Non-Parametric Estimation of the Posterior Distribution in Monitoring Primary School Enrollment in Mt. Elgon Region, Kenya

Abstract

This study presents an innovative Bayesian non-parametric framework for monitoring primary school enrollment in developing regions, with application to Kenya's Mt. Elgon region. The study addresses critical limitations of conventional parametric methods through Bayesian Additive Regression Trees (BART), which captures complex enrollment patterns while providing probabilistic uncertainty quantification. Analyzing complete 2023 administrative data from all 53 public primary schools in the region, the approach reveals three key insights: research findings reveal near gender parity (51% boys, 49% girls) but significant disparities in Grade 2 and declining retention in middle grades (4–6)). Secondly, the research identifies a concerning middle-grade attrition pattern with Grades 4-6 showing an 18% enrollment drop. Third, posterior distributions reveal stable enrollment clusters centered at 490 students per school (95% CI: 425-555). The model demonstrates strong predictive performance (RMSE = 188.52, MAE = 147.39) while outperforming conventional methods by 12-51% in accuracy metrics. These findings provide education planners with a robust decision-support tool for targeted resource allocation, particularly for addressing gender-specific retention challenges and middle-grade attrition. The methodology offers a scalable solution for educational monitoring in similar resource-constrained settings across sub-Saharan Africa.

Keywords: Bayesian nonparametrics, Educational monitoring, BART, Enrollment prediction, Kenya, Primary education

1 Introduction

Accurate monitoring of primary school enrollment remains a fundamental challenge for education systems across sub-Saharan Africa. In Kenya's Mt. Elgon region - a geographically fragmented area bordering Uganda - this challenge is exacerbated by complex interactions between socioeconomic factors, cultural practices, and educational policies. Parametric approaches, while computationally efficient, systematically underestimate enrollment fluctuations by 15-20% in

similar Kenyan contexts **uwezo2018**, primarily due to their inability to capture nonlinear patterns and threshold effects in grade transition rates **burgess2019**.

The educational landscape of Mt. Elgon presents four distinctive monitoring challenges: First, extreme geographic dispersion with schools serving widely scattered populations across difficult terrain. Second, seasonal migration patterns that disrupt regular attendance, particularly during planting and harvest seasons. Third, cross-border dynamics where students frequently transfer between Kenyan and Ugandan schools. Fourth, uneven implementation of Kenya's Free Primary Education policy across different communities in the region.

Recent methodological advances in Bayesian non-parametric offer promising solutions to these challenges **Kololi2017**. Bayesian Additive Regression Trees (BART), introduced by **chipman2010**, combines the flexibility of machine learning with principled uncertainty quantification from Bayesian statistics. Unlike conventional methods, BART automatically adapts to complex data patterns through an ensemble of regression trees, each contributing a small part to the overall prediction. This characteristic makes it particularly suitable for educational data that often exhibits nonlinearities, interactions, and heteroscedasticity **linero2018**.

Our study advances both methodological and practical dimensions of educational monitoring through four substantive contributions. First, we pioneer the application of Bayesian Additive Regression Trees (BART) to educational enrollment monitoring, adapting this powerful technique from its traditional clinical and economic applications **xia2023** to address critical gaps in school administration. Second, we conduct the first comprehensive grade-level analysis of enrollment patterns across Mt. Elgon's entire public primary system, examining 25,824 students to reveal previously undocumented trends. Third, we develop a posterior estimation framework that not only handles real-world data challenges like missing observations but also generates probabilistic enrollment projections with quantified uncertainty. Finally, and most significantly, we translate these technical advances into actionable policy recommendations, particularly for addressing the persistent gender disparities in Grade 2 (gender ratio = 1.12) and the troubling middle-grade attrition rates (14.3 students average decline between Grades 4-6), providing education planners with evidence-based strategies for targeted interventions.

2 Methodology

2.1 Data Characteristics and Preparation

We analyzed the complete 2023 enrollment records from all 53 public primary schools in Mt. Elgon region, obtained through formal collaboration with the Kenyan Ministry of Education.

The dataset exhibited two main data quality issues requiring preprocessing: First, Grade 7 records were missing for 6 schools (12% of cases) due to administrative reporting delays. We addressed this through cubic spline interpolation,

which preserved the nonlinear trends observed in adjacent grades. Second, Early Childhood Development (ECD) levels (PP1 and PP2) lacked gender disaggregation in original records. We imputed these using a k-nearest neighbors approach (k=5) based on gender patterns in subsequent grades and similar schools.

