Mathematical Modelling and Applications
2025, Vol. 10, No. 2, pp. 31-42
https://doi.org/10.11648/j.mma.20251002.12

Research Article

Assessing the Efficiency of Generalized Additive Mixed Models (GAMMs) Using Milk Production Variables: Evidence from Tanzania's National Panel Surveys

Zainabu Bonza[1,*], Rosalia Katapa[2], Amina Msengwa[2]

[1]Department of Statistics, Eastern Africa Statistical Training Centre, Dar es Salaam, Tanzania
[2]Department of Statistics, University of Dar es Salaam, Dar es Salaam, Tanzania

Abstract

The study aimed to evaluate the performance of Generalized Additive Mixed Models (GAMMs) in fitting and predicting cattle milk production variables in Tanzania. The study used the National Panel Survey data collected between 2012 and 2021. GAMM (gamma), GAMM (lognormal), and GAMM (inverse normal) were fitted, and their performance was evaluated using Akaike Information Criterion (AIC), proportional of deviance explained, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The findings show that the GAMM with the lognormal distribution performed best, having a balanced good fit and predictive performance (AIC=4714.36, deviance explained = 94%, RMSE = 4.72, and MAE=2.94). Although the GAMM with the gamma distribution had a slightly better prediction accuracy (RMSE = 4.27 and MAE =2.77), it had a higher AIC (5027.75) and a smaller proportion of deviance explained (59%), making it less suitable for identifying factors affecting cattle milk production. The GAMM with the inverse normal distribution showed the poorest performance, with the highest AIC (5366.34), lowest deviance explained (37.3%), and largest errors (RMSE = 13.64, MAE = 4.46). Despite the poor performance of the GAMM with the inverse normal distribution, the overall results demonstrate that GAMMs are an effective tool for fitting and predicting cattle milk production variables, particularly when data are collected repeatedly from the same units over time. The GAMM with the lognormal distribution not only provided accurate predictions but also offered valuable insights into nonlinear relationships between milk production and household inputs. These insights can guide dairy farmers and policymakers in optimizing milk production by focusing on key factors such as improved breeding practices, proper feeding, health management, and labour efficiency.

Keywords

Generalized Additive Mixed Models, Household Input, Cattle Milk Production

*Corresponding author: zainabu.bonza@eastc.ac.tz (Zainabu Bonza)

Received: 11 June 2025; **Accepted:** 27 June 2025; **Published:** 26 September 2025

1. Introduction

The dairy sector has been recognized as an important part of livestock production, playing a crucial role in supporting the achievement of the Sustainable Development Goals, boosting the Gross Domestic Product (GDP), and providing a significant source of household income. The importance of the dairy sector in Tanzania has been recognized by the United Republic of Tanzania, which reports that the sector contributes 30% to the GDP of the livestock industry [1]. Milk production in Tanzania is projected to increase from 2,037,765 tonnes in 2016 to 3,384,970 tonnes by 2031 [2]. Despite the increase in milk production, cattle productivity in Tanzania is still low [2, 3]. All over the world, livestock keepers intend to produce animal products efficiently and economically to reduce household poverty and foster national GDP. However, poverty incidence among livestock keepers in Tanzania is still high [2]. Therefore, effective investment decisions in the livestock sector are crucial for improving cattle milk productivity. To make informed decisions, it is essential to understand the various factors influencing milk production, including household inputs and management practices. In this regard, analysing the relationships between household inputs and milk production can provide valuable insights for improving efficiency in the sector [4, 5]. These relationships provide more understanding and insights regarding which input parameter is worth more effort to improve per-head cattle milk productivity.

Inputs into livestock production consist of land, housing, water, labour, animal health, and livestock feed [2]. These inputs play a critical role in determining milk production, and many studies have explored their relationships. A significant body of research has examined the effect of feeding practices on milk production. Various feeding practices and their effects on milk yield and composition in cattle and camels in Kenya have been studied [6]. The influence of labour and fodder type (green, dry, and concentrate) on daily milk production has also been examined [7]. These studies emphasize the significance of feed availability and quality in milk production. The impact of restricted water intake on milk yield, body weight, and blood composition in dairy cows has likewise been explored, underscoring water access as another critical factor influencing cattle milk productivity [8].

In addition to feeding, labour and water, housing systems also play a key role. The effects of two major housing systems, loose and tie-stall, on the welfare and milk yield of dairy cows in Italy [9]. This research highlights how housing conditions can affect both the well-being and productivity of livestock. The use of generalized linear models to examine the effects of breeding and feeding systems on milk production and reproductive performance has contributed to a broader understanding of how farm management practices influence milk productivity [10]. Animal health, particularly in relation to disease, is also a key factor. The effect of gastrointestinal nematode infections on daily milk production among smallholder farmers in Kenya has been investigated, demonstrating the direct impact of health management on cattle milk productivity [11].

Despite the wealth of research on these various household inputs, much of it relies on linear models (associating a straight-line relationship between household inputs and milk production), which limits their ability to capture nonlinear relationships. Additionally, these studies have used cross-sectional data, which restricts their capacity to account for both fixed and random effects. Fixed effects refer to variables that are considered to have a constant influence on the outcome across all observations, meaning their impact does not vary between groups (households) or over time. This limitation can lead to biased parameter estimates and less accurate predictions [12].

On the other hand, when relationships are considered nonlinear, Artificial Neural Networks (ANNs) have been widely used [13]. Studies have compared the predictive ability of generalized linear models and ANNs [4, 5, 14-17]. ANNs are known to be effective in capturing nonlinear relationships between variables, but they pose challenges in interpreting these relationships [18]. This limitation has made the available studies effectively predict milk production but not capture the relationship between milk production and household inputs. Therefore, the studies cannot inform on the significant household inputs affecting cattle milk production and hence do not guide dairy farmers and policymakers on how they can optimize milk production.

