DEFF \approx 1.172$ and an adjusted sample size $n' = n_{SRS} \times DEFF \approx 198.26$.

After adjusting for a 10% expected non-response rate, the final sample size is:

$$n_{final} = \frac{n'}{1-0.10} \approx 220. \quad (4)$$

2.3. Sampling Plan

Six sub-prefectures $H_1 = 6$ were selected by systematic probability-proportional sampling (PPS), size being N_h. The probability of inclusion and the associated weight are respectively:

$$p_{SP,h} = \min\left(1, \frac{H_1 \times N_h}{N}\right), \quad w_{SP,h} = \frac{1}{p_{SP,h}}. \quad (5)$$

Within each selected sub-prefecture, a number m_h of districts were drawn without replacement using systematic PPS. The probability of inclusion and the associated weight are respectively:

$$p_{1,hd} = \min\left(1, \frac{m_h \times N_{hd}}{N_h}\right), \quad w_{1,hd} = \frac{1}{p_{1,hd}}. \quad (6)$$

In each selected district, a simple random sample without replacement of n_{hd} fish farmers was drawn. The probability of inclusion and the associated weight are respectively:

$$p_{2|hd} = \frac{n_{hd}}{N_{hd}}, \quad w_{2|hd} = \frac{N_{hd}}{n_{hd}}. \quad (7)$$

The probability of inclusion of a respondent i is:

$$\pi_i = p_{SP,h} \times p_{1,hd} \times p_{2|hd}, \quad (8)$$

and the associated weight is:

$$w_i = \frac{1}{\pi_i}. \quad (9)$$

The initial design weights were adjusted for non-response. Since 212 of the 220 selected fish farmers completed the survey, the global non-response adjustment factor was 220/212 = 1.038. The final analytical weights were obtained by applying this correction to the initial design weights.

A summary table (**Table 1**) presents for each district: N_{hd}, n_{hd}, the probabilities of inclusion at the three levels, the average probability of inclusion and the average weight.

Table 1. Sampling plan.

SP	District	N_{hd}	n_{hd}	$p_{(SP,h)}$	$p_{(1,hd)}$	$p_{(2\|hd)}$	π_i	w_i
Gouecke	Banzou Nord	10	5	0.575	0.833	0.500	0.240	4.170
Gouecke	Gouecke Centre	11	6	0.575	0.917	0.545	0.288	3.475
Gouecke	Nonah Gouecke	17	8	0.575	1.000	0.471	0.271	3.693
Gouecke	Tamoe	14	7	0.575	1.000	0.500	0.288	3.475
Koropara	Bahaita	8	6	0.708	0.747	0.750	0.396	2.524
Koropara	Boma Nord	8	6	0.708	0.747	0.750	0.396	2.524
Koropara	Koroh	5	5	0.708	0.467	1.000	0.330	3.029
Koropara	Lokoa	10	8	0.708	0.933	0.800	0.528	1.893
Koropara	Lomou Koropara	11	8	0.708	1.000	0.727	0.515	1.943
Pale	Oueye	26	14	0.396	1.000	0.538	0.213	4.687
Pale	Pale	8	4	0.396	0.762	0.500	0.151	6.625
Samoe	Banzou Sud	6	4	1.000	0.526	0.667	0.350	2.854
Samoe	Guela	30	25	1.000	1.000	0.833	0.833	1.200
Samoe	Kolyepoulou	7	5	1.000	0.613	0.714	0.438	2.283
Samoe	Koronta	14	9	1.000	1.000	0.643	0.643	1.556
Samoe	Nyema Sud	17	12	1.000	1.000	0.706	0.706	1.417
Samoe	Weya Sud Samoe	10	6	1.000	0.876	0.600	0.526	1.903
Soulouta	Gbouo	14	10	0.623	1.000	0.714	0.445	2.247
Soulouta	Gouh	8	5	0.623	0.727	0.625	0.283	3.533
Soulouta	Kola	7	4	0.623	0.636	0.571	0.226	4.417
Soulouta	Kpagalaye	5	3	0.623	0.455	0.600	0.170	5.889
Soulouta	Souhoule	7	4	0.623	0.636	0.571	0.226	4.417
Soulouta	Voumou	9	5	0.623	0.818	0.556	0.283	3.533
Yalenzou	Galakpaye	6	6	1.000	0.526	1.000	0.526	1.900
Yalenzou	Galaye Nzerekore	18	11	1.000	1.000	0.611	0.611	1.636
Yalenzou	Gbotoye	18	14	1.000	1.000	0.778	0.778	1.286
Yalenzou	Gnalakpale	5	3	1.000	0.439	0.600	0.263	3.800
Yalenzou	Konia Aviation	4	4	1.000	0.351	1.000	0.351	2.850
Yalenzou	Loukele	10	6	1.000	0.877	0.600	0.526	1.900
Yalenzou	Toulemou	11	7	1.000	0.965	0.636	0.614	1.629

