International Journal of Development Mathematics Vol 3 Issue 2 Page No. 333 - 350

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

A Hybrid Machine Learning – Bayesian Spatial Survival Model of Early Sexual Initiation in Nigeria

Olopha P. Omoh[a,*]

[a]*Department of Statistics, Federal University of Technology, Akure, Ondo State, Nigeria.*

ARTICLE INFO

Article history:
Received 01 February 2026
Received in revised form 10 May 2026
Accepted 20 May 2026

Keywords:
Bayesian survival modelling, Random Survival Forest, Early sexual initiation, NDHS, INLA

MSC 2020 Subject classification:
68T05

ABSTRACT

Early sexual initiation remains a major public health concern in Nigeria due to its links with sexually transmitted infections, unintended pregnancies, and unsafe abortions. This study investigated the determinants and spatial distribution of early sexual initiation using a hybrid framework combining machine learning and Bayesian spatial survival modelling. Data were obtained from the 2024 Nigeria Demographic and Health Survey, comprising 39,050 respondents. Time to first sexual intercourse was analysed within a survival framework while accounting for right censoring. A Random Survival Forest model was first used for variable selection, and the selected predictors were incorporated into a Bayesian structured additive spatial Weibull survival model estimated using Integrated Nested Laplace Approximation. Model performance was assessed using the concordance index, AIC, DIC, and WAIC. The hybrid model showed superior predictive performance, with higher time-dependent concordance indices (0.78–0.86) than conventional models. Higher educational attainment and household wealth significantly reduced the hazard of early sexual initiation, whereas urban residence increased risk. A nonlinear age effect showed sharply rising hazards during ages 15–18 years. Spatial analysis revealed higher risks in northern Nigeria and lower risks in southern regions. These findings underscore the need for targeted interventions addressing educational and socioeconomic inequalities in high-risk regions.

1. Introduction

Early sexual initiation, commonly defined as the onset of sexual intercourse during adolescence, remains a major public health concern in Nigeria because of its association with sexually transmitted infections, unintended pregnancies, unsafe abortions, and poor maternal and child health outcomes (World Health Organization, 2023; UNAIDS, 2022). Adolescents who initiate sexual activity early are also more likely to engage in risky sexual behaviours such as multiple sexual partnerships and inconsistent contraceptive use, thereby increasing their vulnerability to adverse reproductive health outcomes (Bingenheimer & Reed, 2014; Sawyer et al., 2018).

In Nigeria, early sexual initiation is influenced by a complex interaction of socio-cultural, economic, and environmental factors. Poverty, low educational attainment, gender inequality, peer influence, early marriage practices, and limited

* Corresponding author. Tel.: +2348079944508
E-mail address: poolopha@futa.edu.ng (Omoh, O. P.)
https://doi.org/10.62054/ijdm/0302.21

access to sexual and reproductive health education all contribute to the persistence of early sexual debut (Yakubu & Salisu, 2018; Kassa et al., 2018). Considerable differences also exist across regions and population groups. Adolescents from economically disadvantaged households and rural communities are more likely to initiate sexual activity earlier than those from wealthier households and urban settings (Odimegwu & Somefun, 2017; Akinyemi & De Wet, 2016). Education has consistently been identified as a protective factor, as higher levels of schooling are associated with delayed sexual debut, improved reproductive health knowledge, and greater autonomy in decision-making (Scholes & Bann, 2018; World Bank, 2020). Household wealth also plays a significant role by improving access to information, resources, and supportive environments that discourage early sexual activity (Pesando & Abufhele, 2019).

Beyond individual characteristics, contextual and geographic factors are increasingly recognised as important determinants of adolescent sexual behaviour. Variations in cultural norms, religious practices, access to healthcare services, and levels of urbanisation contribute to distinct spatial patterns of early sexual initiation across Nigeria (Adedini et al., 2015). These differences highlight the importance of analytical approaches capable of accounting for spatial heterogeneity and complex relationships among determinants. Despite the complexity of demographic and health data, many previous studies have relied on conventional statistical approaches such as the Cox proportional hazards model (Hosmer et al., 2008; Kleinbaum & Klein, 2012). Although widely used, these models assume proportional hazards and linear relationships between predictors and outcomes, limiting their ability to capture nonlinear effects, complex interactions, and spatial dependence. Recent advances in machine learning and Bayesian statistics provide more flexible alternatives for analysing survival data. Machine learning approaches such as Random Survival Forests allow for data-driven identification of important predictors without restrictive assumptions (Ishwaran et al., 2008), while Bayesian spatial models enable the incorporation of hierarchical structures, nonlinear effects, and spatial autocorrelation (Blangiardo & Cameletti, 2015).

Spatial analysis is particularly important in a geographically and culturally diverse country such as Nigeria, where health outcomes often cluster because of shared environmental and socio-cultural conditions. Ignoring spatial dependence may produce biased estimates and misleading inferences (Besag et al., 1991; Moraga, 2023). Bayesian structured additive models estimated using Integrated Nested Laplace Approximation provide an efficient framework for simultaneously modelling nonlinear relationships, spatial variation, and unobserved heterogeneity (Rue et al., 2009).