In light of the limitations of the existing studies that predominantly rely on cross-sectional data and consider nonlinear models that lack interpretability, this study aimed to provide a more comprehensive understanding of household inputs affecting cattle milk production by evaluating the performance of Generalized Additive Mixed Models (GAMMs). The importance of Generalized Additive Mixed Models (GAMMs) has been highlighted in previous research [19, 20]. These models can capture fixed, random, and nonlinear relationships between variables. Furthermore, they provide interpretable smooth functions for nonlinear effects, which can be visualized to gain insights into the relationships among variables. Therefore, this study effectively captured both fixed and random effects, leading to improved parameter estimates and predictive accuracy. The study also captured interpretable nonlinear effects contributing to a better understanding of the significant household inputs affecting cattle milk production. The study's insights into significant household inputs can guide dairy farmers and policymakers in optimizing milk production.

2. Materials and Methods

2.1. Data Source and Study Variables

The study used data from the National Panel Surveys (NPS) waves 3, 4, and 5. The datasets were collected by the National Bureau of Statistics (NBS) in collaboration with the Office of the Chief Government Statistician – Zanzibar. Datasets were requested through the National Bureau of Statistics [21]. The NPS data are collected biannually with the third, fourth, and fifth waves undertaken in the years 2012/2013, 2014/2015, and 2020/2021, respectively. To focus on understanding the key household inputs affecting cattle milk production, the target population of the study includes all households that participated in waves 3, 4, and 5 of the survey and owned at least one milked cow. The study employed a complete enumeration approach, meaning no sampling was necessary. This census methodology ensures comprehensive coverage of all households with at least one milked cow across the three waves, providing a representative dataset for analysing dairy farming practices without sampling bias. In total, the population size of the study included 1266 households, of which 505, 352 and 409 participated in waves 3, 4, and 5, respectively. The pooled dataset was obtained by merging all the variables of interest from section 2A of the Agriculture questionnaire and sections 2 to 6 of the livestock and fisheries questionnaire. Data were merged for each wave, thereafter all three wave datasets were appended.

Table 1 presents the list of twenty-four (24) variables selected for analysis from the NPS questionnaire. The variables include milk production in litres per cow per day and various household inputs affecting milk production. The variables consist of both numerical and categorical measures. Categorical variables with more than three categories from the NPS questionnaire were recoded into new variables with two or three categories to simplify the interpretation of results. The table also includes wave/year of data collection and Household ID.

Table 1. *List of Selected Variables from the Tanzania National Panel Survey (NPS) Questionnaire.*

S/N	Variable
1	Milk production
2	Area owned
3	Number of cattle
4	Total cost of veterinary services
5	Total cost of labour
6	Total cost of feeding
7	Total cost of water
8	Feed purchases
9	Use of vaccine
10	Use of tick treatment
11	Use of dewormer
12	Major Feeding practices
13	Person administered vaccine
14	Person administered tick treatment
15	Person administered dewormer
16	Hiring Labour
17	Average frequency of watering cattle
18	Main source of water
19	Water purchases
20	The main housing system used for cattle
21	Practice mating/breeding
22	Main practised mating/breeding
23	Household ID
24	Year (wave)

2.2. Modelling

To assess multicollinearity among the variables and avoid redundancy, the correlation matrix and Variance Inflation Factor (VIF) were used. When a study includes categorical variables, these variables can be transformed into numeric ones using an indicator matrix of dummy variables [22]. The Spearman's rank correlation coefficient was employed to assess the strength and direction of the relationship between pairs of ranked variables. This method was chosen because it is considered one of the most robust tests for evaluating associations between variable pairs [23]. Variables with a bivariate correlation coefficient exceeding 0.8 were regarded as highly correlated, and one of the variables was removed until the VIF for each independent variable fell below 10 [24]. This threshold was set because a VIF greater than 10 suggests weak estimation of regression coefficients due to multicollinearity [24].

GAMMs are models with higher flexibility and complexity, they are capable of modelling complex datasets with hierarchical structures by considering fixed and random effects, smooth terms, and a non-normal response [20]. The current study fitted Generalized Additive Mixed Models (GAMMs) as represented by the following equation.

$$G\left(\mu^b\right) = \eta = X^*\alpha + \sum_{j=1}^{p} f_j\left(x_j\right) + Zb$$

$G(.)$ is a monotonic differentiable link function.

μ^b is a conditional mean of the response.

η is the linear predictor.

α is the vector of fixed parameters.

X^* is the fixed effects model matrix.

f_j is the smooth function of the covariate x_j.

Z is the random effects model matrix, and

b is the random effects coefficient assumed to be distributed as $N\{0, D(\theta)\}$ and θ is a variance component.

There were six (6) smoothers and each is a function of only one covariate (numeric household input). Each smoother $f_j(x_j)$ is represented by $\sum_{k=1}^{K} \beta_{j,k} b_{j,k}(x_j)$, where $b_{j,k}$ smaller functions are known as basis functions. $\beta_{j,k}$ are corresponding coefficients which need to be estimated and K is known as “basis size,” or “basis complexity,” and it is used to determine the maximum complexity of each smoother. The Effective Degrees of Freedom (EDF) was used to measure the complexity of penalized smooth terms. The EDF is a measure of the model's flexibility in fitting the data. Values greater than 1 suggest a nonlinear relationship between household inputs and milk production.

To estimate the effect of categorical variables on milk production, all categorical variables were dummy-coded. For categorical variables with more than two categories, the most frequent category was chosen as the reference category to enhance the stability and precision of the estimates. The effect of time on milk production was evaluated by treating the year (wave) variable as a categorical fixed effect, with the first wave (2013) serving as the reference category.

Statistical inference in GAMMs is concerned with the estimation of nonparametric functions $f_j(.)$, the smoothing parameters λ and the parametric terms. Estimation of λ was achieved under the restricted maximum likelihood method, since it provides a good fit to the data [19]. A diagnostic check for the fitted models was done by first inspecting the value of k which is a smooth term that controls for extra flexibility.