2.4. Data Collection

The survey was conducted from August 20 to October 10, 2025, via KoboToolbox. A pilot study in three locations was used to validate the questionnaire. A preliminary test in three locations confirmed the questionnaire's comprehensibility.

The questionnaire focused on four aspects: the socio-demographic characteristics of fish farmers (age, sex, level of education, experience), and the technical characteristics of the farms (number and surface area of ponds, feeding, fry stocking, sexing, quantity of production).

2.5. Target Variable

The dependent variable used in this study is annual fish production, defined as the total quantity of fish produced per farm over one year and expressed in kilograms. Since this measure is not standardized by pond surface area, input use, or labor, the manuscript consistently uses the term "annual fish production" rather than "fish farming productivity" when referring to the empirical outcome. Annual fish production was dichotomized at the weighted median into two categories: *Weak* (<median) and *High* (≥median). Predictors included: sexing, stocking method, feeding, fertilization, total number of fry, total pond area, experience class, total number of ponds, and fish farmer's age.

2.6. Data Preprocessing and Bivariate Screening

Prior to modelling, the data underwent several preparation steps. Extreme values were handled using Winsorization at the 5th and 95th percentiles [16] [17], in accordance with recommended approaches to limit the influence of outliers while preserving sample representativeness. Categorical variables were recoded into binary indicator variables for logistic regression. The two continuous predictors retained in the final models were rescaled before estimation: the total number of fingerlings was expressed per 100 fingerlings, and total pond surface area was expressed per 100 m^2. This rescaling was introduced to obtain more interpretable odds ratios.

Before fitting the multivariable model, all candidate predictors were examined in survey-weighted bivariate logistic regressions to assess their crude association with annual fish production. The initial set of candidate predictors included sexing practice, stocking method, fish feeding, fertilization, total number of fingerlings, total pond surface area, experience class, total number of ponds, and fish farmer's age. Variables associated with annual fish production at $p < 0.20$ were considered for the multivariable modelling stage, together with variables considered theoretically relevant. Based on this screening step, sexing practice, stocking method, fertilization, experience class, number of fingerlings per 100 units, pond surface area per 100 m^2, and total number of ponds were considered, whereas fish feeding and fish farmer's age were not retained because of weak bivariate associations.

Multicollinearity was assessed using the generalized variance inflation factor

(GVIF). Adjusted GVIF values were below 2 for all predictors retained in the final model, indicating the absence of problematic multicollinearity. Survey weights were included in all estimations.

2.7. Models

Weighted binary logistic regression. Weighted logistic regression was used to identify the key factors associated with the probability of achieving high annual fish production. The model is defined as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \quad (10)$$

where p denotes the probability of high annual fish production, X_i are the explanatory variables, and β_i, the coefficients are estimated by maximum likelihood under survey weights. Candidate predictors selected after the bivariate screening step were entered into the multivariable modelling process. The final model was chosen according to statistical stability, theoretical relevance, parsimony, and the absence of problematic multicollinearity.

