








ARTICLE

## Dealing with highly kurtotic count data with excess zeroes: comparing different treatments in the control of dairy cattle gastrointestinal parasites

 Fernanda V. Roquim,<sup>\*</sup><sup>1</sup>  Renato R. Lima,<sup>2</sup>  Luiz R. Nakamura,<sup>2</sup>  Thiago G. Ramires,<sup>3</sup> and  Yuly A.C. Blanco<sup>4</sup>

<sup>1</sup>Universidade Federal de Lavras, Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Lavras-MG, Brazil

<sup>2</sup>Universidade Federal de Lavras, Departamento de Estatística, Lavras, Minas Gerais, Brazil.

<sup>3</sup>Universidade Tecnológica Federal do Paraná, Departamento de Matemática, Apucarana-PR, Brazil

<sup>4</sup>Universidad Cooperativa de Colombia UCC. Programa Académico de Medicina Veterinaria y Zootecnia, Santander, Colombia

\*Corresponding author. Email: fer\_venturato@yahoo.com.br

(Received: November 22, 2022; Revised: March 22, 2023; Accepted: April 14, 2023; Published: December 1, 2023)

### Abstract

In this paper we provide a flexible statistical framework to compare different treatments, conventional and selective strategic, in the gastrointestinal parasites control of dairy cattle through the count of *Giardia* cysts. Distributional regression models are considered, allowing the modelling of not only the average of these counts, but also the extra probability that calves do not present any cysts. Our findings show a positive relationship between the count of cysts with the animal body temperature and in animals until the age 150 days (and then, the count decreases). Higher responses are observed during summertime. Animals submitted to the selective strategic treatment present a lower count of cysts than the ones submitted to the conventional option. Further, there is a smaller chance that the sample does not present any cyst during wintertime. Conversely, this chance increases if the animal is submitted to the ST and in the earliest ages. Finally, the probability that no cysts are observed in the sample is roughly constant up to 40 °C and then rapidly increases. Hence, distributional regression models provide a great alternative to explicitly select features to model different aspects (average and extra probability of zero) of the count of *Giardia* cysts.

**Keywords:** *G. duodenalis*; GAMLSS; Parasitic diseases; Statistical modelling

## 1. Introduction

Brazil is considered one of the largest milk producers in the world (Barbero *et al.*, 2022). Livestock management has been improved in the past years, increasing the quality due to disease control. However, climatic changes and incorrect use of medicines are causing the increasing of parasitic diseases with high resistance, and consequently, increasing costs, implying millionaire losses in beef production (Oliveira *et al.*, 2017).

*Giardia duodenalis* is a unicellular protozoan (Plutzer *et al.*, 2010) and in cattle clinical infection mainly affects young animals (Geurden *et al.*, 2010). Giardiasis is a public health problem (Coelho *et al.*, 2017), with outbreaks reported by the use of contaminated water for consumption or recreation with parasite cysts and estimated to cause 280 million diarrhoea infections annually (Einarsson *et al.*, 2016; Efstratiou *et al.*, 2017). Its cycle is related to the gastrointestinal tract of vertebrate hosts (Xiao & Fayer, 2008), generating intestinal disorders, with diarrhoea that does not respond to the treatment using antibiotics and coccidiostats (Geurden *et al.*, 2010). The control of this disease is characterised using prophylactic treatments such as fenbendazole, showing a control of giardiasis in young calves (Blanco *et al.*, 2017).

Blanco (2015) presented a study where they compared the efficiency between a so-called selective strategical treatment (ST) against the conventional treatment (CT) in gastrointestinal parasites control (*G. duodenalis*) of dairy cattle through the count of *Giardia* cysts. Due to some of its characteristics, the authors performed a transformation and used as the response, a binary variable, which may imply a loss of information and negative consequences, leading researchers to a not reliable conclusion. In fact, in Blanco's dissertation, the author has not detected any difference between both treatments (one of its main hypotheses).

Hence, in this paper we revisit the data available in Blanco (2015) and propose a different and more robust statistical framework in order to understand whether the ST performs better than the CT, considering all the available information in the variability of count of *Giardia* cysts, i.e., instead of simplifying this variable into a dichotomous response, we model its original values using a distributional regression approach. In addition to which treatment was applied to explain the number of cysts, we also consider the season the data was collected and the age, body weight and height of the calf as possible candidates to explain such behaviour.

## 2. Material and methods

### 2.1 The data

In this paper we are considering the same data set studied by Blanco (2015). In the author's experiment, the efficiency of a selective strategical treatment (ST) against the conventional treatment (CT) in gastrointestinal parasites control of dairy cattle has been evaluated. The efficiency of both treatments was measured by the count of *Giardia* cysts.

Faecal samples from the heifers (30 up to 50 g) were collected directly from the rectum every 15 days, massaging the rectal walls using disposable gloves, which were then removed in an inverted way, eliminating the air, tied and identified with their respective name and number of the animal, with further analyses at the Laboratory of Parasitic Diseases of DMV/UFLA. These stool samples were used to count cysts per gram of stool based on Gordon & Whitlock (1939), and the zinc-sulphate flotation technique at 33% to search for cysts of *Giardia* spp.