The quantile-quantile (Q-Q) plot of residuals and a plot of residuals versus linear predictor were used to assess the validity of the models. The residuals should approximately follow a straight line in the plot to indicate that parts which are not explained by the model (the residuals) are assumed to represent a random noise. Heteroscedasticity was assessed by the residuals versus linear predictor plot. The t-value was used to test the significance of parametric terms in GAMMs, and the significance of smoothing functions was tested by using the F-value. The analyses used R version 4.4.1 (2024-06-14, Universal C Runtime). The Mixed GAM Computational Vehicle – mgcv R package version 1.9-1 was used to fit the models.

The milk production variable is characterized by non-negative, positively skewed continuous data, with zero being an impossible value [25]. Consequently, Gamma, Lognormal, and Inverse Normal distributions were selected as the potential distributions for the data. The Log link function was chosen for all three distributions to facilitate clear interpretation. Performance evaluation of the models was done during the model fitting and testing stages. To achieve this, the dataset was randomly partitioned into training and testing datasets. The training dataset had 90% of the data (1140 households), while the test dataset had 10% of the remaining data (126 households). During model fitting, the best GAMM amongst the three fitted models was selected based on the smallest Akaike Information Criterion (AIC) and highest deviance explained. During model testing, the best GAMM was selected based on the smallest values of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

3. Results and Discussion

3.1. Results

The study transformed fifteen (15) categorical variables into twenty-three (23) dummy variables, resulting in a total of twenty-nine (29) household inputs (independent variables). However, only twenty-three (23) variables were included in the modelling process. This was due to the Spearman rank correlation analysis, which identified six (6) dummy variables (use of vaccine_No, use of dewormer_No, use of tick treatment_ Yes, all animals at least once, hiring labour_No, water purchases_No and feed purchases_No) as redundant. After removing the six (6) variables, the Variance Inflation Factors (VIFs) for the remaining twenty-three (23) variables fell below 10, indicating that the collinearity issue had been resolved.

Before evaluating the performance of GAMMs, the fitted models were diagnosed by checking if the default basis complexity k for all smoothing functions was sufficient. The diagnostic check results show that the default basis complexity (k=10) was sufficient for all smoothing functions of the three GAMMs. This is because EDF for all estimated smoothers were well below their maximum possible degree of freedom, the p-values for the observed index (a measure of residuals pattern) were greater than 0.05, and all k-indices were very close to 1. Therefore, there is no extra structure that can be captured by increasing the value of k.

Figure 1 shows diagnostic plots of GAMMs presenting the Q-Q plot and Residuals vs. Linear Predictor. The upper, middle and lower panels of the figure present diagnostic plots for the GAMM (gamma), GAMM (lognormal) and GAMM (inverse normal), respectively.

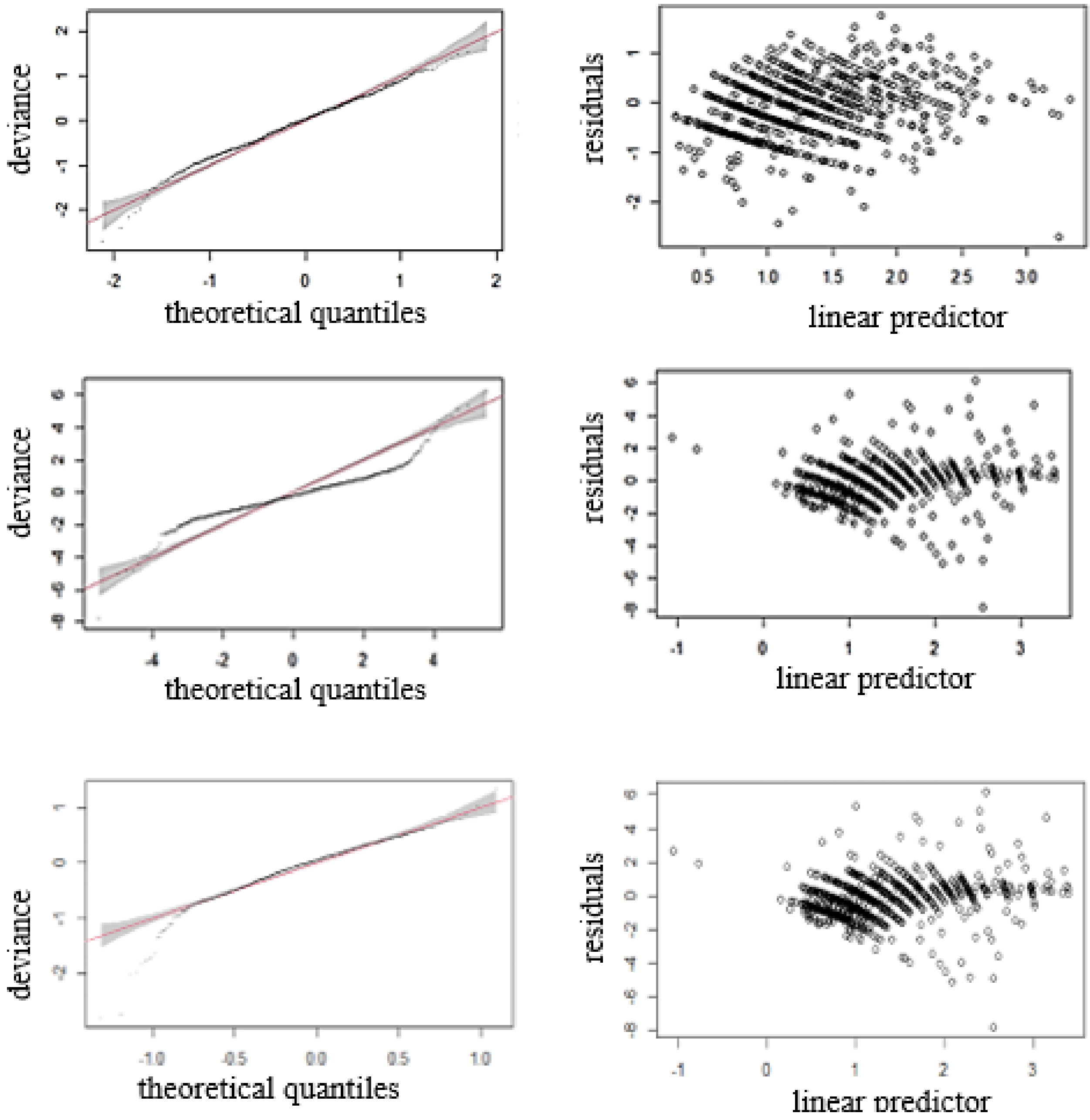

Figure 1. *Diagnostic Plots of Generalized Additive Mixed Models: QQ Plot and Residuals vs. Linear Predictor.*