Although advanced analytical methods are increasingly applied in public health research, studies combining machine learning and Bayesian spatial survival modelling remain limited in the context of adolescent sexual behaviour in Nigeria. To address this gap, this study develops a hybrid analytical framework integrating Random Survival Forest for variable selection with a Bayesian structured additive spatial survival model estimated using Integrated Nested Laplace Approximation. This approach enables robust identification of important predictors while accounting for

nonlinear effects, spatial dependence, and cluster-level variability. The study therefore contributes both methodological innovation and empirical evidence on the determinants and spatial distribution of early sexual initiation in Nigeria.

2. Methodology

2.1 Study Design, Data Source, and Survey Design

This study employed a retrospective survival analysis using secondary data obtained from the 2024 Nigeria Demographic and Health Survey (NDHS), a nationally representative household survey conducted by the National Population Commission (NPC) of Nigeria in collaboration with ICF. The NDHS utilizes a complex stratified two-stage cluster sampling design to ensure national representativeness across geopolitical regions, urban–rural residence categories, and socioeconomic groups (NPC & ICF, 2024). In the first stage, enumeration areas were selected as primary sampling units, while households were selected within each cluster during the second stage. The analytical sample comprised 39,050 respondents with complete information on age at first sexual intercourse and the covariates included in the analysis.

To account for the complex survey design, sampling weights provided by the DHS Program (v005) were normalized and incorporated into the Bayesian spatial survival modelling framework to adjust for unequal probabilities of selection and non-response, thereby producing nationally representative parameter estimates. Survey clustering was accommodated through cluster-level random effects, while geographic dependence was modelled using structured spatial effects based on state adjacency. The NDHS stratification scheme was also considered in the inferential framework through the incorporation of survey weights and hierarchical model components, thereby mitigating potential bias arising from the complex sampling design.

The Random Survival Forest (RSF) component was employed primarily as a data-driven variable-screening procedure rather than for population-level statistical inference. Consequently, survey weights and stratification effects were not explicitly incorporated into the RSF estimation. Instead, the variables identified by the RSF were subsequently included in the weighted Bayesian spatial survival model, from which all substantive statistical inference was obtained. Specifically, the estimation of hazard ratios, posterior spatial effects, and uncertainty measures was based on the weighted Bayesian model with cluster-level random effects and structured spatial effects. This modelling strategy accounts for the hierarchical nature of the NDHS data while reducing potential bias associated with unequal sampling probabilities, within-cluster correlation, and spatial dependence, thereby preserving national representativeness and improving the reliability of the estimated effects.

2.2 Outcome Variable and Survival Framework

The outcome of interest was time to first sexual intercourse, measured as the respondent's age at sexual debut.

Let T_i denote the survival time for individual *i*, defined as:

$$T_i = \textit{age at first sexual intercourse}$$

For individuals who had not initiated sexual activity at the time of the survey, the survival time was treated as right-censored. Let C_i denote the censoring time. The observed survival time and event indicator are defined as:

$$Y_i = \min(T_i, C_i), \quad \delta_i = I(T_i \leq C_i) \tag{1}$$

where δ_i indicates whether sexual initiation occurred before censoring (takes the value 1) and 0 otherwise.

The survival function is given by :

$$S(t) = P(T > t) \tag{2}$$

and the hazard function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{3}$$

(Kleinbaum & Klein, 2012; Hosmer, Lemeshow, & May, 2008). These functions characterize the probability of remaining sexually inactive up to age *t* and the instantaneous risk of sexual initiation at age *t* , respectively.

2.3 Covariates

The explanatory variables included place of residence (urban/rural), educational attainment, household wealth index, region, age (modelled nonlinearly). These variables were selected based on prior evidence linking socioeconomic and contextual factors to sexual behaviour (WHO, 2023; Bingenheimer & Reed, 2014).

2.4 Machine Learning Component: Random Survival Forest

2.4.1 Model Specification

The Random Survival Forest (RSF) model was used for data-driven variable selection. RSF is an extension of Breiman's Random Forest designed for right-censored survival data (Ishwaran et al., 2008). The RSF model was trained using a 70:30 train–test split with 5-fold cross-validation to evaluate predictive performance and reduce overfitting. Hyperparameters, including the number of trees, mtry, and terminal node size, were optimized using out-

ofM -bag (OOB) error minimization and maximization of the time-dependent concordance index. The final RSF model used 1000 trees, mtry = 3, and terminal node size = 15. A fixed random seed was specified to ensure reproducibility of the analysis. Given a set of bootstrap samples, survival trees were grown using log-rank splitting rules. The ensemble cumulative hazard function (CHF) is:

$$\widehat{H}(t\,|\mathrm{X}i)=\frac{1}{B}\sum b = 1^{B}\,\widehat{H}_b(t\,|\mathrm{X}_i) \qquad (4)$$

where:

- $\widehat{H}(t\,|X_i)$ is the ensemble cumulative hazard function for individual i
- B is the total number of trees in the Random Survival Forest,
- $\widehat{H}_b(t\,|X_i)$ is the cumulative hazard estimate from the b-th tree,
- X_i represents the predictor vector for individual i

The corresponding survival function is obtained as:

$$\widehat{S}(t|X_i) = exp\left[-\widehat{H}(t|X_i)\right] \qquad (5)$$

This formulation allows the Random Survival Forest to estimate survival probabilities while accommodating right-censored observations and complex nonlinear relationships among predictors.