Positive samples of *G. duodenalis* were detected on both ST (115/365, 31.51%) and CT (117/314, 37.26%). Regardless of the applied treatment, the dry season returned the highest number of positive samples (137/365, 38.16%). Once detected positive, an antiparasitic with the active ingredient of Fenbendazole at 10% 3 mL/20 kg pv (15 mg/kg) (Panacur® at 10%) was administered orally.

In addition to the two treatments (ST and CT), the following explanatory variables were considered as possible candidates to explain the response:

- treatment: a two-levels factor indicating whether the calf received the CT or ST;
- season: a four-levels factor indicating in which season the data was collected (summer, autumn, winter or spring);
- age: a quantitative variable indicating the age of the calf;
- weight: a quantitative variable indicating the body weight of the calf;
- temperature: a quantitative variable indicating the body temperature of the calf.

Due to the complexity of these data, we have used a very flexible class of regression models known as distributional regression models, which as stated by Heller *et al.* (2022), were first proposed as generalised additive models for location, scale and shape (GAMLSS) (Rigby & Stasinopoulos, 2005). The key advantage is that they are able to consider different flexible distributions and distinct regression structures (subsets of covariates) in each of these distribution parameters.

## 2.2 Statistical modelling

GAMLSS are a very flexible semiparametric class of regression models, which involves a distribution for the response variable and may involve non-parametric smoothing terms when modelling any and/or all parameters of the response variable distribution, generalising the well-established generalised linear (Nelder & Wedderburn, 1972) and generalised additive (Hastie & Tibshirani, 1990) models. Mathematically, let  $Y \sim \mathcal{D}(\boldsymbol{\theta})$ , where  $\mathcal{D}$  is the response variable density and  $\boldsymbol{\theta}$  is its vector of parameters of length  $p$ . The GAMLSS can be written as

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{jk}(\mathbf{x}_{jk}), \quad (1)$$

for  $k = 1, \dots, p$ , where  $g(\cdot)$  is a known monotonic link function,  $\eta_k$  denotes a predictor,  $\mathbf{X}_k$  is a known design matrix,  $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \dots, \beta_{J_k k})$  is a parameter vector of length  $J_k k$  and  $s_{jk}(\cdot)$  is a smooth non-parametric function of explanatory variable  $\mathbf{x}_{jk}$ , e.g., P-splines (Eilers & Marx, 1996; Eilers *et al.*, 2015). We shall highlight that when a smoothing term is included in the model, its coefficient (and standard error) should not be interpreted, as stated by Ramires *et al.* (2019).

The estimation process of GAMLSS consists in maximising the penalised log-likelihood function, which is achieved through three different options, based on an iteratively reweighted penalised least squares algorithm: CG algorithm, RS algorithm and a combination of both methods (Stasinopoulos *et al.*, 2017).

### 2.2.1 Selecting the response variable distribution

As it will be highlighted in Section 3, we are dealing with count data characterised by excess zero responses and long tail. Hence, a natural choice here is to select zero inflated distributions to explain the response variable counts of *Giardia* cysts. The most used distributions in literature with this feature are the zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) distributions (some examples of their applications may be seen in Morales & Higuchi (2018) and Ostfeld & Keesing (2018)). Nevertheless, in this paper we will also consider the zero inflated Poisson inverse Gaussian (ZIPIG) distribution that has longer upper tails than the ZINB distribution.

Within the GAMLSS framework and the `gamlss` package (Stasinopoulos & Rigby, 2007) in R (R Core Team, 2022), most distributions are reparametrized (Rigby *et al.*, 2019) to facilitate the interpretation of the results. The ZIP distribution has two parameters where  $\mu > 0$  and  $0 < \sigma < 1$  are the mean of the Poisson component and the inflated probability that  $Y = 0$  (i.e., there are no

cysts found and the calf is healthy), respectively. Its probability mass function (pmf) is given by

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma + (1 - \sigma) e^{-\mu} & , \text{ if } y = 0 \\ (1 - \sigma) \frac{e^{-\mu} \mu^y}{y!} & , \text{ if } y = 1, 2, 3, \dots \end{cases}$$

On the other hand, ZINB is a three-parameter distribution with pmf given by

$$P(Y = y|\mu, \sigma, \nu) = \begin{cases} \nu + (1 - \nu) P(Y_1 = 0|\mu, \sigma) & , \text{ if } y = 0 \\ (1 - \nu) P(Y_1 = y|\mu, \sigma) & , \text{ if } y = 1, 2, 3, \dots \end{cases} \tag{2}$$

where  $\mu > 0$  is the mean of the negative binomial component,  $\sigma > 0$  is a dispersion parameter related to the negative binomial component,  $0 < \nu < 1$  is the extra probability that  $Y = 0$  and

$$P(Y = y|\mu, \sigma) = \frac{\Gamma\left(y + \frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right) \Gamma(y + 1)} \left(\frac{\mu\sigma}{1 + \mu\sigma}\right)^y \left(\frac{1}{1 + \mu\sigma}\right)^{\frac{1}{\sigma}}$$

and  $\Gamma(\omega) = \int_0^\infty t^{\omega-1} \exp(-t) dt$  is the gamma function.