Weighted random forest. The random forest model was constructed from an ensemble of 500 decision trees with parameter $mtry = 2$, which provided a reasonable trade-off between bias and variance. Survey weights were incorporated, and variable importance was assessed using the permutation-based Mean Decrease in Accuracy, a method known for its robustness to correlations and its direct connection to predictive performance.

The total sample was randomly split into 70.

2.8. Evaluation

Predictive performance was evaluated on the independent test set for both models. To compare the quality of the binary predictions (High/Low), a confusion matrix was constructed for each of the two models considered (Table 2). The positive reference class corresponds to high annual fish production, while the negative class designates low annual fish production.

Table 2. General structure of the confusion matrix for binary classification.

Predicted/Observed	Low production	High production
Low production	TN	FN
High production	FP	TP

Note: TN = true negatives; FN = false negatives; FP = false positives; TP = true positives. The positive class corresponds to high annual fish production.

From this matrix, several performance indicators were calculated to validate and compare the models. The formulas used are presented in Table 3. The indicators derived directly from the matrix include overall accuracy, sensitivity, specificity, the F1 score, balanced accuracy (BA), and the Youden (*J*) index. The area under the ROC curve (AUC), obtained from the ROC curve, measures the overall discriminatory ability of the model independently of the classification threshold.

Table 3. Performance indicators used for model evaluation.

Indicator	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity/Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
F1 score	$\frac{2\times \text{Precision}\times \text{Sensitivity}}{\text{Precision}+\text{Sensitivity}}$
Balanced accuracy	$\frac{\text{Sensitivity}+\text{Specificity}}{2}$
Youden index J	$\text{Sensitivity}+\text{Specificity}-1$
AUC	Area under the ROC curve

Note: **TP** = true positives; **TN** = true negatives; **FP** = false positives; **FN** = false negatives.

For both models, confusion matrices and threshold-dependent indicators were computed using survey weights. Therefore, the entries of the confusion matrices represent weighted counts and may take non-integer values. Survey-weighted ROC curves and AUC values were also computed on the independent test set using the predicted probabilities and final survey weights. The weighted logistic regression was evaluated at the conventional classification threshold of 0.50, whereas for the weighted random forest, the final classification threshold was set to 0.63, corresponding to the value that maximized Youden's index on the test set.

All analyses were performed using **R** (version 4.5.0) with the survey packages survey [18], ranger [19], caret [20], and WeightedROC.

3. Results

3.1. Descriptive Characteristics of the Sample

Of the 220 fish farmers selected for the survey, 212 completed the questionnaire, corresponding to a response rate of 96.36%. After applying the final survey weights and restricting the analysis to eligible respondents with complete information on annual fish production and the selected explanatory variables, the analytical sample represented a weighted population of approximately 530 producers across the six sub-prefectures of N'Zérékoré. Table 4 shows that annual fish production varies from 30 kg to 205.6 kg, with a weighted median of 81 kg. Nearly half of the fish farmers (48.11%) have low annual fish production, while 51.89% belong to the high annual fish production category.

Table 4. Descriptive statistics of the main continuous variables used in the analysis.

Variable	Average	Standard deviation	Min	Q1	Med.	Q3	Max	IQR
Number of fingerlings/100	10.01	9.86	0.82	3.00	5.00	16.00	33.54	13.00
Pond surface area/100 m^2	19.41	16.96	3.00	6.00	12.00	28.00	59.30	22.00
Annual fish production (kg)	92.93	50.40	30.00	53.00	81.00	120.00	205.60	67.00

Note: IQR = Q3 − Q1. The number of fingerlings and pond surface area were rescaled before modelling in order to obtain interpretable odds ratios. Thus, the corresponding odds ratios refer respectively to an increase of 100 fingerlings and 100 m^2 of pond surface area. Values are rounded to two decimal places for continuous measurements. **Source:** Survey on fish farming in the prefecture of N'Zérékoré (2025), author's calculations.

The average annual fish production (92.93 kg) is close to the median (81 kg), indicating moderate positive skewness (0.899). After weighting, 26.3% of fish farmers recorded low production, 47.2% average production, and 26.5% high production. Thus, nearly three-quarters of producers exhibited modest annual fish production, while approximately one-quarter stood out with superior performance, revealing marked technical and economic heterogeneity among farms.