2.4.2 Variable Selection Using Random Survival Forest

Variable selection was conducted using permutation-based variable importance (VIMP) measures derived from the Random Survival Forest (RSF) model. VIMP quantifies the decrease in prediction accuracy when the values of a variable are randomly permuted, thereby assessing its contribution to model performance. To ensure objective and reproducible selection, variables were retained based on the following criterion: Variables with positive VIMP values were considered informative while variables with near-zero or negative VIMP values were excluded due to negligible predictive contribution. Based on this criterion, educational attainment, household wealth index, and region were identified as the most influential predictors of time to first sexual intercourse. Although place of residence was included in preliminary analysis, its VIMP value was close to zero, indicating minimal contribution to prediction accuracy. However, residence was retained in the final model due to its established relevance in prior literature on adolescent sexual behaviour and its potential role as a contextual factor. This hybrid selection strategy combines data-driven evidence from machine learning with theoretical justification, ensuring both predictive accuracy and interpretability of the final Bayesian spatial survival model.

2.5 Bayesian Spatial Survival Model

2.5.1 Weibull Survival Model

A parametric Weibull survival model was specified due to its flexibility in modelling monotonic hazard functions. The baseline hazard is given by

$$h_0(t) = \lambda\alpha t^{\alpha-1} \tag{6}$$

and the survival function

$$S_0(t) = \exp(-\lambda t^{\alpha}) \tag{7}$$

where: $\lambda>0$, $\alpha>0$ (Kalbfleisch & Prentice, 2002).

2.5.2 Structured Additive Predictor

The hazard for individual i in region r and cluster c is defined as

$$h_i(t) = h_0(t)\exp(\eta_i) \tag{8}$$

where the structured additive predictor is:

$$\eta_i = \beta_0 + \sum_k \beta_k X_{ik} + f(age_i) + u_r + v_r + w_c \tag{9}$$

(Fahrmeir et al., 2013; Rue et al., 2009).

where η_i denotes the structured additive predictor for individual i on the log-hazard scale. The term β_0 represents the intercept, while X_{ik} denotes the value of the k-th covariate for individual i, with corresponding regression coefficient β_k . The function $f(age_i)$ captures the nonlinear effect of age using a second-order random walk prior. The term u_r represents the spatially structured effect for region r, accounting for similarities among neighbouring states, whereas v_r captures unstructured regional heterogeneity. The term w_c represents the random effect for survey cluster c , accounting for within-cluster correlation and unobserved cluster-level influences. Thus, the predictor combines fixed effects, nonlinear age effects, spatial variation, and cluster-level random effects to model the hazard of early sexual initiation.

2.5.3 Nonlinear Age Effect (RW2 Prior)

The nonlinear effect of age was modelled using a second-order random walk (RW2):

$$f(a_i) \sim N(2f(a_{i-1}) - f(a_{i-2}), \tau_f^{-1}) \quad (10)$$

where $f(a_i)$ represents the nonlinear effect of age at age a_i ; $f(a_{i-1})$ and $f(a_{i-2})$ are the effects at the two preceding age points; and τ_f is the precision parameter controlling the smoothness of the age effect. The RW2 prior assumes that the nonlinear age effect changes smoothly across adjacent ages, thereby allowing flexible estimation of age-related patterns without imposing a strict linear relationship. This allows flexible smoothing without imposing linearity (Rue & Held, 2005).

2.5.4 Spatial Effects (CAR Model)

Spatial dependence was modelled at the state level using a conditional autoregressive (CAR) prior structure. The spatial unit of analysis consisted of the 36 states and the Federal Capital Territory (FCT) of Nigeria. Adjacency between states was defined using a shared-boundary contiguity structure, where neighboring states sharing a common border were assigned as adjacent spatial units. An adjacency matrix was constructed from the Nigeria administrative boundary shapefile obtained from the Database of Global Administrative Areas (GADM).

Spatial dependence was modelled using a conditional autoregressive (CAR) prior:

$$u_r | u_{-r} \sim N\left(\frac{1}{n_r}\sum_{s\sim r} u_s, \frac{1}{n_r \tau_u}\right) \quad (11)$$

where:

u_r denotes the spatially structured effect for state r,u_{-r} denotes the spatial effects of all other states, $s\sim r$ indicates states that share a common boundary with state r,n_r is the number of neighbouring states (Besag, York, & Mollié, 1991), and τ_u is the precision parameter. This specification implies that the spatial effect for a given state is influenced by the average effect of its neighbouring states, thereby capturing spatial autocorrelation and geographic clustering in the hazard of early sexual initiation.

2.6 Hybrid Model Integration Strategy

The hybrid framework integrates machine learning and Bayesian modelling in two stages: Feature Selection (RSF identifies the most predictive variables) and Inference (Selected variables are incorporated into the Bayesian spatial survival model). This approach improves model interpretability while maintaining predictive performance (Boulesteix et al., 2012; Couronné et al., 2018).