The ZIPIG is a three-parameter distribution as well, and its pmf can also be expressed as (2). However, in this case,  $\mu > 0$  is the mean of the Poisson inverse Gaussian (PIG) component,  $\sigma > 0$  is the dispersion of the PIG component,  $0 < \nu < 1$  is the extra probability that  $Y = 0$ ,

$$P(Y = y|\mu, \sigma) = \sqrt{\frac{2\alpha}{\pi}} \frac{\mu^y e^{\frac{1}{\sigma}} K_{y-0.5}(\alpha)}{y! (\alpha\sigma)^y}$$

$\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$  and  $K_\lambda(t) = 0.5 \int_0^\infty x^{\lambda-1} \exp[-0.5t(x + x^{-1})] dx$  is the Bessel function of the second kind.

### 2.2.2 Selecting regression structures and residual analysis

In summary, the main advantages of the GAMLSS model are that any distribution can be assumed to model the response variable and we may select different subsets of explanatory variables to explain each of its parameters (e.g., a parameter related to the inflated probability that no cysts are observed). These regression structures selection can be performed by several distinct methods. Within GAMLSS framework, the most frequently used method for model selection is a stepwise-based procedure (called Strategy A) which uses the generalised Akaike information criterion (GAIC) (Voudouris *et al.*, 2012) given by  $GAIC(\kappa) = -2 \times \hat{l} + \kappa \times df$ , where  $\hat{l}$  is the fitted log-likelihood function,  $df$  are the effective degrees of freedom of the fitted model and  $\kappa$  is a penalty. If  $\kappa = 2$  or  $\kappa = \log(n)$ , it reduces to the Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978), respectively. For further details regarding the stepwise-based procedure in GAMLSS, check Nakamura *et al.* (2017), Stasinopoulos *et al.* (2017) and Ramires *et al.* (2021b).

The adequacy of the final fitted model can be checked through the normalised randomised quantile residuals (Dunn & Smyth, 1996) within the GAMLSS framework. The main advantage of these residuals is that despite the distribution assumed to explain the response variable, if the fitted model is reasonable, they follow a standard normal distribution. As stated by Stasinopoulos *et al.* (2017), when dealing with a discrete response, they can be defined as

$$r = \Phi^{-1}(u),$$

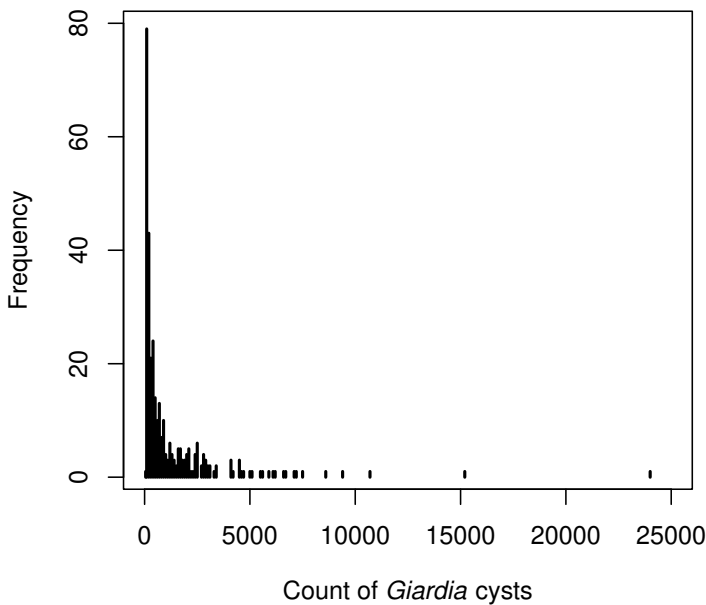
where  $\Phi^{-1}(\cdot)$  is the inverse cumulative distribution function of the standard normal distribution,  $u$  is defined as a random value from the uniform distribution on the interval  $[u_1, u_2] = [F(y - 1|\theta), F(y|\theta)]$  and  $F(y|\theta)$  is the model cumulative distribution function. Because of this randomization, it is advisable to compute the median of a collection of residual realisations.

### 3. Results and discussion

Table 1 displays some descriptive statistics for the response variable. We note that the count of *Giardia* cysts presents a large range as its minimum and maximum values are zero and 24,000, respectively. The average count of cysts is 605.33 and its standard deviation (SD) is 1,670.30. This variable is highly positively skewed (6.81) and has an extreme excess of kurtosis (70.55), i.e., the count's density is very asymmetrical and heavy-tailed (leptokurtic) as displayed in Figure 1. It is noteworthy that more than 50% of the observations are zeroes (342 cases) and only 25% of the count are higher than 400 (even though the maximum value is 24,000). The high number of outliers present in the data is a characteristic of this variable rather than a measurement error.

**Table 1.** Some descriptive statistics for the count of *Giardia* cysts

Mean	SD	Skewness	Kurtosis	Minimum	1st Qu.	Median	3rd Qu.	Maximum
605.33	1,670.30	6.81	70.55	0.00	0.00	0.00	400.00	24,000.00



**Figure 1.** Bar plot of the count of *Giardia* cysts. For clarity of presentation zeroes (342 cases) are not displayed.