3.2. Key Factors (Weighted Logistic Regression)

Following the bivariate screening and multivariable adjustment steps, the final weighted logistic regression model retained sexing practice, number of fingerlings per 100 units, pond surface area per 100 m^2, and experience class. Although stocking method, fertilization, and total number of ponds were associated with annual fish production in the bivariate screening step, they were not retained in the final multivariable model because their contribution became limited after adjustment for the retained predictors.

The results in Table 5 indicate that the practice of sexing, the number of fingerlings stocked per 100 units, and total pond surface area per 100 m^2 were significantly associated with high annual fish production. Farmer experience showed a positive but non-significant association in the final model.

Table 5. Determinants of high annual fish production.

Characteristic	Modality	OR	95% CI	p-value	Adj. GVIF
Sexing practice	No	Ref.	–	–	–
	Yes	5.304	[1.742; 16.145]	0.0056	1.232
Number of fingerlings/100	Continuous	1.092	[1.030; 1.157]	0.0053	1.119
Pond surface area/100 m^2	Continuous	1.048	[1.020; 1.078]	0.0019	1.574
Experience class	Beginner (0 - 7 years)	Ref.	–	–	–
	Intermediate (8 - 14 years)	1.570	[0.532; 4.639]	0.3930	1.207
	Experienced (15 years and above)	1.981	[0.505; 7.771]	0.3075	–

Note: OR = odds ratio; CI = confidence interval; GVIF = generalized variance inflation factor. Adjusted GVIF values were computed as $GVIF^{1/(2\times Df)}$. The variables "number of fingerlings" and "pond surface area" were rescaled before modelling; therefore, their odds ratios refer respectively to an increase of 100 fingerlings and 100 m^2 of pond surface area. All adjusted GVIF values are below 2, indicating the absence of problematic multicollinearity among the predictors retained in the final model. **Source:** Fish farming survey in the N'Zérékoré prefecture (2025), author's calculations.

Fish farmers who practice sexing are 5.3 times more likely to achieve high annual fish production than those who do not. Similarly, each additional 100 fingerlings stocked increased the odds of high annual fish production by 9.2%, while each additional 100 m^2 of pond surface area increased the odds by 4.8%. Although intermediate and experienced fish farmers showed higher odds of high annual fish production compared with beginners, these associations were not statistically significant in the final model. Collinearity indices (adjusted GVIF < 2) confirm the absence of problematic multicollinearity.

3.3. Importance of Variables (Weighted Random Forest)

Figure 1 shows the relative importance of variables according to the permutation approach in the weighted random forest. It shows that the number of fingerlings stocked per 100 units was the most important predictor of high annual fish production, with an importance value of 0.0817. It was followed by pond surface area per 100 m^2 (0.0316) and sexing practice (0.0256). Experience class had a much lower importance value (0.0032), suggesting a limited contribution to the overall predictive performance of the model.

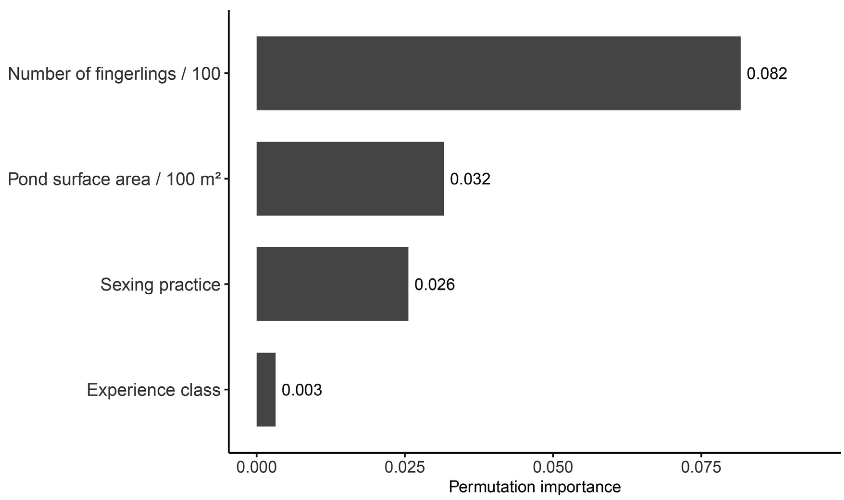

Figure 1. Relative importance of explanatory variables in the weighted random forest.