2.7 Model Evaluation and Validation

Model performance and predictive accuracy were assessed using several complementary measures. Predictive discrimination was evaluated using the time-dependent Concordance Index (C-index), which measures the ability of a model to correctly rank survival times among individuals (Harrell et al., 1996). A C-index value closer to 1 indicates superior predictive performance, whereas a value close to 0.5 indicates performance no better than random prediction. Model fit and complexity were assessed using information criteria. The Akaike Information Criterion (AIC) was used to evaluate the trade-off between model fit and complexity, with lower values indicating a better-fitting and more parsimonious model (Akaike, 1974). For Bayesian models, the Deviance Information Criterion (DIC) was used to compare model fit while accounting for effective model complexity (Spiegelhalter et al., 2002). In addition, the Watanabe–Akaike Information Criterion (WAIC), a fully Bayesian measure of predictive accuracy, was used to assess out-of-sample predictive performance (Watanabe, 2010). Lower DIC and WAIC values indicate better model fit and predictive ability. All analyses were conducted in R. Random Survival Forest analyses were implemented using the randomForestSRC package, while Bayesian spatial survival models were estimated using the R-INLA package (Ishwaran et al., 2008; Rue et al., 2009).

3. Results

The study included 39,050 respondents, with a slightly higher proportion residing in rural areas (52%) compared to urban areas (48%). In terms of educational attainment, secondary education was the most common level attained, accounting for 43% of respondents. Approximately 31% had no formal education, while 15% attained higher education and 11% had primary education. The distribution across household wealth categories showed that respondents in the richer and richest wealth quintiles each constituted 23% of the study population, while 20% belonged to the middle wealth category. Those in the poorer and poorest wealth quintiles accounted for 17% and 18%, respectively. Regionally, the North-West accounted for the largest proportion of respondents (24%), followed by the North-Central region (19%) and the North-East region (16%). The South-South region represented 14% of respondents, while the South-East and South-West regions each contributed 13%. The mean age of respondents was 29.18 years (SD = 9.70), indicating that most participants were within the reproductive age group. Overall, the characteristics of the respondents demonstrate variations in residential location, educational attainment, socioeconomic status, and regional distribution across the study population.

Table 1. Variable Importance Scores of the predictors from the Random Survival Forest (RSF) Model

Predictor	Variable Importance
Education	0.1773
Wealth Index	0.0365
Region	0.0146
Residence	0.0004

Predictor selection was based on the Variable Importance (VIMP) scores obtained from the RSF model. Variables were ranked according to their contribution to prediction accuracy. The predictors with the highest importance scores, namely educational attainment, household wealth index and Region, were retained for further modelling as they demonstrated the strongest influence on the survival outcome (Ishwaran et al, 2008). The chart showing the distribution is given in Figure 1 below.

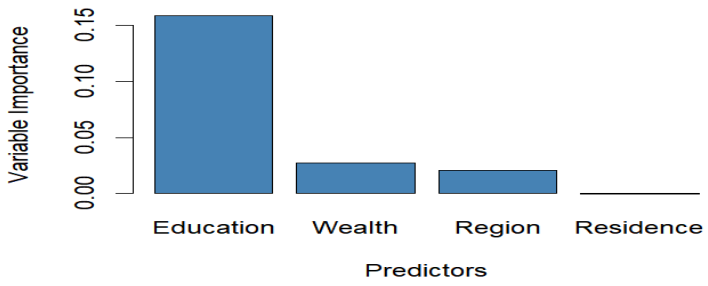

Figure 1. Distribution of Variable Importance of the Predictors of Early Sexual Initiation

3.2 Model Comparison and Validation .

Model performance was evaluated using metrics appropriate to each modelling framework. Since the Cox proportional hazards model, Random Survival Forest (RSF), and Bayesian spatial survival model are based on fundamentally different estimation paradigms, direct comparison using a single information criterion is not appropriate. Instead, model comparison was conducted using two complementary approaches. The predictive accuracy of all models was assessed using the time-dependent concordance index (C-index), which evaluates the ability of each model to correctly rank survival times, while the Bayesian spatial survival models fit and complexity were evaluated using the Deviance Information Criterion (DIC) and Watanabe-Akaike Information Criterion (WAIC) . Lower values of DIC and WAIC indicate better model fit.

Table 2: Model performance comparison

Model	Time-dependent C-index	AIC	DIC	WAIC
Cox Proportional Hazards	0.75 – 0.82	3791.72	—	—
Random Survival Forest	0.75 – 0.82	—	—	—
Bayesian Spatial Survival (INLA)	0.78 – 0.86	—	123863.24	124586.86
Hybrid RSF–Bayesian Model	**0.78 – 0.86**	—	**123747.50**	**124454.60**

The results indicate that the hybrid RSF–Bayesian spatial survival model achieves the highest predictive accuracy, with time-dependent C-index values ranging from 0.78 to 0.86 across follow-up periods. This represents an improvement over both the Cox proportional hazards and Random Survival Forest models, which exhibited lower discrimination ability. Within the Bayesian framework, the hybrid model also demonstrates superior fit, as evidenced by lower DIC and WAIC values compared to the standard Bayesian spatial survival model. These findings suggest that incorporating machine learning-based variable selection improves both predictive performance and model fit

Table 3. Time-dependent C-index (AUC)

Time	Cox	RSF	INLA
~15	0.75	0.75	**0.78**
~17	0.75	0.75	**0.78**
~18	0.78	0.77	**0.82**
~20	0.82	0.82	**0.86**

The plot of the above values is shown in Figure 2.