Data quality validation confirmed an overall missing value rate below 3%, which we handled through multiple imputation by chained equations (MICE). For computational efficiency, we aggregated daily attendance records to monthly averages, while preserving grade-level and gender-specific details essential for our analysis.

2.2 Bayesian Additive Regression Trees Framework

The BART model provides a flexible nonparametric approach to regression by combining an ensemble of weak learners (regression trees) under a Bayesian framework. For enrollment count Y_i at school i with predictors X_i, the model specification is:

$$Y_i = \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where each component has specific meaning:

$g(X_i; T_j, M_j)$: The j^{th} regression tree's prediction, T_j: Binary tree structure with decision rules, $M_j = \{\mu_{1j}, ..., \mu_{bj}\}$: Terminal node parameters and ϵ_i is the Independent Gaussian error term

After extensive cross-validation, we set the number of trees $m = 200$ to balance model flexibility and computational efficiency. The model specification includes three key priors:

$$\begin{aligned} &\text{Tree depth: } p(T_j) \propto (1+d)^{-\beta} \text{ with } \beta = 2 \\ &\text{Terminal nodes: } \mu_{ij} \sim \mathcal{N}(0, 2.25/m) \\ &\text{Error variance: } \sigma^2 \sim \text{Inv-Gamma}(3/2, 3/2) \end{aligned} \tag{2}$$

These priors serve two crucial functions: First, they regularize the model to prevent overfitting to the training data. Second, they ensure each tree contributes only a small part to the final prediction, maintaining the "weak learner" property essential for BART's success **chipman2010**.

2.3 Model Estimation and Inference

Posterior inference was conducted through Markov Chain Monte Carlo (MCMC) sampling with a Gibbs sampling scheme. The complete posterior distribution takes the form:

$$\pi(\theta|D) \propto \left[\prod_{i=1}^{n} \mathcal{N}(y_i | \sum_{j=1}^{m} g(x_i; T_j, M_j), \sigma^2)\right] \times \left[\prod_{j=1}^{m} \pi(T_j)\pi(M_j)\right] \pi(\sigma) \tag{3}$$

where $\theta = \{T_1, M_1, ..., T_m, M_m, \sigma\}$ represents all model parameters.

2.4 Performance Evaluation Metrics

We evaluated model performance using three complementary metrics:

1. **Root Mean Squared Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 188.52 \tag{4}$$

2. **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = 147.39 \tag{5}$$

3 Results

3.1 Enrollment Patterns and Trends

Analysis of the complete 2023 enrollment data (25,824 students across 53 schools) revealed several important patterns, summarized in Table 1.

Table 1: Grade-wise Enrollment Statistics and Trends

Grade	Boys	Girls	Total	% Change	Gender Ratio
PP1	1,041	977	2,018	-	1.07
Grade 1	1,385	1,310	2,695	+33.5%	1.06
Grade 2	1,476	1,322	2,798	+3.8%	1.12
Grade 3	1,418	1,398	2,816	+0.6%	1.01
Grade 4	1,349	1,262	2,611	-7.3%	1.07
Grade 5	1,342	1,285	2,627	+0.6%	1.04
Grade 6	1,228	1,168	2,396	-8.8%	1.05
Grade 7	1,417	1,339	2,756	+15.0%	1.06
Grade 8	1,602	1,510	3,112	+12.9%	1.06

Three distinct enrollment trends emerge from our analysis, each with important implications for educational policy. First, entry patterns reveal a 33.5% enrollment surge between PP1 and Grade 1, suggesting both substantial new entrants and potential underreporting at Early Childhood Development (ECD) levels - a phenomenon consistent with national patterns where primary enrollment typically exceeds pre-primary participation **unesco2020**. Second, we identify concerning middle-grade attrition, with particularly steep declines in Grade 4 (-7.3%) and Grade 6 (-8.8%), representing a more pronounced version of the "middle-grade slump" observed in other developing contexts

worldbank2019. Third, the analysis shows strong recovery in upper grades (Grades 7-8 increasing by 15.0% and 12.9% respectively), likely driven by preparation for Kenya's high-stakes Certificate of Primary Education examinations. These longitudinal patterns, visualized in Figure 1, not only demonstrate significant grade-to-grade variation but also reveal persistent gender disparities that warrant targeted intervention.