These differences between models do not reflect a contradiction, but rather illustrate the complementarity between the explanatory effects derived from logistic regression and the overall predictive effects from the random forest model. The convergence on the variables: number of fingerlings stocked, pond surface area, and sexing practice, confirms the robustness and consistency of the identified determinants. The weak contribution of experience class is also consistent with the weighted logistic regression results, where experience showed a positive but non-significant association with high annual fish production.

3.4. Performance Evaluation and Comparison

The performance of the models was evaluated from the weighted confusion matrices, from which the synthetic indicators presented in Table 7 were calculated. In this study, the positive reference class corresponds to high annual fish production (production greater than or equal to the weighted median), while the negative reference class corresponds to low annual fish production.

For both models, the entries of the confusion matrices correspond to survey-weighted counts. Therefore, the values may be non-integer and are directly comparable within the same weighted evaluation framework. The logistic regression and random forest models were evaluated on the same independent test set of 63 observations, using the same final survey weights.

Figure 2 presents the survey-weighted ROC curves of the two models computed on the independent test set. Both models showed strong discriminatory ability, with close weighted AUC values. The weighted logistic regression showed a slightly higher AUC than the weighted random forest (0.861 vs. 0.856), indicating a marginally stronger overall discrimination across all possible thresholds.

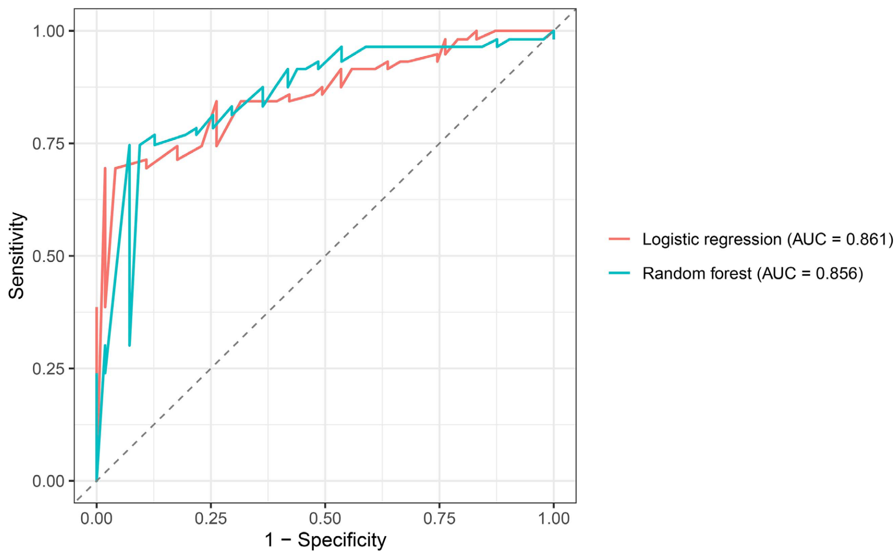

Figure 2. Survey-weighted ROC curves of the weighted logistic regression and weighted random forest models.

Table 6 highlights a strong consistency in the classifications between the two weighted models. The distribution of values shows that the majority of fish farmers are correctly classified within their annual fish production category, reflecting good stability of the predictions and overall reliability of the tested models.

The overall performance, summarized in Table 7, confirms the methodological consistency between the two weighted models.

The weighted logistic regression correctly classified 77.3% of producers in the test set, while the weighted random forest reached an accuracy of 81.7% at the optimized threshold of 0.63. Logistic regression showed higher sensitivity (0.844 vs. 0.746), reflecting a stronger ability to detect high-production fish farmers. In contrast, the random forest exhibited higher specificity (0.906 vs. 0.684), indicating a better identification of low-production producers.

