

**UNDERSTANDING PATTERNS OF AGGREGATION IN  
COUNT DATA**

**BY**

**PHUTI SEBATJANE**

**SUBMITTED IN ACCORDANCE WITH THE REQUIREMENTS OF THE  
DEGREE**

**MASTER OF SCIENCE**

**IN THE SUBJECT OF**

**STATISTICS**

**AT THE**

**UNIVERSITY OF SOUTH AFRICA**

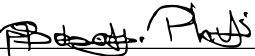


**SUPERVISOR: PROF PETER NJUHO**

**JUNE 2016**

## Declaration

I declare that “Understanding patterns of aggregation in count data” has been completed by me, the undersigned and that all the sources used or quoted have been indicated and acknowledged by means of a complete reference. I further declare that the work has not been submitted for the purpose of academic examination, either in its original or similar form, anywhere else.

Signature:   
Student name: Phuti Sebatjane  
Student number: 53755383

## Acknowledgements

I would like to thank the Agricultural Research Council for funding the initial part of my research and the University of South Africa for giving me a post graduate assistant position to help with my personal finances. I would also like to express my sincere gratitude to the following people:

Mogaswane K.H.R., Mtsali M.S. and Tsotetsi A. from the University of the North, Qwaqwa campus whose data formed the basis of my study.

My parents; Piet and Maselaelo Sebatjane, for bringing me to this world. A special thanks to my mother for all the hardwork and sacrifices, I am what I am today because you are. Like you always saw “perseverance is the mother of success”.

My sister Selaelo Sebatjane and her family, you have shown me such support and love, I honestly have no idea where I would be if it was not for you. No words can describe how grateful I am.

My partner Nkamogeleng Motswage and my very good friend Duduetsang Koloi, my life changed for the better in 2007 and 2015 when I met you.

My supervisor Prof P. Njuho, thank you so much for the time you put in our work and your balanced comments. I could not have done it without you.

My siblings Ezra, Ivy, Sello, Itumeleng and Makoena, I know we hardly say this but I love you.

**TO GOD THE FATHER THE SON AND THE HOLY SPIRIT, THANK  
YOU FOR CARRYING ME THROUGH.**

# Contents

<b>Contents</b>	<b>iii</b>
<b>List of Abbreviations</b>	<b>vi</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>viii</b>
<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Background . . . . .	2
1.2 Data Description . . . . .	3
1.3 Justification . . . . .	5
1.4 Purpose . . . . .	6
1.5 Problem statement . . . . .	6
1.6 Research objectives . . . . .	7
<b>2 Literature Review</b>	<b>8</b>
2.1 General approach to modelling count data . . . . .	8
2.2 Lognormal distribution and logistic regression . . . . .	8
2.3 Common models for count data . . . . .	9
2.4 Models for excess zeroes in count data . . . . .	13
2.5 Summary of the reviewed literature . . . . .	14
<b>3 Statistical Models on Count Data</b>	<b>17</b>
3.1 The exponential family . . . . .	18
3.2 The Poisson distribution . . . . .	18
3.2.1 Goodness of fit . . . . .	21
3.2.2 Model selection . . . . .	21
3.2.3 Overdispersion . . . . .	22
3.2.4 Model validation . . . . .	22
3.3 The negative binomial distribution . . . . .	23
3.4 Excess zeroes . . . . .	28
3.4.1 Zero inflated models . . . . .	29
3.4.2 Zero-altered models (hurdle models) . . . . .	30

<b>4</b>	<b>Fitting the Poisson, NB, ZIP and ZINB distributions</b>	<b>32</b>
4.1	Preliminary data exploration . . . . .	32
4.1.1	The presence of excessive zeroes in the data . . . . .	33
4.1.2	Variability of counts . . . . .	34
4.1.3	Characterising aggregation . . . . .	34
4.1.4	Nonparametric tests for differences between factors . . . . .	36
4.2	Analysis of <i>Cooperia isospora</i> egg counts . . . . .	41
4.2.1	The Poisson model . . . . .	41
4.2.2	The overdispersion test . . . . .	47
4.2.3	The negative binomial model . . . . .	48
4.2.4	Zero inflated models . . . . .	52
4.2.5	Zero inflated Poisson model (ZIP) . . . . .	53
4.2.6	Zero inflated negative binomial (ZINB) model . . . . .	56
4.2.7	Logistic regression model . . . . .	61
4.3	Analysis of Dictyocaulus filaria egg counts . . . . .	63
<b>5</b>	<b>Simulation Study</b>	<b>68</b>
5.1	Gibbs sampler . . . . .	69
5.2	Simulation example . . . . .	70
5.3	Evaluation of predictive performance . . . . .	71
<b>6</b>	<b>Zero Inflated Time Series Counts</b>	<b>74</b>
6.1	Observation-driven models . . . . .	74
6.1.1	ZIP Autoregression (ZIP-AR) . . . . .	75
6.1.2	ZINB Autoregression (ZINB-AR) . . . . .	76
6.2	Illustrative examples . . . . .	78
6.2.1	Application to <i>Haemonchus contortus</i> egg counts . . . . .	78
6.2.2	Application to <i>Fasciola hepatica</i> egg counts . . . . .	83
<b>7</b>	<b>Discussion and conclusion</b>	<b>85</b>
7.1	Discussion . . . . .	85
7.1.1	Quantifying aggregation and zero inflation . . . . .	85
7.1.2	Characterise distributions applicable to count data and assessing their performance under zero-inflation . . . . .	85
7.1.3	Assessing the nature of seasonality . . . . .	87
7.1.4	Significance of covariates in explaining FEC . . . . .	90
7.2	Conclusion . . . . .	90
	<b>Appendices</b>	<b>92</b>