Figure 2 displays plots of the response against each of the considered explanatory variables, which may give us an idea of the relationships in the data despite the interplay between the covariates. Panels (a) and (b) show boxplots of the explanatory factors where animals submitted to the conventional

treatment and during summer presented higher counts than the other levels. Panels (c) and (d) display the response against age and body weight, respectively, where we can note that the highest counts were collected in individuals where the date of collection is in the age range of 100 to 200 days and with 100 to 150 kg. Finally, as expected, higher counts were detected in animals that present higher body temperatures, as can be seen in Panel (e). It is noteworthy the presence of excess zeroes in the response variable that can be checked in all these panels as already highlighted in Table 1.

We then proceed to the distributional regression approach. The stepwise-based procedure (Strategy A) was performed in all GAMLSS based on the ZIP, ZINB and ZIPIG distributions. Table 2 displays values of the global deviance (GD, equals minus twice the fitted log-likelihood), AIC and BIC for each model, which were used to compare the fitted model. As we can see, the ZIPIG model clearly outperformed the others since it returned the lowest AIC and BIC values (5,689 and 5,808, respectively).

**Table 2.** Comparison of different marginal fitted distributions to the response variable

Distribution	Global deviance	AIC	BIC
ZIPIG	5,636	5,689	5,808
ZINB	61,237	61,377	61,693
ZIP	368,121	368,266	386,593

The final ZIPIG fitted model based on the GAMLSS framework (1) is given by

$$\begin{aligned} \log(\hat{\mu}) &= -14.766 + 0.545 \text{ temperature} - 0.310[\text{if treatment=ST}] \\ &+ 0.098[\text{if season=spring}] + 0.548[\text{if season=summer}] \\ &+ 0.315[\text{if season=autumn}] + s(\text{age}) \\ \log(\hat{\sigma}) &= 0.621 \end{aligned} \tag{3}$$

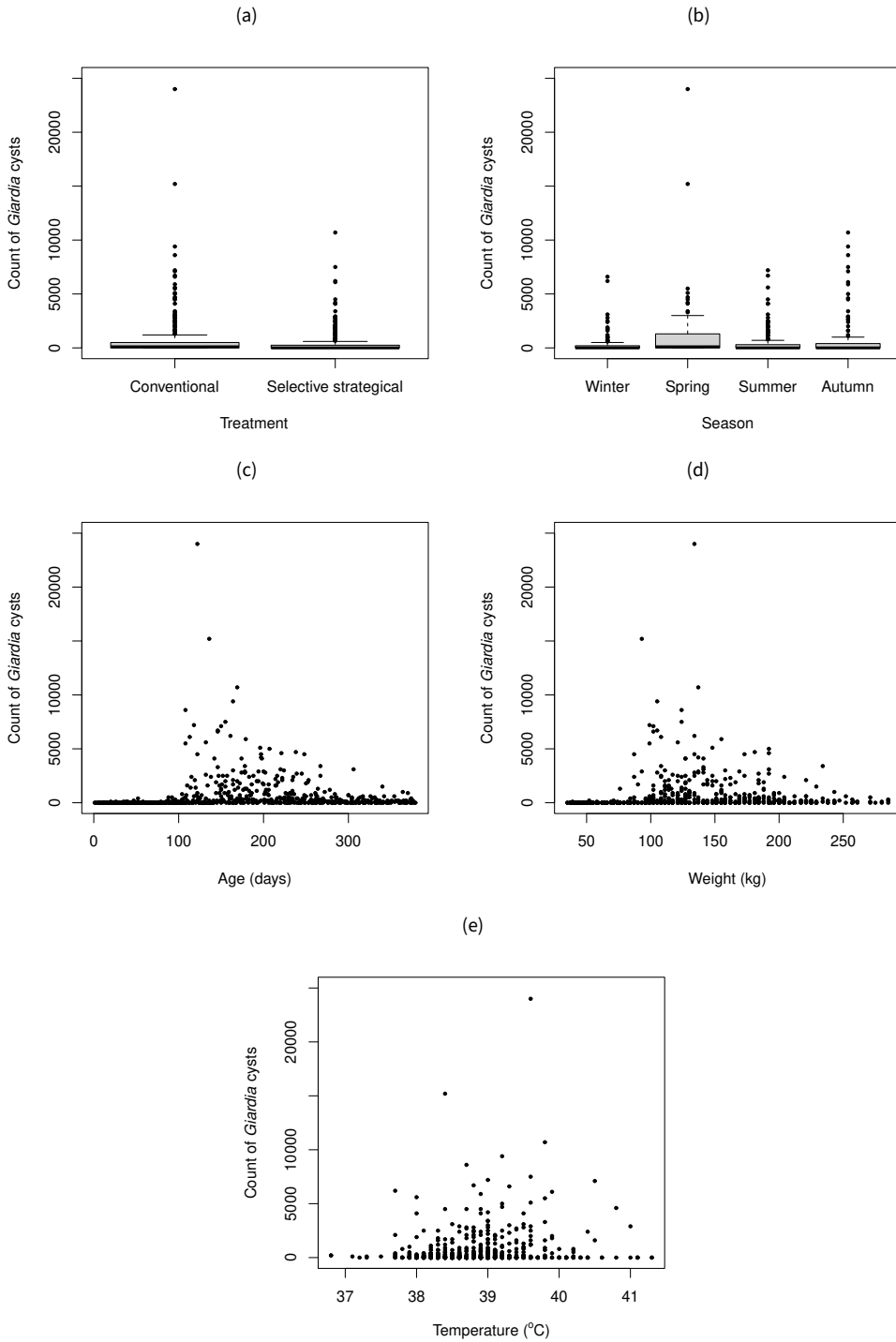
$$\begin{aligned} \text{logit}(\hat{\nu}) &= -1.697 + 0.613[\text{if treatment=ST}] + s(\text{age}) + s(\text{temperature}) \\ &- 0.455[\text{if season=spring}] - 0.878[\text{if season=summer}] \\ &- 0.627[\text{if season=autumn}]. \end{aligned}$$