The Q-Q plots are presented on the left panel while the right panel presents the plots of residuals versus linear predictor. The Q-Q plots on the left panel indicate an approximately normal distribution of residuals (deviance) for each fitted GAMM. This is because the residuals approximately follow the straight line in the Q-Q plot of each GAMM, which indicates that parts not explained by the model (i.e. the residuals) are assumed to represent a random noise. On the right panel of the figure, residuals versus linear predictor plots of each GAMM indicate residual variance was not increasing much causing an obvious V-shape (heteroscedasticity). It can be seen, that some of the residuals were nearly zero with increased values of linear predictor. Generally, diagnostic plots indicate the models best fit the given data.

Table 2 presents performance metrics during model development and testing stages for the three fitted GAMMs. It was found that among these models, GAMM (lognormal) had the smallest AIC (4714.36) and highest deviance explained (94%), indicating the best overall fit to the data despite slightly higher RMSE (4.72) and MAE (2.94). GAMM (gamma) performed moderately, with AIC=5027.75, deviance explained=59%, and the smallest RMSE (4.27) and MAE (2.77). GAMM (inverse normal) with the highest AIC (5366.34), lowest deviance explained (37.3%) and the largest errors (RMSE = 13.64 and MAE = 4.46), demonstrated the poorest fit and predictive accuracy. Therefore, GAMM (lognormal) is the best-performing model based on its low AIC and high deviance explained, while its RMSE and MAE are slightly higher than those of GAMM (gamma), the significantly greater deviance explained suggests it captures the relationship between cattle milk production and household inputs more effectively than GAMM (gamma).

Table 2. Performance Comparison of GAMMs: AIC, Deviance Explained, RMSE, and MAE for GAMM (gamma), GAMM (lognormal), and GAMM (inverse Normal).

Models	AIC	Deviance Explained (%)	RMSE (litre/per cow/per day)	MAE (litre/per cow/per day)
GAMM (gamma)	5027.75	59	4.27	2.77
GAMM (lognormal)	4714.36	94	4.72	2.94
GAMM (inverse normal)	5366.34	37.3	13.64	4.46

Note: AIC = Akaike Information Criterion, RMSE = Root Mean Squared Error, MAE = Mean Absolute Error

Table 3 presents the significance of nonlinear effects (smoothing functions) for cattle milk production from the GAMM (lognormal). The table shows the model had six smoothing functions (total cost of feeding, total cost of veterinary services, number of cattle, total cost of labour, total cost of water, and area owned). All smoothing functions were significant since their p values of the F test were less than 0.05. The estimated Effective Degree of Freedom (EDF) values ranged from 2.76 (area owned) to 6.99 (total cost of veterinary services), indicating varying levels of nonlinearity among the household inputs affecting cattle milk production. Additionally, the random effect (Household ID) was significant ($p \prec 0.05$), indicating substantial variation in cattle milk production attributable to household differences.

Table 3. Significance of Nonlinear Effects for Cattle Milk Production from the GAMM (lognormal).

Smooth function of a variable	EDF	F-Value	P-value
Total Cost of Feeding	6.38	5.57	<0.001
Total Cost of Veterinary Services	6.99	10.37	<0.001
Number of Cattle	5.47	20.74	<0.001
Total Cost of Labour	4.93	5.44	<0.001
Total Cost of Water	3.40	3.88	0.004
Area Owned	2.76	4.09	0.006
Household ID	616.11	3.86	<0.001

Figure 2 presents nonlinear relationships between smoothing functions and cattle milk production, highlighting how significant continuous household inputs affect cattle milk production in a complex, nonlinear fashion, supporting the findings presented in Table 3. The figure includes six plots, each illustrates the individual effect of a smoothing function on the log-transformed scale, with all other terms in the model held constant at zero. Figure 2a and 2b show that the total cost of feeding (TCF) and area owned (AO) exhibit positive nonlinear relationships with cattle milk production, as indicated by the plots above zero. However, for higher values of both the total cost of feeding and the area owned, the relationships tend to show downward trends. Figure 2c and 2f reveal that the total cost of veterinary services (TCV) and the number of cattle (NC) also have positive nonlinear relationships with cattle milk production, with the plots remaining above zero. In contrast, figure 2d and 2e illustrate that the total cost of labour (TCL) and total cost of water (TCW) have predominantly negative nonlinear effects on cattle milk production (the plots are below zero), showing downward trends for lower and intermediate values of labour and water costs. However, for higher values, both relationships show upward trends.