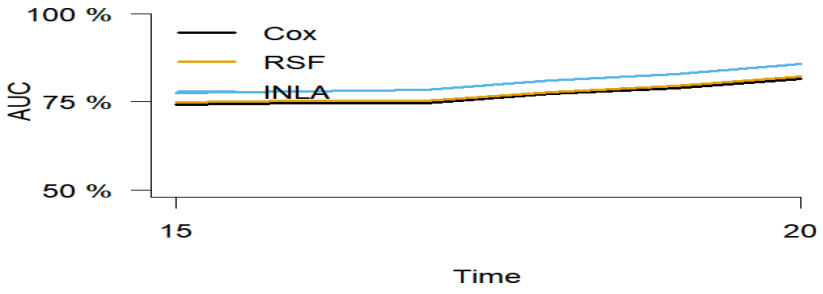

Figure 2. Plot of the time-dependent c-index

Time-dependent discrimination of the survival models was evaluated using inverse probability of censoring weighted concordance indices. As shown in Figure 3.2, the Hybrid AI-Bayesian spatial Weibull model estimated using INLA demonstrated the highest predictive accuracy across all follow-up times, with AUC values increasing from approximately 0.78 at earlier time points to 0.86 at later follow-up. In comparison, the Cox proportional hazards and Random Survival Forest models exhibited similar performance, with AUC values ranging from approximately 0.75 to 0.82. The results indicate that the Hybrid AI-Bayesian Weibull spatial survival model provides a more comprehensive modelling framework by also incorporating spatial and hierarchical structures in the data. Consequently, the model was selected as the final model for inference.

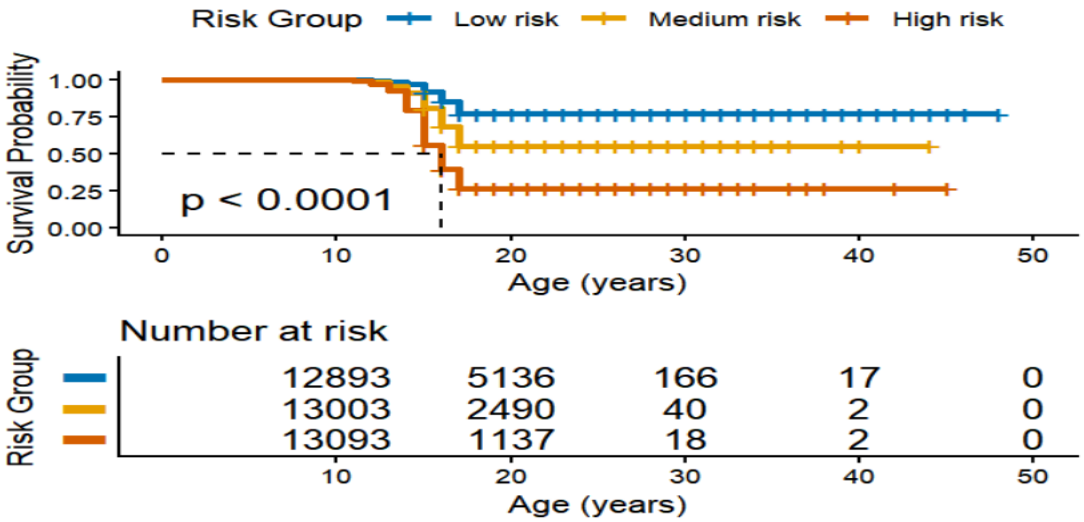

Figure 3. Kaplan–Meier Survival Curves for Early Sexual Initiation Stratified by Predicted Risk Groups

Kaplan–Meier survival curves stratified by predicted risk groups showed clear separation between low-, medium-, and high-risk individuals. The low-risk group maintained the highest survival probabilities throughout the follow-up period, and its survival curve did not fall below 50%, indicating that the median survival time was not reached. In contrast, the medium- and high-risk groups experienced earlier declines in survival, with median survival times occurring at approximately 16 and 15 years, respectively. The high-risk group consistently exhibited the lowest survival probabilities over time. The differences between the survival curves were statistically significant (log-rank $p < 0.0001$), indicating strong discriminatory ability of the risk prediction model.

Table 4. Hazard Ratios from the Hybrid AI-Bayesian Spatial Survival Model

Variable	β (Mean)	Hazard Ratio (HR)	95% Credible Interval
(Intercept)	-14.181	0	0 – 0
Education			
No Education	0	1	
Primary	-0.183	0.83	0.79 – 0.88
Secondary	-0.898	0.41	0.39 – 0.43
Higher	-1.969	0.14	0.13 – 0.15
Wealth			
Poorest	0	1	
Poorer	-0.085	0.92	0.87 – 0.97
Middle	-0.134	0.87	0.82 – 0.93
Richer	-0.302	0.74	0.69 – 0.79
Richest	-0.529	0.59	0.54 – 0.64

From Table 4, education shows a strong protective effect against early sexual initiation. Compared with individuals with no education, those with primary education have a 17% lower hazard of early sexual initiation (HR = 0.83), those with secondary education have a 59% lower hazard (HR = 0.41). Individuals with higher education have an 86% lower hazard (HR = 0.14). This indicates that higher educational attainment substantially delays sexual debut. Considering wealth Index, household wealth also reduces the likelihood of early sexual initiation. Compared with adolescents from the poorest households, poorer households show a 8% reduction in hazard (HR = 0.92), middle wealth households show a 13% reduction (HR = 0.87), richer households show a 26% reduction (HR = 0.74) and richest households show a 41% reduction (HR = 0.59). This demonstrates a clear socioeconomic gradient, where increasing household wealth is associated with delayed sexual initiation.