All linear coefficients were statistically significant at a 5% level (i.e., p-value<0.05), except for spring and autumn levels in the  $\mu$  structure, indicating that there is no statistical difference between the observed average of count of *Giardia* cysts during these seasons and wintertime.

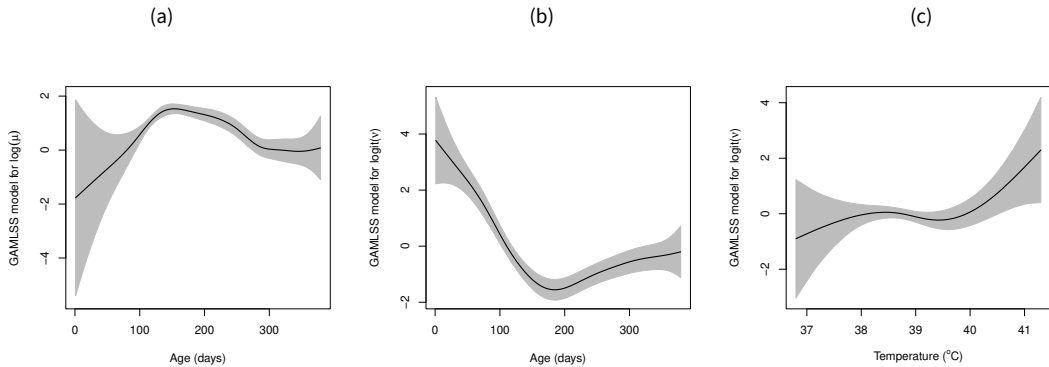
The fitted model for the mean of the PIG component  $\mu$ , in equation (3), indicates that, as expected, as the body temperature of the animal increases, the count of cysts also increases. More precisely, for each degree Celsius it is expected that the average count increases in  $\exp(0.545) = 1.725$  times, considering calves in the same condition (i.e., setting all other covariates as constants). Furthermore, in average, animals submitted to the selective strategical treatment present a lower count of *Giardia* cysts than the ones submitted to the conventional option.

Another covariate considered to model the mean of the PIG component was the season time. During summer, the average of counts is  $\exp(0.548) = 1.730$  times greater than the ones observed during wintertime. As previously stated, there are no statistical differences between the other levels and winter, which contradicts the findings of Ehsan *et al.* (2015), who found that there was no greater presentation of positive cattle for *Giardia* spp. During the humidity and precipitation season.

The last considered covariate to model the average of the counts is age. Since a non-linear effect of covariate age in the parameter was detected, a smoothing function (p-spline) was considered to



**Figure 2.** Relationship between the count of *Giardia* cysts and each of the covariates: (a) treatment; (b) season; (c) age; (d) weight; and (e) temperature.



**Figure 3.** Relationship between: (a) age against the average count of *Giardia* cysts, and (b) age and (c) temperature against the extra probability of zero (i.e., healthy animal), respectively.

model this relationship (Figure 3(a)). As we can see, the expected count of cysts increases until age 150 days, then decreases up to age 275 days, after which is roughly linear until age 365 days.

Regarding the dispersion parameter  $\sigma$ , it can be seen from equation (3) that no covariates were selected to compose its regression structure. Hence,  $\sigma$  was modelled as a constant, where  $\hat{\sigma} = \exp(0.621) = 1.861$ .

The extra probability of zeroes  $\nu$ , i.e., the probability that no cysts are observed in the sample, was modelled using four different covariates. Based on the estimated coefficients in equation (3), this probability will be higher if the calf is submitted to the selective strategical treatment. Regarding the seasons of the collection, during the summer there is a smaller probability of zero than in the other three seasons and this probability is higher during the winter. Smoothing functions were considered for both age and temperature in the  $\nu$  structure.