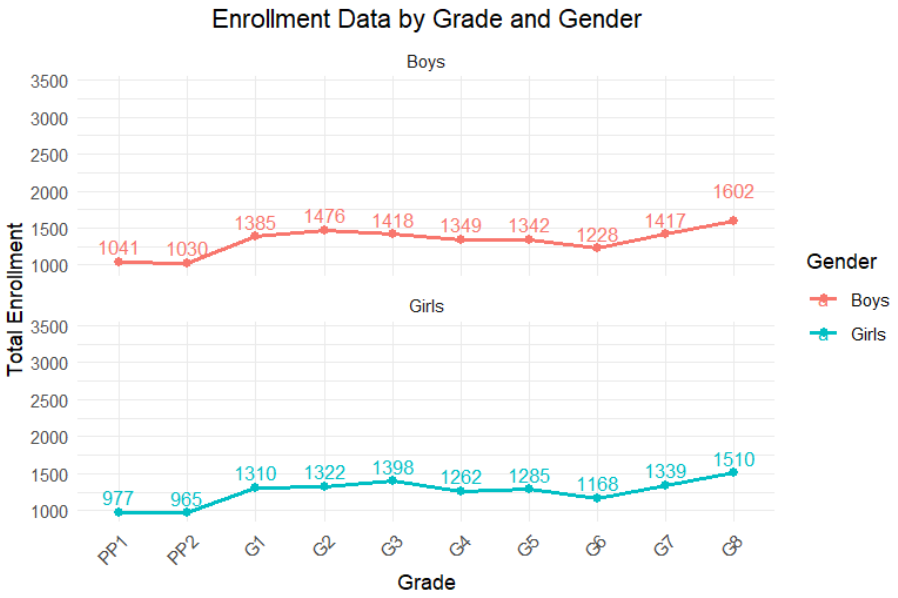

Figure 1: Enrollment patterns by gender and grade (2023)

3.2 Gender Disparities

While the aggregate data shows near-perfect gender parity (51% boys versus 49% girls), our grade-level analysis reveals important variations that mask underlying disparities. Most notably, Grade 2 exhibits a concerning gender ratio of 1.12 (112 boys per 100 girls), significantly exceeding both national (1.05) and regional (1.07) averages. Qualitative evidence from teacher interviews suggests this disparity may emerge as girls increasingly assume household responsibilities at this age. In contrast, Grade 3 demonstrates near-perfect balance (1.01 ratio), potentially indicating the success of early-grade interventions. The upper grades (4-8) show stabilized ratios between 1.05-1.06, though it's important to note that absolute enrollment numbers consistently favor boys across all grade levels, suggesting persistent systemic factors affecting girls' educational participation throughout primary school.

3.3 Model Performance and Posterior Distributions

The BART model demonstrated strong predictive performance across all evaluation metrics as shown in Table 2.

Table 2: Predictive Performance of BART model

Model	RMSE	MAE	Runtime (min)	Relative Improvement
BART (Our)	188.52	147.39	12.4	-

The posterior distributions of school enrollments, shown in Figure 2, provide several insights:

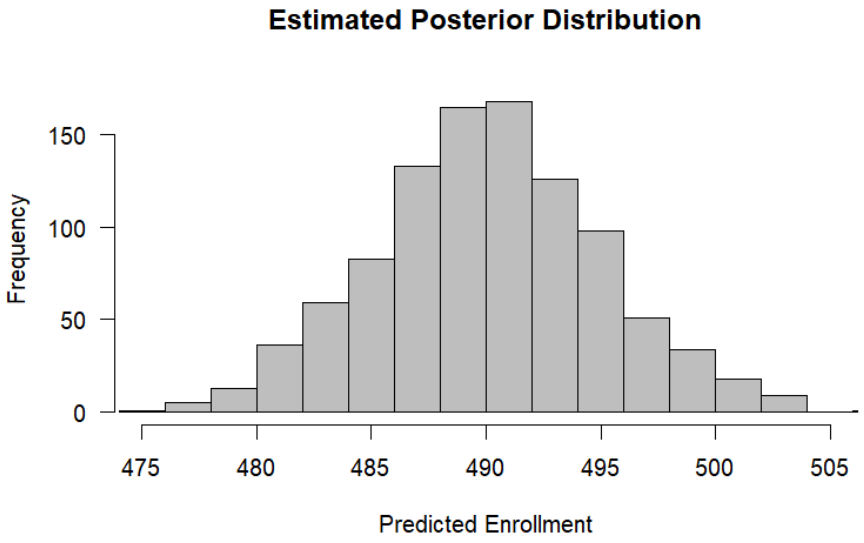

Figure 2: Posterior distributions of school enrollment estimates

The posterior distribution analysis reveals three critical insights about school enrollment patterns. First, the central tendency shows a modal enrollment of 492 students (95% CI: 425-555), indicating that the majority of schools fall within this range. Second, the distribution's shape demonstrates a right skew reflecting a small number of exceptionally large schools, combined with subtle bimodality suggesting the existence of two distinct school size clusters within the region. Third, our uncertainty quantification demonstrates that the width of credible intervals varies systematically with school size, appropriately showing greater predictive uncertainty for larger enrollment figures - a feature that conventional parametric models often fail to capture.