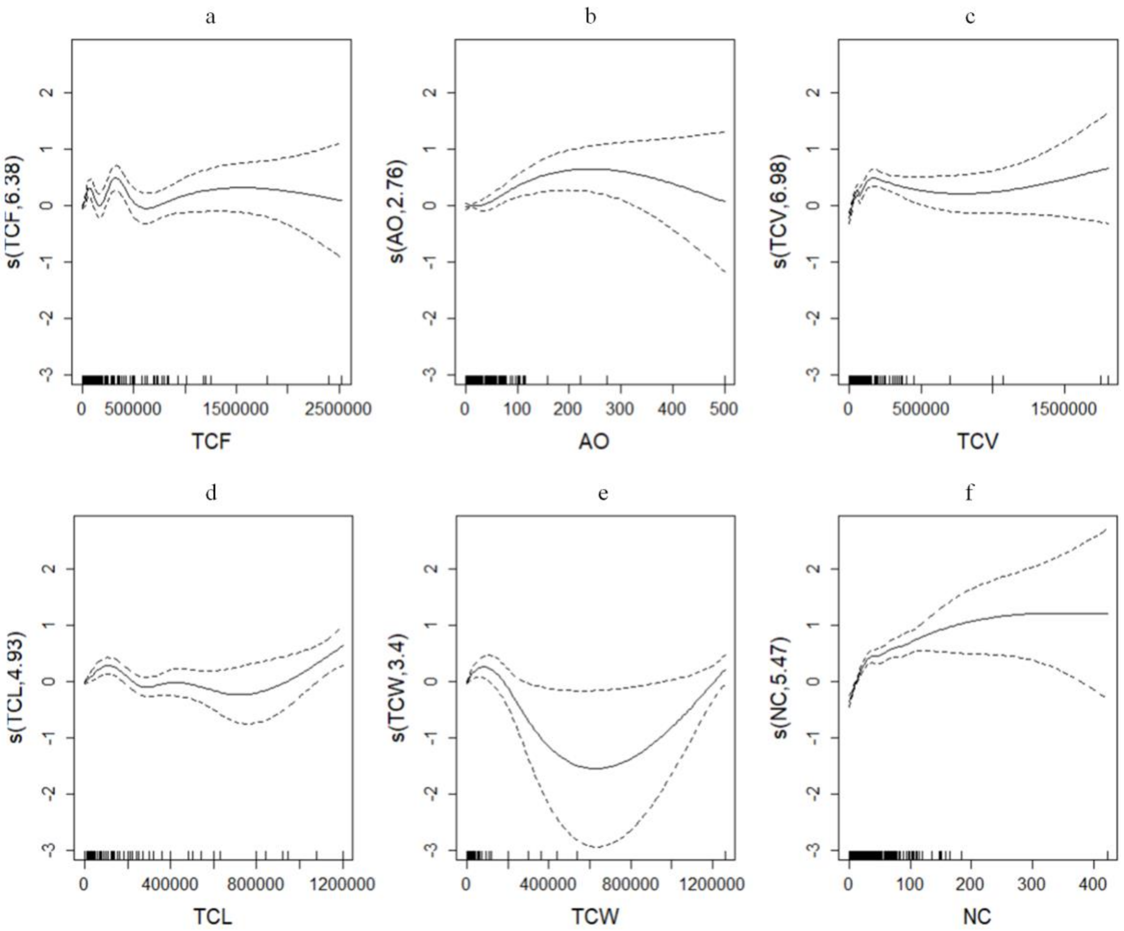

Figure 2*. Plots of Nonlinear Relationships between smoothing functions and cattle milk production.*

Table 4 presents the estimates of fixed effects for factors (household inputs) affecting cattle milk production from the GAMM (lognormal). Each fixed effect quantified its unique contribution to the logarithmic changes in the average cattle milk production relative to the reference category of each categorical household input. The significant $(p \prec 0.05)$ household inputs identified by GAMM (lognormal) include average frequency of watering cattle, major feeding practices, use of vaccines, use of tick treatments, person administered tick treatment, person administered de-wormer, main practiced mating/breeding, and main housing system. These household inputs showed significant associations with cattle milk production, with varying directions and magnitudes of effects. The estimates, standard errors, t-values, and significance levels for the significant household inputs are presented in Table 4.

Table 4*. Fixed Effects Estimates for Factors Affecting Cattle Milk Production, GAMM (lognormal).*

Parameter	Estimate	Standard error	t-value	P-value
Intercept	1.37			
Average Frequency of watering cattle				
Cattle were given water: reference				
Cattle get water on their own	-0.19	0.05	-4.10	<0.001
Major feeding practices				
Grazing: reference				
Feeding	0.46	0.08	5.62	<0.001
Use of vaccine				

Parameter	Estimate	Standard error	t-value	P-value
Yes, all animals at least once: reference				
Yes, some	0.32	0.07	4.48	<0.001
Use of tick treatment				
No: reference				
Yes, some	-0.22	0.10	-2.18	0.03
Person administered tick treatment				
Not Applicable: reference				
Veterinary clinic	0.14	0.06	2.51	0.01
Others	0.06	0.05	1.27	0.2
Person administered de-wormer				
Not applicable: reference				
Veterinary clinic	-0.03	0.05	-0.64	0.02
Others	0.14	0.06	2.51	0.14
Main practised mating/breeding				
None: reference				
Natural mating	-0.02	0.08	-0.23	0.82
Artificial insemination	0.63	0.12	5.32	<0.001
Main housing system				
Confined: reference				
Others	-0.14	0.20	-0.68	0.5
None	0.22	0.08	2.58	0.01

3.2. Discussion

The primary aim of this study was to evaluate the performance of Generalized Additive Mixed Models (GAMMs) in fitting and predicting cattle milk production variables in Tanzania. The study employed panel data, which allowed for the consideration of both random and fixed effects, offering a thorough examination of the relationships over time and across various households. The performance of the GAMM (gamma), GAMM (lognormal), and GAMM (inverse normal) was compared based on model fitting and prediction performances. The GAMM (lognormal) emerged as the best-performing model among the models. While the GAMM (gamma) model demonstrated slightly better predictive accuracy based on lower RMSE and MAE values, we prioritized explanatory power in selecting the GAMM (lognormal) as the preferred model. This decision aligns with the primary objective of the study, which was to understand the relationship between cattle milk production and household inputs more effectively, rather than solely to optimize out-of-sample prediction. The lognormal model explained 94% of the deviance, substantially outperforming the gamma model in capturing the variability inherent in the data. This higher explanatory power offers deeper insight into the relationships between predictors and response, which is crucial for informing policy and theoretical implications [26].

Studies have compared prediction techniques like Artificial Neural Networks (ANNs) and Multiple Regression in forecasting milk production [5, 15, 16]. While ANNs are effective at capturing complex nonlinear relationships, the GAMMs models compared in this study provided more interpretable results that can be directly linked to management practices. This interpretability stems from the models providing estimates for both fixed and random effects, allowing stakeholders to understand how changes in household inputs affect cattle milk production. Additionally, the models allowed visualization of nonlinear relationships between household inputs and cattle milk production.