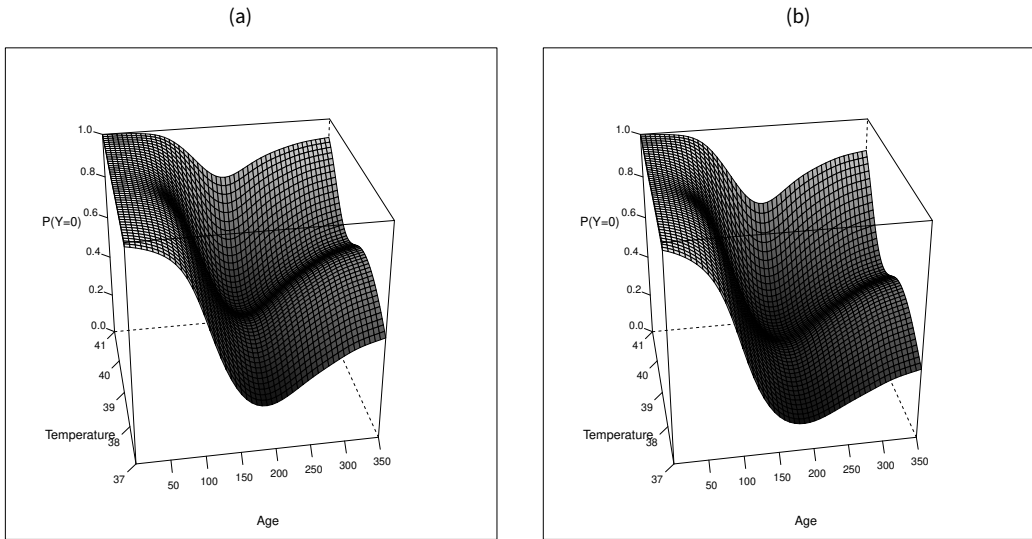
Cattle feeding varies according to their nutritional requirements based on race, sex and age, and in the higher phase of growth and consumption of these parasites, it increases in moderate amounts in the feed or pasture provided, with a lower risk of characteristic signs of the disease, as it allows the development of protective immunity in the animal (Taylor *et al.*, 2013). Volpato *et al.* (2017) concluded that the infection of *Giardia* spp. is higher in calves, which interferes with the animal's growth and low yield. Nonetheless, results in our paper show that there is a rapid decrease of the probability of zero up to age 200 days (the largest probabilities are observed during the early days of the animal age), and then a slight increase until age 365 days (Figure 3 (b)).

The extra probability that no cysts are observed is roughly constant up to 40 °C and then rapidly increases from that point (Figure 3(c)). The increase in temperature that have been shown by the animals may be directly related to the presence of other agents found when conducting this study, as presented by Blanco *et al.* (2017) and Barbieri *et al.* (2017), or good livestock practices that are of great relevance for the development of animals (Hernández-Gallo & Cortés-Vecino, 2012), so the presence of some symptoms cannot be directly related to parasitosis, since the loads found do not affect animal development.

In the way the GAMLSS is defined in equation (1), since we are modelling all parameters of the response variable distribution, different sets of values of the selected regression structures for each parameter may result in different shapes. Figure 4 summarises the behaviour of all predicted  $\nu$  values in our data, divided by the two distinct treatments applied to the animals. In general, the selective strategical treatment returned higher  $\nu$  values, suggesting that it may be more effective than the conventional treatment, in the sense that higher values of  $\nu$  imply a higher probability of counting zero *Giardia* cysts - which occurs when the calf is healthy - and a smaller average count.

As highlighted in Section 1, one of the main hypotheses tested by Blanco (2015) was how se-





**Figure 4.** Estimated probability that the calf is healthy fixing winter season, changing age and temperature, considering (a) selective strategical treatment and (b) conventional treatment.

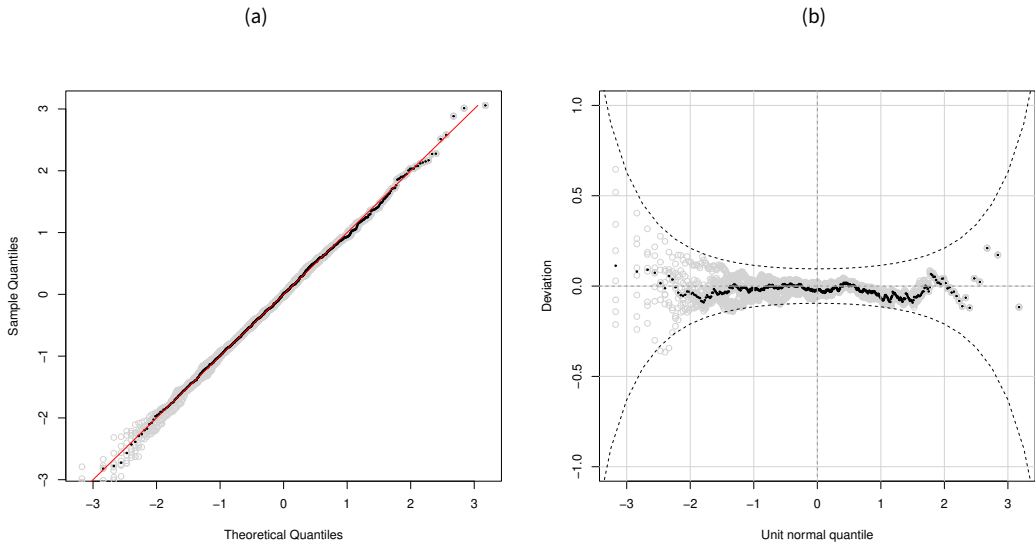
lective strategical treatment works over the conventional one in the count of *Giardia* cysts. Despite the author having not found any statistical difference between both treatments, as can be seen in equation (3), the covariate treatment was statistically significant to explain both average cysts count and the extra probability that the animal is healthy. Note that this finding was only possible since we are considering the GAMLSS framework, where different regression structures may be considered in all the response variable distribution parameters. As stated by Ramires *et al.* (2021a), considering a more sophisticated regression structure such as GAMLSS may result in a better (and more parsimonious) model than when using a more complex location model (e.g., modelling only the mean of cysts count).