Table 6. Survey-Weighted confusion matrices for the two models.

	WLR		WRF	
Predictions/Realities	**Weak**	**High**	**Weak**	**High**
Predicted: Weak	47.38	13.50	62.69	21.88
Predicted: High	21.85	72.80	6.53	64.41

Note: WLR = weighted logistic regression; WRF = weighted random forest. The positive class is high annual fish production. Because survey weights were applied, the entries of the confusion matrices represent weighted counts and may therefore take non-integer values. The WRF classification threshold was set to 0.63, corresponding to the value that maximized Youden's index.

The random forest also achieved a higher precision (0.908), F1 score (0.819), balanced accuracy (0.826), and Youden index (0.652), while the logistic regression retained a slightly higher survey-weighted AUC (0.861 vs. 0.856). This suggests that the logistic regression model provides slightly stronger overall discrimination across thresholds, whereas the random forest performs better for threshold-based classification after optimization.

Overall, weighted logistic regression stands out for its explanatory power and interpretative transparency, while random forests, thanks to their non-parametric structure, offer greater robustness in the face of farm heterogeneity. This convergence of performance strengthens the reliability of the adopted analytical framework, combining statistical interpretability with the predictive power of machine learning.

Table 7. Comparison of performance indicators: logistic regression and random forest (weighted).

Indicator	Weighted logistic regression	Weighted random forest
Accuracy	0.773	0.817
Sensitivity (recall)	0.844	0.746
Specificity	0.684	0.906
Precision	0.769	0.908
F1 score	0.805	0.819
Balanced accuracy	0.764	0.826
Survey-weighted AUC	0.861	0.856
Youden index (J)	0.528	0.652

4. Discussion

Annual fish production is a key driver of food security and rural development in the Forest Region of Guinea. However, the observed disparities between farms reflect substantial technical and organizational heterogeneity. This study sought to identify the main determinants of annual fish production and to compare the explanatory and predictive power of two survey-weighted statistical approaches.

The results show that weighted logistic regression and weighted random forest converge on the same core variables: the number of fingerlings stocked, total pond surface area, and sexing practice. These factors exert a positive influence on the probability of achieving high annual fish production. In the weighted logistic regression model, sexing practice, the number of fingerlings stocked per 100 units, and pond surface area per 100 m^2 were statistically significant, while farmer experience showed a positive but non-significant association. This convergence reinforces the robustness of the identified determinants and suggests that these variables constitute structural levers of annual fish production rather than context-specific artefacts. Our findings are consistent with those of [21] and [22], who report a positive effect of stocking intensity, farm management practices, and technical conditions on fish production performance, while also showing that pond size may be less beneficial when technical management is suboptimal.

From a methodological standpoint, the comparison between the two models highlights complementary strengths. The weighted logistic regression exhibits excellent discriminatory power (survey-weighted AUC = 0.861; Accuracy = 0.773), providing interpretable effect estimates that are directly usable for targeting interventions, for example, by quantifying how much the odds of high annual fish production increase with sexing practice, additional fingerlings stocked, or larger pond surface area. The weighted random forest yields higher threshold-based classification performance at the optimized threshold of 0.63 (Accuracy = 0.817; Balanced Accuracy = 0.826; Youden Index = 0.652), while maintaining a very good survey-weighted AUC of 0.856. This reflects its ability to capture complex nonlinear relationships and interactions, although at the cost of reduced interpretability compared with logistic regression.

These patterns are in line with recent work, particularly that of [23], who compared logistic regression, random forest, and Support Vector Machines (SVMs) for predicting disease stage from clinical data. Their study showed that combining interpretable parametric models with flexible non-parametric algorithms improves both predictive performance and classification reliability in high-dimensional settings. Our results extend this insight to the context of weighted agricultural survey data, confirming the relevance of a mixed approach for modelling annual fish production under a complex sampling design.