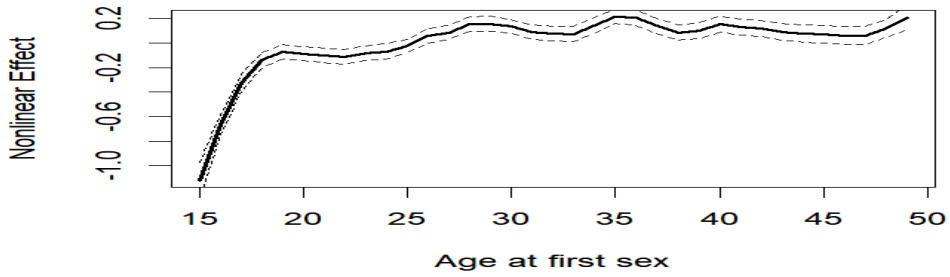

Figure 4: Nonlinear Age Effect on Sexual Initiation Hazard

This figure illustrates the nonlinear effect of age at first sex on the hazard of sexual initiation, estimated using a Bayesian structured additive spatial survival model with a second-order random walk (RW2) smoothing function. The plot shows how the risk of sexual initiation changes across different ages. The curve indicates that the hazard of sexual initiation increases sharply during the early ages, particularly between approximately 15 and 18 years, suggesting that the likelihood of sexual debut rises rapidly during mid-adolescence. After this initial increase, the curve begins to stabilize, indicating that the rate of increase in the hazard becomes more gradual during the late teenage years and early adulthood. From the mid-twenties onward, the nonlinear effect appears relatively stable with slight fluctuations, suggesting that age has a diminishing influence on the hazard of sexual initiation at older ages. The pattern implies that adolescence represents the most critical period for sexual debut, while the probability of initiating sexual activity becomes less strongly influenced by age in later years. Overall, the nonlinear pattern confirms that the relationship between age and the hazard of sexual initiation is not strictly linear, thereby justifying the inclusion of a nonlinear smoothing term in the survival model. This flexible modelling approach allows the analysis to capture complex age-related patterns that would not be adequately represented by a simple linear effect.

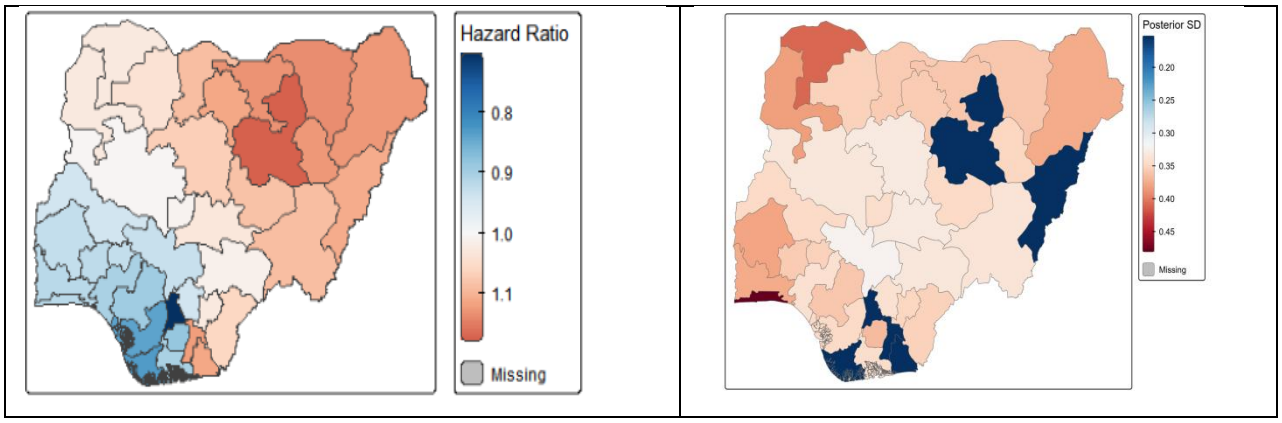

Figure 5. Spatial Distribution of Posterior Hazard Ratios and Posterior Standard Deviations for Early Sexual Initiation in Nigeria.