Different combinations may be made to obtain the extra probability of zeroes and/or the expected number of cysts. Moreover, several statistical distribution shapes may be obtained depending on the combinations of covariates in both  $\mu$  and  $\nu$  parameters. Thus, an interactive plot to find these probabilities, values and shapes is provided by the authors through the link: <https://ramires.shinyapps.io/EPG-paper/>.

The adequacy of the final ZIPIG fitted model is performed through the normalised randomised quantile residuals. In order to graphically check these residuals, we display a set of randomised residual realisations (greyish dots) in a QQ-plot and a worm plot (van Buuren & Fredriks, 2001) in Figure 5. In the worm plot (Panel (b)), if the assumed distribution for the response variable is correct, then at least 95% of the median residuals (black points) must lie between the elliptical 95% pointwise interval band curves and no particular shape may be observed. With that said, and given that the median residuals (black points) in the QQ-plot (Panel (a)) are essentially linear, the final fitted model is reasonable and the ZIPIG regression model provides a good fit to these data.

## 4. Concluding remarks

This paper detailed the analysis of a positively skewed and highly kurtotic count data with excess zeroes from a real experiment. Although zero inflated Poisson and/or zero inflated negative binomial regression models may be promptly considered in such cases, we have shown that they were not able



**Figure 5.** (a) QQ-plot and (b) worm plot of the final ZIPIG fitted model based on the GAMLSS framework.

to model such characteristics, even under the generalised additive models for location, scale and shape (GAMLSS) framework, a very flexible class of semi-parametric regression models. In this sense, a distribution with longer upper tails than the previously mentioned ones was required, and thus the zero inflated Poisson inverse Gaussian (ZIPIG) was considered. Due to the ease of interpretation of each of the parameters of the ZIPIG distribution, based on the GAMLSS framework, we were able to understand how different features (age, temperature, treatment and season) impacted the average count of cysts and the extra probability that the animal is healthy. This analysis showed that GAMLSS are a competitive approach to model highly complex data and can be applied to other problems that face these situations.

## Acknowledgments

First and last authors gratefully acknowledge grants from CAPES (Brazil) and FAPEMIG (Brazil), respectively.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
2. Barbero, M. M. D., Fort, N. M., Schultz, E. B., Melo, A. L. P. & Moura, A. M. Estimation of genetic parameters in dairy production in girolando cattle. *Cien. Anim. Bras.* **23**, e-72300E (2022).
3. Barbieri, J. M., Blanco, Y. A. C., Lima, R. R., Lopes, M. A., Reis, E. M. B., Rocha, C. M. B. M., Coutinho, A. S. & Guimarães, A. M. Giardiasis as a neglected disease in Brazil: Systematic review of 20 years of publications. *PLOS Negl. Trop. Dis.* **11**, e0006005 (2017).