In addition to identifying the best-performing model, this study highlighted several key household inputs as significant predictors of cattle milk production. These predictors included linear and nonlinear relationships with cattle milk production, demonstrating the complex nature of the factors influencing dairy performance. Nonlinear relationships were observed with predictors such as the total cost of feeding and total labour costs. These findings are consistent with earlier research [7]. However, our study identified additional household inputs, including the total cost of water, total cost of veterinary services, number of cattle, and area owned that also significantly affected cattle milk production. Moreover, the nonlinear nature of these relationships could not be captured by linear models.

The results also revealed a significant random effect for household ID (EDF= 616.11, F-value =3.86, p<0.001), indicating that differences between households contribute significantly to the variability in cattle milk production. This suggests that household-specific factors, which may not have been fully captured by the fixed effects, play an important role in shaping cattle milk production. Consequently, the variability due to household differences could not be explained by the fixed effects alone. This finding highlights the importance of incorporating random effects in models when analysing data with hierarchical or grouped structures, such as livestock data at the household level. Furthermore, it justifies the use of GAMMs, which are well-suited to account for both fixed and random effects and capture complex relationships between predictors and the outcome. This approach provides a more accurate understanding of the underlying factors influencing cattle milk production, especially when there are unobserved variations across different households.

The results further indicated that major feeding practices had a positive relationship with milk production. This finding is consistent with earlier research that explored the effect of various feeding practices on smallholder dairy cattle and pastoral camel herds [6]. Despite using cross-sectional data, their study identified several feeding practices that were associated with higher milk yields in smallholder dairy cattle herds. In contrast, our study categorized feeding practices based on whether cattle primarily relied on grazing or purchased feed, offering a broader perspective on cattle feeding strategies. Additionally, our study revealed that the average frequency at which cattle sourced water on their own (i.e., cattle getting water independently) had a negative relationship with milk production. This highlights the importance of not only optimizing feeding practices but also ensuring that cattle have consistent and adequate access to water. The findings are in line with previous research, which concluded that milk yield decreased significantly with increasing water deprivation [8].

In addition, our study identified cattle health variables as key predictors of milk production, suggesting that animal health management plays a crucial role in determining milk production. These results are consistent with previous research, which similarly demonstrated the importance of animal health in optimizing milk production [27]. Our findings are also in line with the broader literature, which emphasizes the importance of veterinary services in improving production efficiency within extensive livestock systems. In particular, practices such as tick treatment and vaccinations were found to be essential for boosting milk yields. This highlights the critical need for integrating veterinary services into livestock management, not only to improve animal health but also to enhance food production. The significance of these veterinary services is further supported by research emphasizing the necessity of continued multidisciplinary collaboration and coordinated action across sectors to ensure global food security and equitable access to animal-based nutrition [28].

It is important to note that artificial insemination, as the primary breeding method, has the largest positive effect on cattle milk production in this study. This finding is particularly notable given the relatively lower influence of other management practices. Notably, this result aligns with findings from a study conducted in the same area, which reported that 89% of respondents recognized artificial insemination as an effective breeding method for increasing cattle milk production [29]. However, while that study relied on descriptive statistics, our analysis builds on this by employing more advanced statistical modelling.

The study also revealed a negative association between cattle milk production and two household-level interventions: partial tick treatment of herds and deworming carried out by veterinary clinics. While these practices are generally intended to support animal health, this finding may seem counterintuitive. One possible explanation is that treating only a few animals within a herd may contribute to acaricide resistance in tick populations, leading to suboptimal health outcomes and reduced milk production in the study area [30]. Similarly, the observed association with deworming through veterinary clinics could be related to issues in the application or choice of dewormers. For example, Tanzanian farmers often rely on inexpensive options such as albendazole, which, due to frequent and prolonged use, has shown reduced efficacy against gastrointestinal nematodes [31]. These potential mechanisms, while consistent with existing literature, remain speculative and warrant further investigation to confirm their role in influencing milk productivity.

A fundamental limitation of this study is the model's inability to consider interactions between household inputs. While the model successfully evaluates the independent effects of individual variables on cattle milk production, it does not account for the potential combined effects or interdependencies between these household inputs. In real-world settings, factors such as feed purchases, water usage, and the main housing system may not operate independently but rather interact in complex ways, influencing cattle milk production in ways that cannot be captured by the current model. Future studies could benefit from incorporating interaction terms to better reflect the multifaceted relationships between these inputs, providing a more comprehensive understanding of their collective impact on cattle milk production.

Another significant limitation of this study arises from the absence of detailed animal-level data. The NPS dataset, which was used for this analysis, primarily focuses on household-level information and does not provide specific details about individual animals, such as breed, lactation stage, age, or parity. These animal characteristics are known to be important factors influencing productivity and milk production in cattle. Consequently, our analysis could not account for the potential variation in productivity due to differences between individual animals. Despite this, we believe that the household-level variables we included still provide valuable insights into factors that influence cattle milk production at a broader scale. For future research, the inclusion of detailed animal-level data such as breed, lactation stage, and age would enhance the understanding of the specific factors that influence cattle milk production.

Furthermore, given the unexpected negative associations between cattle milk production and the use of tick treatment on only a few cattle within a herd, as well as the application of dewormers by veterinary clinics, future studies should aim to explore these relationships in greater depth. Specifically, research could examine the effectiveness of partial or selective tick treatment strategies within herds to determine whether more uniform treatment protocols would lead to improved cattle health and milk production outcomes. Previous studies have examined the timing, frequency, and dosage of dewormer treatments; however, further research is needed in Tanzania, where climate variability may influence the effectiveness of these treatments [32, 33].

Additionally, the study recommends that any strategy aiming at increasing milk production and overall productivity should prioritize key factors such as improved breeding practices, proper feeding, effective health management, and enhanced labour efficiency. Specifically, the study recommends that improved breeding strategies, particularly the use of artificial insemination, should be emphasized. Also, proper feeding strategies should include providing balanced and nutritious diets that meet the specific needs of lactating cows and ensuring that feed quality is consistently maintained to support optimal milk yields. Furthermore, livestock keepers should collaborate with local veterinary services to ensure that accurate and effective tick treatment, deworming, and vaccination programs are followed appropriately. Lastly, labour efficiency can be enhanced through better training of farm workers, investment in labour-saving technologies, and streamlined management practices to maximize productivity per unit of labour. By focusing on these key factors, farmers can create a sustainable approach to increasing milk production.