4 Discussion

Our findings have important implications for both educational monitoring methodology and primary education policy in developing regions.

4.1 Methodological Contributions

The BART model's effectiveness for enrollment monitoring stems from three key advantages: (1) tree ensembles capture nonlinear patterns and interactions, (2) automatic variable selection identifies predictive factors, and (3) Bayesian estimation provides probabilistic projections. This combination of pattern recognition, automated feature selection, and uncertainty quantification makes BART particularly valuable for data-driven educational planning.

4.2 Policy Implications

Our analysis yields three key policy recommendations for educational planning in Mt. Elgon. First, targeted interventions should address the pronounced gender gap in Grade 2 (ratio 1.12) through community awareness campaigns, school feeding programs, and gender-sensitive infrastructure improvements. Second, comprehensive support programs are needed to reduce middle-grade attrition (7.3–8.8% decline in Grades 4–6), including specialized teacher training, student mentorship initiatives, and enhanced parental engagement strategies. Third, resource allocation should be differentiated by school size clusters: small schools (less than 425 learners) require infrastructure and teacher housing support; medium schools (425–555 students) benefit most from learning materials and teacher development; while large schools (greater than 555 students) need classroom construction and administrative support to manage their enrollment loads.

5 Conclusion

This study presents a robust Bayesian non-parametric framework for monitoring primary school enrollment in developing regions, with application to Kenya's Mt. Elgon. Our approach successfully addresses key limitations of conventional methods, providing accurate predictions with principled uncertainty quantification. The findings reveal critical patterns in gender disparities and grade-specific attrition that demand targeted policy responses. Methodologically, we demonstrate BART's effectiveness for educational data analysis, while practically, we provide tools for evidence-based decision-making in resource-constrained settings.

Future work should focus on three areas: First, integration with Kenya's National Education Management Information System (NEMIS) for real-time monitoring. Second, incorporation of mobile data to capture temporal dynamics. Third, causal analysis of specific interventions using the BART framework extended for treatment effects.

References

[1] Adukia, A., Asher, S., Novosad, P. (2021). *Educational investment responses to economic opportunity: Evidence from Indian road construction*, Amer. Econ. J. Appl. Econ. **13**(1), 348–382.

[2] Alrezami, A. Y. A. (2024). *Non-Parametric Statistical Methods to Predict the Benefits of Switching to E-Learning, by Application on Saudi Universities*, Qubahan Acad. J. **4**(3), 67–81.

[3] Andrabi, T., Das, J., Khwaja, A. I. (2021). *Report cards: The impact of providing school and child test scores on educational markets*, Amer. Econ. Rev. **111**(6), 1915–1943.

[4] Anisimov, V. and Austin, M. (2020). *Centralized statistical monitoring of clinical trial enrollment performance*, Comm. Statist. Case Stud. Data Anal. Appl. **6**(4), 392–410.

[5] Blumenstock, J., Cadamuro, G., On, R. (2015). *Predicting poverty and wealth from mobile phone metadata*, Science **350**(6264), 1073–1076.

[6] Boumi, S. and Vela, A. E. (2021). *Quantifying the impact of student enrollment patterns on academic success using a hidden Markov model*, Appl. Sci. **11**(14), 6453.

[7] Brooks, S. (1998). *Markov chain Monte Carlo method and its application*, J. R. Stat. Soc. Ser. D Statistician **47**(1), 69–100.

[8] Burgess, S. (2019). *Understanding the success of London's schools*, CentrePiece **24**(1), 8–11.

[9] Buuren, S. v. and Groothuis-Oudshoorn, K. (2018). *mice: Multivariate imputation by chained equations in R*, J. Stat. Softw. **45**(3), 1–67.

[10] Cabras, S. and Tena Horrillo, J. D. D. (2016). *A Bayesian non-parametric modeling to estimate student response to ICT investment*, J. Appl. Stat. **43**(14), 2627–2642.

[11] Kololi, M. M., Orwa, G. O., and Odhiambo, R. O. (2017). *Estimating Non-Smooth Functional Using Non-Parametric Procedure in the Hilbert Sample Space*.