4. Blanco, Y. A. C. *Efeito e custos do tratamento estratégico seletivo no controle de parasitoses gastrointestinais em bezerras leiteiras* Master's dissertation (Federal University of Lavras, Brazil, 2015).
5. Blanco, Y. A. C., Barbieri, J. M., Lima, R. R., Lopes, M. A., Reis, E. M. B., Rocha, C. M. B. M., Coutinho, A. S. & Guimarães, A. M. Economic evaluation and efficacy of strategic-selective treatment of gastrointestinal parasites in dairy calves. *Rev. Bras. Parasitol. Vet.* **26**, 123–128 (2017).
6. Coelho, H. C., Durigan, M., Leal, D. A. G., Schneider, A. B., Franco, R. M. B. & Singer, S. M. Giardiasis as a neglected disease in Brazil: Systematic review of 20 years of publications. *PLOS Negl. Trop. Dis.* **11**, e0006005 (2017).
7. Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **5**, 236–244 (1996).
8. Efstratiou, A., Ongerth, J. E. & Karanis, P. Waterborne transmission of protozoan parasites: Review of worldwide outbreaks – An update 2011–2016. *Water Res.* **114**, 14–22 (2017).
9. Ehsan, A. M., Geurden, T., Casaert, S., Parvin, S. M., Islam, T. M., Ahmed, U. M., Levecke, B., Vercruyse, J. & Claerebout, E. Assessment of zoonotic transmission of Giardia and Cryptosporidium between cattle and humans in rural villages in Bangladesh. *Plos ONE* **10**, e0118239 (2015).
10. Eilers, P. H. C. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, 89–121 (1996).
11. Eilers, P. H. C., Marx, B. D. & Durbán, M. Twenty years of P-splines. *SORT* **39**, 149–186 (2015).
12. Einarsson, E., Ma'ayeh, S. & Svärd, S. G. An up-date on Giardia and giardiasis. *Curr. Opin. Microbiol.* **34**, 47–52 (2016).
13. Geurden, T., Vercruyse, J. & Claerebout, E. Is Giardia a significant pathogen in production animals? *Exp. Parasitol.* **124**, 98–106 (2010).
14. Gordon, H. & Whitlock, H. V. A new technique for counting nematode eggs in sheep faeces. *J. Counc. Sci. Ind. Res.* **12**, 50–52 (1939).
15. Hastie, T. J. & Tibshirani, R. J. *Generalized additive models* (Chapman & Hall, London, 1990).
16. Heller, G. Z., Robledo, K. P. & Marschner, I. C. Distributional regression in clinical trials: treatment effects on parameters other than the mean. *BMC Med. Res. Methodol.* **22**, 56 (2022).
17. Hernández-Gallo, N. & Cortés-Vecino, J. A. Cryptosporidium spp. and Giardia spp. prevalence and risk factors in dairy calves of the north-western zone of the Bogota Savanna. *Rev. Salud Publica* **14**, 169–181 (2012).
18. Morales, L. E. & Higuchi, A. Is fish worth more than meat? – How consumers' beliefs about health and nutrition affect their willingness to pay more for fish than meat. *Food. Qual. Prefer.* **65**, 101–109 (2018).
19. Nakamura, L. R., Rigby, R. A., Stasinopoulos, D. M., Leandro, R. A., Villegas, C. & Pescim, R. R. Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized t distribution. *J. Data Sci.* **15**, 221–237 (2017).
20. Nelder, J. A. & Wedderburn, R. W. M. Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **135**, 370–384 (1972).
21. Oliveira, P., Ruas, J. L., Riet-Correa, F., Coelho, A. C. B., Santos, B. L., Marcolongo-Pereira, C., Sallies, E. S. V. & Schild, A. L. Doenças parasitárias em bovinos e ovinos no sul do Brasil: frequência e estimativa de perdas econômicas. *Pesq. Vet. Bras.* **37**, 797–801 (2017).

22. Ostfeld, J. K. & Keesing, F. Impacts of large mammals on movements of the pouched mouse (*Saccostomus mearnsi*) in central Kenya. *Afr. J. Ecol.* **57**, 2–9 (2018).
23. Plutzer, J., Ongerth, J. & Karanis, P. *Giardia* taxonomy, phylogeny and epidemiology: facts and open questions. *Int. J. Hyg. Environ. Health* **213**, 321–333 (2010).
24. R Core Team. *R: A language and environment for statistical computing* Software available at <https://www.R-project.org>. 2022.
25. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Carvalho, R. J., Vieira, L. A. & Pereira, C. A. B. Comparison between highly complex location models and GAMLSS. *Entropy* **23**, 469 (2021).
26. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J. & Cordeiro, G. M. A new semiparametric Weibull cure rate model: fitting different behaviors within GAMLSS. *J. Appl. Stat.* **46**, 2744–2760 (2019).
27. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J., Rigby, R. A. & Stasinopoulos, D. M. Validation of stepwise-based procedure in GAMLSS. *J. Data Sci.* **19**, 96–110 (2021).
28. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. C Appl. Stat.* **54**, 507–554 (2005).
29. Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. & De Bastiani, F. *Distributions for modelling location, scale, and shape: using GAMLSS in R* (CRC Press, Boca Raton, 2019).
30. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
31. Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in *R*. *J. Stat. Softw.* **23**, 1–46 (2007).
32. Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. & De Bastiani, F. *Flexible regression and smoothing: using GAMLSS in R* (CRC Press, Boca Raton, 2017).
33. Taylor, M., Coop, R. L. & Wall, R. L. *Veterinary parasitology* 3rd (Wiley-Blackwell, New York, 2013).
34. van Buuren, S. & Fredriks, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Stat. Med.* **20**, 1259–1277 (2001).
35. Volpato, A., Tonin, A. A., Machado, G., Stefani, L. M., Campigotto, G., Glombowsky, P., Gasli, G. M., Favero, J. F. & Silva, A. S. Gastrointestinal protozoa in dairy calves: identification of risk factors for infection. *Rev. MVZ. Córdoba* **22**, 5910–5924 (2017).
36. Voudouris, V., Gilchrist, R., Rigby, R., Sedwick, J. & Stasinopoulos, D. Modelling skewness and kurtosis with the BCPE density in GAMLSS. *J. Appl. Stat.* **39**, 1279–1293 (2012).
37. Xiao, L. & Fayer, R. Molecular characterisation of species and genotypes of *Cryptosporidium* and *Giardia* and assessment of zoonotic transmission. *Int. J. Parasitol.* **38**, 1239–1255 (2008).