4. Conclusion

This study advances the understanding of predictive modelling in cattle milk production by examining the relationships between cattle milk production and household inputs using generalized additive mixed models (GAMMs). A key contribution of this study lies in its novel application of a GAMM framework to nationally representative panel data on livestock production. Previous studies have largely relied on linear or less flexible modelling techniques, whereas our approach captures complex nonlinear relationships while accounting for hierarchical data structure. To our knowledge, this is the first time GAMMs have been used to analyze this dataset in the context of smallholder dairy production in Tanzania, offering new insights into household inputs affecting milk production. The study provides practical recommendations for optimizing farm management, including the use of artificial insemination, improved labour management, adherence to vaccination, deworming, and tick treatment schedules. Additionally, sustainable feeding practices are emphasized. The methodology used in this study provides key insights and lays the groundwork for future research, especially in improving predictive models by exploring how interactions between household inputs affect cattle milk production.

Abbreviations

AIC	Akaike Information Criterion
ANNs	Artificial Neural Networks
AO	Area Owned
EDF	Effective Degrees of Freedom
GAMM	Generalized Additive Mixed Models
GDP	Gross Domestic Product
MAE	Mean Absolute Error
NBS	National Bureau of Statistics
NC	Number of Cattle
NPS	National Panel Surveys
Q-Q	Quantile-Quantile
RMSE	Root Mean Squared Error
TCF	Total Cost of Feeding
TCL	Total Cost of Labour
TCV	Total Cost of Veterinary Services
TCW	Total Cost of Water
VIF	Variance Inflation Factor

Supplementary Material

The supplementary material can be accessed at https://doi.org/10.11648/j.mma.20251002.12

Author Contributions

Zainabu Bonza: Conceptualization, Data curation, Methodology, Formal Analysis, Writing – original draft

Rosalia Katapa: Conceptualization, Supervision, Writing – review & editing

Amina Msengwa: Conceptualization, Supervision, Writing – review & editing

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The data used in this study were obtained from the Tanzania National Panel Survey (NPS) waves 3, 4, and 5, available through the World Bank Microdata Library. The data can be accessed by creating an account and requesting access through the Microdata Library portal at https://microdata.worldbank.org. Please note that access to these datasets may require approval from the data provider.

Additional Information

The Tanzania National Panel Survey (NPS) used the same questionnaire for waves 3, 4, and 5, which is available for reference. Descriptive statistics derived from the data are included in this study. For further details on the questionnaire and descriptive statistics, refer to the supplementary materials provided.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] The United Republic of Tanzania, "LIVESTOCK SECTOR TRANSFORMATION PLAN (LSTP) 2022/23 - 2026/27." Accessed: Nov. 28, 2022. Available: https://faolex.fao.org/docs/pdf/tan223748.pdf

[2] United Republic of Tanzania- Ministry of Livestock and Fisheries, "Tanzania Livestock Sector Analysis) 2016/2017-3031/2032)," =, p. 150, 2017,. Available: https://www.mifugouvuvi.go.tz/uploads/documents/sw-1749142748-1553602287-LIVESTOCK%20SECTOR%20ANALYSIS_compressed.pdf

[3] B. K. Paul *et al.*, "Reducing agro-environmental trade-offs through sustainable livestock intensification across smallholder systems in Northern Tanzania," *Int. J. Agric. Sustain.*, vol. 18, no. 1, pp. 35–54, 2020, https://doi.org/10.1080/14735903.2019.1695348

[4] Q. T. Nguyen, R. Fouchereau, E. Frénod, C. Gerard, and V. Sincholle, "Comparison of forecast models of production of dairy cows combining animal and diet parameters," *Comput. Electron. Agric.*, vol. 170, 2020, https://doi.org/10.1016/j.compag.2020.105258

[5] L. Sanzogni and D. Kerr, "Milk production estimates using feed forward artificial neural networks," 2001. https://doi.org/10.1016/S0168-1699(01)00151-X

[6] O. B. Kashongwe, B. O. Bebe, J. W. Matofari, and C. G. Huelsebusch, "Effects of feeding practices on milk yield and composition in peri-urban and rural smallholder dairy cow and pastoral camel herds in Kenya," *Trop. Anim. Health Prod.*, vol. 49, no. 5, pp. 909–914, 2017, https://doi.org/10.1007/s11250-017-1270-3

[7] S. M. Feroze, R. Singh, and S. Sirohi, "Economics of milk production and factors affecting milk yield in Meghalaya: Estimating the seasonal effect," *Indian J. Dairy Sci.*, vol. 72, no. 3, pp. 328–335, 2019, https://doi.org/10.33785/ijds.2019.v72i03.014

[8] W. Little, B. F. Sansom, R. Manston, and W. M. Allen, "Effects of restricting the water intake of dairy cows upon their milk yield, body weight and blood composition," *Cambridge Univ. Pres*, vol. 22, no. 3, pp. 329–339, 2010, https://doi.org/https://doi.org/10.1017/S0003356100035601

[9] S. Popescu, C. Borda, E. A. Diugan, M. Niculae, R. Stefan, and C. D. Sandru, "The effect of the housing system on the welfare quality of dairy cow," *Ital. J. Anim. Sci.*, vol. 13, no. 1, pp. 15–22, 2014, https://doi.org/10.4081/ijas.2014.2940

[10] S. Walsh, F. Buckley, K. Pierce, N. Byrne, J. Patton, and P. Dillon, "Effects of breed and feeding system on milk production, body weight, body condition score, reproductive performance, and postpartum ovarian function," *J. Dairy Sci.*, vol. 91, no. 11, pp. 4401–4413, 2008, https://doi.org/10.3168/jds.2007-0818