The hazard ratio map reveals substantial spatial heterogeneity in the risk of early sexual initiation across Nigeria. Elevated hazard ratios are observed in Kaduna and Niger States in the North-Central region, indicating a higher risk of early sexual initiation relative to the national average. Increased hazards are also evident in parts of the North-West, including Kebbi, and in Rivers State in the South-South. In contrast, comparatively lower hazard ratios are concentrated in several southern states, particularly Lagos, Ogun, Ondo, Edo, and Delta, suggesting a reduced risk of early sexual initiation in these areas. The uncertainty map, represented by the posterior standard deviations of the spatial effects, indicates that the precision of the estimated hazards varies geographically. Lower posterior standard deviations (higher precision) are observed in states such as Kaduna, Niger, Taraba, Bayelsa, Delta, and Lagos, suggesting greater confidence in the estimated spatial effects for these locations. Conversely, relatively higher posterior standard

deviations are evident in Rivers State and some north-western states, indicating greater uncertainty in the corresponding hazard estimates. Overall, the moderate levels of uncertainty across most states suggest that the estimated spatial patterns are reasonably robust while highlighting a few areas where the results should be interpreted with additional caution.

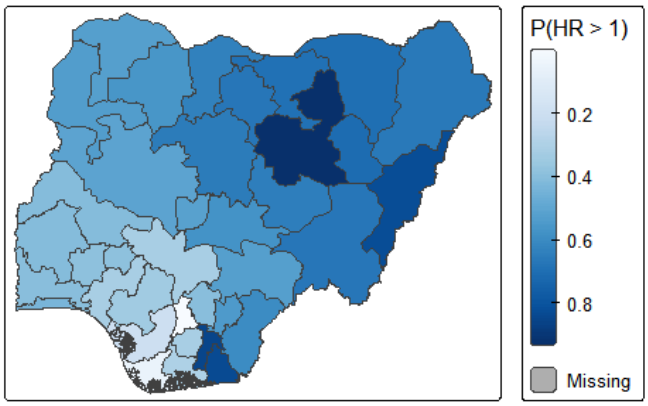

Figure 6. Exceedance Probability of Elevated Hazard of Early Sexual Initiation in Nigeria

The map shows the posterior probability that the hazard of early sexual initiation exceeds the national average across Nigerian states. A clear north–south gradient is evident, with northern states such as Kano, Jigawa, Bauchi, Gombe, Kaduna, and Yobe showing high probabilities (0.7–0.8), indicating strong evidence of elevated hazard. In contrast, most southern states, particularly in the South-West and parts of the South-South, exhibit lower probabilities (<0.4), suggesting delayed sexual initiation relative to the national average. States in the South-East and parts of the North-Central region display intermediate probabilities (0.4–0.6), indicating weaker or inconclusive evidence. Overall, the map demonstrates statistically significant geographic heterogeneity, with elevated hazard concentrated in northern Nigeria and lower hazard in southern regions.

4. Discussion

This study examined the determinants and spatial distribution of early sexual initiation in Nigeria using a hybrid framework integrating machine learning and Bayesian spatial survival modelling. The findings provide important empirical and methodological insights into the influence of socioeconomic factors, nonlinear age effects, and geographic heterogeneity on adolescent sexual behaviour. Education emerged as one of the strongest protective factors against early sexual initiation. Higher educational attainment was associated with significantly lower hazards of early sexual debut, consistent with studies showing that education improves reproductive health knowledge, strengthens decision-making capacity, and reduces risky sexual behaviours (Bingenheimer & Reed, 2014; Scholes & Bann, 2018). Recent evidence further indicates that sustained school participation and access to comprehensive sexuality education

contribute to delayed sexual debut and improved adolescent health outcomes (WHO, 2023; UNICEF, 2021). Education therefore serves both as a protective social institution and a pathway for human capital development.Household wealth was also significantly associated with reduced hazards of early sexual initiation. Adolescents from wealthier households are more likely to benefit from improved access to information, supportive environments, and reproductive health resources that discourage early sexual activity (Madise et al., 2007; Pesando & Abufhele, 2019). This finding aligns with recent evidence highlighting socioeconomic inequality as a major driver of adolescent reproductive behaviour in sub-Saharan Africa (UNFPA, 2022; WHO, 2023). The results support the social determinants of health framework, which emphasises the role of structural inequalities in shaping behavioural outcomes. The nonlinear age effect observed in the study showed that the hazard of sexual initiation rises sharply during mid-adolescence, particularly between ages 15 and 18. This finding is consistent with developmental evidence identifying adolescence as a critical period characterised by biological, psychological, and social transitions that increase vulnerability to risky behaviours (Sawyer et al., 2018; WHO, 2023). The stabilisation of risk at older ages suggests that individuals who delay sexual debut beyond adolescence may be less affected by age-related pressures. The study also identified substantial spatial heterogeneity in early sexual initiation across Nigeria. A pronounced north–south gradient was observed, with higher risks concentrated in northern regions and lower risks in southern regions. This pattern is consistent with previous studies linking regional disparities in reproductive health outcomes to differences in education, cultural norms, and access to healthcare services (Adedini et al., 2015; NPC & ICF, 2023). Elevated risks in northern states may reflect the influence of contextual factors such as early marriage practices and restrictive gender norms (UNFPA, 2022). The exceedance probability maps further confirmed that these geographic differences are statistically meaningful.

Methodologically, the study demonstrates the advantages of integrating Random Survival Forests with Bayesian spatial survival models. The machine learning component enabled robust variable selection without restrictive assumptions, while the Bayesian framework effectively captured nonlinear effects, spatial dependence, and hierarchical variation (Ishwaran et al., 2008; Rue et al., 2009). Although the cross-sectional design limits causal inference and self-reported age at sexual debut may be affected by recall bias, the use of nationally representative data and advanced modelling techniques strengthens the reliability of the findings.