Taken together, these findings underscore the dual contribution of this work. Substantively, they document the central role of stocking intensity, pond surface area, and sexing practice in driving annual fish production in N'Zérékoré. Farmer experience remains relevant from a practical perspective, but its association was not statistically significant in the final weighted logistic regression model and should therefore be interpreted with caution. Methodologically, they illustrate the benefits of integrating survey weights into both classical and machine-learning models to obtain valid inferences and realistic predictive assessments at the population level. This combination of interpretability and predictive performance provides a solid basis for evidence-based decision-making in aquaculture policies and programmes.

5. Conclusions

This study shows that annual fish production in the prefecture of N'Zérékoré is mainly associated with technical and structural factors, particularly sexing practice, the number of fingerlings stocked, and pond surface area. Farmer experience showed a positive but non-significant association in the final weighted logistic regression model and should therefore be interpreted with caution. By combining a weighted logistic regression model with a survey-weighted random forest, we provide a coherent and complementary analytical framework in which the former delivers interpretable effect estimates under a complex multi-stage sampling design, while the latter captures nonlinear relationships and variable interactions that are difficult to specify parametrically.

From a methodological standpoint, this dual modelling strategy substantially improves the analytical reliability of weighted survey data. The excellent discriminatory performance of the weighted logistic regression model (survey-weighted AUC = 0.861; Accuracy = 0.773) and the robust predictive accuracy of the weighted random forest (Accuracy = 0.817 at the optimized threshold of 0.63; survey-weighted AUC = 0.856) confirm the added value of integrating classical statistical modelling with modern machine-learning techniques. These findings are consistent with the results of [23], who demonstrated that combining interpretable parametric models with flexible non-parametric algorithms improves both classification accuracy and decision support in heterogeneous empirical contexts.

Operationally, the results highlight concrete avenues for strengthening aquaculture performance in Guinea, including improved technical training, better control of stocking and pond management practices, and enhanced institutional support for producers. Beyond the specific case of fish farming in N'Zérékoré, this research proposes a reproducible statistical framework for analyzing weighted agricultural survey data in West Africa, thereby offering a methodological contribution of broader relevance for annual fish production assessment in resource-constrained environments.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Amian, A.F., Wandan, E.N., Blé, M.C., Vanga, A.F. and Kaudhjis, P.J.A. (2017) Etude des Determinants Socioeconomiques et Techniques de la Psciculture Extensive en Cote d'Ivoire. *European Scientific Journal, ESJ*, **13**, 389-409. https://doi.org/10.19044/esj.2017.v13n6p389

[2] Bodian, A. (2011) Variabilité pluviométrique et impacts sur les écosystèmes aquatiques. *Revue Ouest Africaine de Climatologie*, **5**, 56-70.

[3] Lazard, J. (2019) Les systèmes aquacoles face au changement climatique. *Cahiers*

Agricultures, **26**, Article No. 34001. https://doi.org/10.1051/cagri/2017018

[4] Chan, C.Y., Tran, N., Pethiyagoda, S., Crissman, C.C., Sulser, T.B. and Phillips, M.J. (2019) Prospects and Challenges of Fish for Food Security in Africa. *Global Food Security*, **20**, 17-25. https://doi.org/10.1016/j.gfs.2018.12.002

[5] Mitra, S., Khan, M.A. and Nielsen, R. (2019) Credit Constraints and Aquaculture Productivity. *Aquaculture Economics & Management*, **23**, 410-427. https://doi.org/10.1080/13657305.2019.1641571

[6] Kimou, N.B., Koumi, R.A., Koffi, M.K., Atsé, C.B., Ouattara, I.N. and Kouamé, P.L. (2016) Utilisation des sous-produits agroalimentaires dans l'alimentation des poissons d'élevage en Côte d'Ivoire. *Cahiers Agricultures*, **25**, Article No. 25006. https://doi.org/10.1051/cagri/2016012