[11] W. Kabaka, G. Gitau, P. Kitala, N. Maingi, and J. Van Leeuwen, "The prevalence of gastrointestinal nematode infection and their impact on cattle in Nakuru and Mukurweini districts of Kenya," *Ethiop. Vet. J.*, vol. 17, no. 1, p. 95, 2014, https://doi.org/10.4314/evj.v17i1.8

[12] O. Lipps, "Panel Surveys: advantages and disadvantages compared with repeated cross-sectional surveys. FORS Guide No. 14, Version 1.0. Lausanne (FORS).," no. 14, pp. 1–13, 2021, https://doi.org/10.24449/FG-2021-00014

[13] T. K. Maity and A. K. Pal, "Subject specific treatment to neural networks for repeated measures analysis," Lecture Notes in Engineering and Computer Science. Accessed: Jun. 12, 2023. Available: https://www.iaeng.org/publication/IMECS2013/IMECS2013_pp60-65.pdf

[14] R. S. Gandhi, T. V. Raja, A. P. Ruhil, and A. Kumar, "Artificial neural network versus multiple regression analysis for prediction of lifetime milk production in sahiwal cattle," *Journal of Applied Animal Research*, vol. 38, no. 2. pp. 233–237, 2010. https://doi.org/10.1080/09712119.2010.10539517

[15] O. Gorgulu, "Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks," *S. Afr. J. Anim. Sci.*, vol. 42, no. 3, 2012, https://doi.org/10.4314/sajas.v42i3.10

[16] W. Grzesiak, R. Lacroix, J. Wójcik, and P. Błaszczyk, "A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records," *Can. J. Anim. Sci.*, vol. 83, no. 2, pp. 307–310, 2003, https://doi.org/10.4141/A02-002

[17] M. D. Murphy, M. J. O'Mahony, L. Shalloo, P. French, and J. Upton, "Comparison of modelling techniques for milk-production forecasting," *J. Dairy Sci.*, vol. 97, no. 6, pp. 3352–3363, 2014, https://doi.org/10.3168/jds.2013-7451

[18] V. Hassija *et al.*, "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024, https://doi.org/10.1007/s12559-023-10179-8

[19] E. J. Pedersen, D. L. Miller, G. L. Simpson, and N. Ross, "Hierarchical generalized additive models in ecology: An introduction with mgcv," *PeerJ*, vol. 2019, no. 5, 2019, https://doi.org/10.7717/peerj.6876

[20] V. Mamouridis, "Additive Mixed Models applied to the study of red shrimp landings : comparison between frequentist and Bayesian perspectives. University of Coruna." Accessed: Mar. 10, 2021. Available: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_393.pdf

[21] The United Republic of Tanzania, "Tanzania National Panel Survey (NPS) Waves 3, 4, and 5 (2012-2021). National Bureau of Statistics." Accessed: Oct. 02, 2022. Available: https://microdata.worldbank.org/index.php/auth/login/?destination=catalog/2252/get-microdata

[22] J. Hair, R. Anderson, B. Babin, and W. Black, "Multivariate Data Analysis." Accessed: Feb. 27, 2020. Available: https://www.drnishikantjha.com/papersCollection/Multivariate Data Analysis.pdf

[23] M. W. Watkins, "Exploratory Factor Analysis: A Guide to Best Practice," *J. Black Psychol.*, vol. 44, no. 3, pp. 219–246, 2018, https://doi.org/10.1177/0095798418771807

[24] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, 2020, https://doi.org/10.12691/ajams-8-2-1

[25] M. G. Pizarro, V. Landi, F. J. Navas González, J. M. León, and J. V. Delgado, "Non-parametric analysis of the effects of αS1-casein genotype and parturition non-genetic factors on milk yield and composition in Murciano-Granadina goats," *Ital. J. Anim. Sci.*, vol. 18, no. 1, pp. 1021–1034, 2019, https://doi.org/10.1080/1828051X.2019.1611388

[26] G. Shmueli, "To explain or to predict?," *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, 2010, https://doi.org/10.1214/10-STS330

[27] Z. Bonza, R. Katapa, and A. Msengwa, "Bayesian predictive modelling to ascertain factors affecting cattle milk production in Tanzania: Evidence from the national panel surveys 2012 – 2021," *Vet. Anim. Sci.*, vol. 27, 2024, https://doi.org/https://doi.org/10.1016/j.vas.2024.100404

[28] J. H. Carag, C. K. Lesorogol, and L. L. Iannotti, "Global nutrition security: The role of Veterinary Services," *OIE Rev. Sci. Tech.*, vol. 40, no. 2, pp. 523–532, 2021, https://doi.org/10.20506/rst.40.2.3242

[29] I. Kashoma and A. Ngou, "Insight into the adoption and success of artificial insemination services in smallholder dairy farming systems: A cross-sectional study.," Tanzania Veterinary Journal. Accessed: Nov. 30, 2024. Available: https://tvj.sua.ac.tz/vet2/index.php/TVJ/article/view/596

[30] Y. P. Nagagi, E. G. Kimaro, and V. Temba, "Practical application and the possible emergence of tick resistance to commonly usedacaricides in various districts of Tanzania," Livestock Research for Rural Development. Accessed: Oct. 09, 2024. Available: http://lrrd.cipav.org.co/lrrd32/8/petnag32127.html

[31] J. S. Chang'a, R. H. Mdegela, R. Ryoba, T. Løken, and O. Reksen, "Calf health and management in smallholder dairy farms in Tanzania," *Tropical Animal Health and Production*, vol. 42, no. 8. pp. 1669–1676, 2010. https://doi.org/10.1007/s11250-010-9619-x

[32] M. Catangui, "Dewormers for beef and diary cattle." Accessed: Dec. 19, 2024. Available: https://www.worldcourier.com/jp/sitecore/コンテンツ/mwianimalhealth/home/our-insights/-/media/assets/mwianimalhealth/insights-articles/doc_dewormersforbeefanddairycattle.pdf

[33] T. Smith, "Proper Timing And Proper Products Needed For Deworming Pasture Cattle." Accessed: Dec. 19, 2024. Available: https://www.beefmagazine.com/cattle-disease/proper-timing-and-proper-products-needed-for-deworming-pasture-cattle