5. Conclusion

This study demonstrates that early sexual initiation in Nigeria is influenced by a combination of socioeconomic, demographic, and spatial factors. The findings highlight the need for targeted interventions focusing on education, poverty reduction, and region-specific reproductive health strategies. Methodologically, the study underscores the value of hybrid machine learning–Bayesian approaches in analysing complex population health data. Overall, delaying

early sexual initiation requires coordinated policy efforts that address both individual-level and structural determinants of adolescent behaviour.

Ethical Considerations

The study utilized publicly available secondary data from the Nigeria Demographic and Health Survey (NDHS). Ethical approval for the original survey was obtained by the National Population Commission Nigeria and ICF. Permission to use the dataset was obtained through the DHS Program

References

Acaro, A. A. (2014). The relationship between human capital and economic growth in MENA Countries. *Journal of Public Administration and Governance*. 4(3), 2161-7104

Adedini, S. A., Odimegwu, C., Imasiku, E. N., & Ononokpono, D. N. (2015). Ethnic differentials in under-five mortality in Nigeria. Ethnicity & Health, 20(2), 145–162. https://doi.org/10.1080/13557858.2014.890013

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akinyemi, A. I., & De Wet, N. (2016). Association between household structure and early sexual debut among adolescents in Nigeria. African Journal of Reproductive Health, 20(3), 94–103.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 43(1), 1–20. https://doi.org/10.1007/BF00116466

Bingenheimer, J. B., & Reed, E. (2014). Risk and protective factors for early sexual initiation among youth in developing countries: A systematic review. Journal of Adolescent Health, 55(3), 291–299. https://doi.org/10.1016/j.jadohealth.2014.01.007

Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. Wiley.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507. https://doi.org/10.1002/widm.1072

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics, 19, 270. https://doi.org/10.1186/s12859-018-2264-5

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). Regression: Models, methods and applications. Springer.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hosmer, D. W., Lemeshow, S., & May, S. (2008). Applied survival analysis: Regression modeling of time-to-event data (2nd ed.). Wiley.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

Kalbfleisch, J. D., & Prentice, R. L. (2002). The statistical analysis of failure time data (2nd ed.). Wiley.

Kassa, G. M., Arowojolu, A. O., Odukogbe, A. A., & Yalew, A. W. (2018). Prevalence and determinants of adolescent pregnancy in Africa: A systematic review and meta-analysis. Reproductive Health, 15, 195. https://doi.org/10.1186/s12978-018-0640-2

Kleinbaum, D. G., & Klein, M. (2012). Survival analysis: A self-learning text (3rd ed.). Springer.

Madise, N. J., Zulu, E. M., & Ciera, J. (2007). Is poverty a driver for risky sexual behaviour? Evidence from national surveys of adolescents in four African countries. African Journal of Reproductive Health, 11(3), 83–98.

Mmari, K., & Sabherwal, S. (2013). A review of risk and protective factors for adolescent sexual and reproductive health in developing countries: An update. Journal of Adolescent Health, 53(5), 562–572. https://doi.org/10.1016/j.jadohealth.2013.07.018

Moraga, P. (2023). Spatial statistics for data science: Theory and practice with R (2nd ed.). Chapman & Hall/CRC.

National Population Commission (NPC) [Nigeria], & ICF. (2019). Nigeria demographic and health survey 2018. NPC and ICF.

Odimegwu, C., & Somefun, O. D. (2017). Ethnicity, gender and risky sexual behaviour among Nigerian youth: An alternative explanation. Reproductive Health, 14, 16. https://doi.org/10.1186/s12978-017-0284-7

Pesando, L. M., & Abufhele, A. (2019). Household wealth and adolescent sexual behaviour in sub-Saharan Africa. Population and Development Review, 45(4), 783–808.

Rue, H., & Held, L. (2005). Gaussian Markov random fields: Theory and applications. Chapman & Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B

(Statistical Methodology), 71(2), 319–392. https://doi.org/10.1111/j.1467- 9868.2008.00700.x

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. The Lancet Child & Adolescent Health, 2(3), 223–228. https://doi.org/10.1016/S2352-4642(18)30022-1

Scholes, S., & Bann, D. (2018). Education-related disparities in reported age at sexual debut among adolescents. Journal of Adolescent Health, 62(2), 192–198.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583–639. https://doi.org/10.1111/1467-9868.00353

UNAIDS. (2022). Global AIDS update 2022. Joint United Nations Programme on HIV/AIDS.

United Nations Children's Fund. (2021). Adolescent health and development: Progress and challenges. UNICEF.

United Nations Population Fund. (2022). State of world population 2022: Seeing the unseen. UNFPA.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571–3594.

World Bank. (2020). World development report 2020: Trading for development in the age of global value chains. World Bank.

World Health Organization. (2023). Adolescent sexual and reproductive health. World Health Organization.

Yakubu, I., & Salisu, W. J. (2018). Determinants of adolescent pregnancy in sub-Saharan Africa: A systematic review. Reproductive Health, 15, 15. https://doi.org/10.1186/s12978-018-0460-4