[7] Le Boucher, R., Dupont-Nivet, M., Laureau, S., Labbé, L., Geurden, I., Médale, F., *et al.* (2013) Amélioration génétique et utilisation des aliments à base de végétaux en pisciculture. *INRAE Productions Animales*, **26**, 317-326. https://doi.org/10.20870/productions-animales.2013.26.4.3160

[8] Labbé, L., Lefèvre, F., Bugeon, J., Fostier, A., Jamin, M. and Gaumé, M. (2014) Conception d'un système innovant de production de truites en eau recirculée. *INRAE Productions Animales*, **27**, 135-146. https://doi.org/10.20870/productions-animales.2014.27.2.3061

[9] Vallod, D. and Sarrazin, B. (2010) Caractérisation de l'effluent de vidange d'un étang de pisciculture extensive. *Hydrological Sciences Journal*, **55**, 394-402. https://doi.org/10.1080/02626661003683272

[10] Trintignac, P., Le Berre, M., Haury, J. and Lambert, E. (2020) Les étangs piscicoles, des réservoirs de biodiversité végétale. *Dynamiques environnementales*, **45**, 22-34. https://doi.org/10.4000/dynenviron.3783

[11] N'dri, K.M., Yao, K. and Ibo, G.J. (2016) La pisciculture continentale dans la région du gontougo (Côte d'Ivoire): Caractérisation et aspects socio-économiques. *Tropicultura*, **13**, 300-312.

[12] Alam, M.A., Guttormsen, A.G. and Roll, K.H. (2019) Production Risk and Technical Efficiency of Tilapia Aquaculture in Bangladesh. *Marine Resource Economics*, **34**, 123-141. https://doi.org/10.1086/704129

[13] Milon, A. (2023) État des lieux de la pisciculture en Guinée forestière. Rapport Technique du Ministère de la Pêche, République de Guinée.

[14] Rangé, C. and Pallière, A. (2017) Intégration et Contribution de la Pisciculture au Développement Agricole du Sud de la Guinée Forestière. Technical Report, APDRA Pisiculture Paysanne.

[15] Dabis, F., Desenclos, J.C., *et al.* (2017) Épidémiologie de Terrain (2e Édition): Méthodes et Applications. John Libbey.

[16] Tukey, J.W. (1962) The Future of Data Analysis. *The Annals of Mathematical Statistics*, **33**, 1-67. https://doi.org/10.1214/aoms/1177704711

[17] Cheng, K. and Young, D.S. (2023) An Approach for Specifying Trimming and Winsorization Cutoffs. *Journal of Agricultural, Biological and Environmental Statistics*, **28**, 299-323. https://doi.org/10.1007/s13253-023-00527-4

[18] Lumley, T. (2020) The Survey Package: Analysis of Complex Survey Samples in R. https://CRAN.R-project.org/package=survey

[19] Wright, M.N. and Ziegler, A. (2017) Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**, 1-17. https://doi.org/10.18637/jss.v077.i01

[20] Kuhn, M. (2008) Building Predictive Models in *R* Using the Caret Package. *Journal of Statistical Software*, **28**, 1-26. https://doi.org/10.18637/jss.v028.i05

[21] Gwazani, R., Gandiwa, E. and Poshiwa, X. (2022) Rural Small-Scale Aquaculture: An Assessment of Farmers' Perceptions and Technical Efficiency in Masvingo, Zimbabwe. *Cogent Food & Agriculture*, **8**, Article ID: 2139817. https://doi.org/10.1080/23311932.2022.2139817

[22] Byabasaija, S., Limuwa, M. and Semyalo, R. (2025) Optimizing Small-Scale Aquaculture Systems in the Lake Victoria Basin Uganda through Analysis of Profitability Drivers. *Discover Sustainability*, **6**, Article No. 768. https://doi.org/10.1007/s43621-025-01706-8

[23] Doğançay, M.M., Ege Oruç, Ö., Şırlancı Tüysüzoğlu, M. and Altın, Z. (2024) Prediction of Disease Stage by Machine Learning Classification Methods for Covid-19 Patients. *Austrian Journal of Statistics*, **53**, 1-10. https://doi.org/10.17713/ajs.v53i5.1799