<b>A</b>	<b>Moment Generating Function (MGF) of the Negative Binomial Dis- tribution</b>	<b>92</b>
<b>B</b>	<b>R Codes</b>	<b>94</b>
	<b>References</b>	<b>98</b>

## List of Abbreviations

AIC	Akaike Information Criterion
ACF	Autocorrelation Function
AR	Autoregressive
FEC	Faecal Egg counts
GLM	Generalised Linear Models
HSD	Honest Significant Difference
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
NB	Negative binomial
NBD	Negative binomial distribution
OD VAL	Optical Density Value
OLS	Ordinary Least Squares
PCV	Packed Cell Volume
PMF	Probability Mass Function
QP	Quasi-Poisson
ZANB	Zero Altered Negative Binomial
ZINB	Zero Inflated Negative Binomial
ZINB-AR	Zero Inflated Negative Binomial Autoregression
ZAP	Zero Altered Poisson
ZIP	Zero Inflated Poisson
ZIP-AR	Zero Inflated Poisson Autoregression

## List of Figures

1	Composition of rural and commercial livestock in South Africa ( <i>Source:</i> Department of Agriculture, Forestry and Fisheries, 2015) . . . . .	3
2	Parasite species frequency distribution . . . . .	38
3	<i>Coccidia eimeria</i> and <i>Ostertagia pinnata</i> boxplots . . . . .	39
4	Boxplots of all parasite species . . . . .	39
5	Observed and fitted values for <i>C. isospora</i> . . . . .	46
6	Poisson residual plot . . . . .	46
7	Negative binomial residual plot . . . . .	52
8	ZIP model residual plot . . . . .	56
9	<i>Cooperia isospora</i> rootogram (Poisson and NB model) . . . . .	59
10	<i>Cooperia isospora</i> rootogram (ZIP and ZINB model) . . . . .	60
11	Predicted probabilities of testing negative for an infection . . . . .	62
12	<i>Haemonchus contortus</i> time series plot (time in months) . . . . .	78
13	ACF and Partial ACF of <i>H. contortus</i> . . . . .	79
14	<i>H. contortus</i> observed and fitted egg counts . . . . .	81
15	<i>Fasciola hepatica</i> time series plot (time in months) . . . . .	83
16	ACF and Partial ACF of <i>F. hepatica</i> . . . . .	83



## List of Tables

1	Parasite species names and type . . . . .	4
2	Summary of the reviewed literature . . . . .	15
3	Variable description . . . . .	33
4	Parasite species index of discrepancy and variance to mean ratio . . .	36
5	Kruskal-Wallis Test . . . . .	40
6	Poisson model selection . . . . .	43
7	Poisson model parameter estimates . . . . .	44
8	Overdispersion test . . . . .	47
9	Negative binomial model selection . . . . .	49
10	Negative binomial model parameter estimates . . . . .	50
11	ZIP model selection . . . . .	53
12	ZIP model parameter estimates . . . . .	54
13	ZINB model selection . . . . .	57
14	ZINB model parameter estimates . . . . .	57
15	Logistic regression parameter estimates . . . . .	61
16	Predicted percentage of zeroes . . . . .	64
17	Poisson and negative binomial parameter estimates . . . . .	65
18	ZIP and ZINB model parameter estimates . . . . .	66
19	Mean and Variance changes as sample size (N) increase . . . . .	69
20	Simulation results with MSE of parameters . . . . .	71
21	Pearson's correlation ( $\rho_p$ ) and pseudo $R^2$ . . . . .	72
22	Observation-drive time series models for overdispersion and zero inflation	77
23	ZINB autoregression parameter estimates . . . . .	80
24	Frequency distribution of observed and expected <i>H. contortus</i> egg counts	81
25	5-step forecasting measures of accuracy . . . . .	82
26	ZIP-AR parameter estimates . . . . .	84
27	Parasite species AIC and AIC weights . . . . .	86
28	Model comparison for 15 parasite species: AIC . . . . .	89
29	Percentage of explaine deviance . . . . .	91

# Abstract

The term aggregation refers to overdispersion and both are used interchangeably in this thesis. In addressing the problem of prevalence of infectious parasite species faced by most rural livestock farmers, we model the distribution of faecal egg counts of 15 parasite species (13 internal parasites and 2 ticks) common in sheep and goats. Aggregation and excess zeroes is addressed through the use of generalised linear models. The abundance of each species was modelled using six different distributions: the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-altered Poisson (ZAP) and zero-altered negative binomial (ZANB) and their fit was later compared. Excess zero models (ZIP, ZINB, ZAP and ZANB) were found to be a better fit compared to standard count models (Poisson and negative binomial) in all 15 cases. We further investigated how distributional assumption affects aggregation and zero inflation. Aggregation and zero inflation (measured by the dispersion parameter  $k$  and the zero inflation probability  $\pi$ ) were found to vary greatly with distributional assumption; this in turn changed the fixed-effects structure. Serial autocorrelation between adjacent observations was later taken into account by fitting observation driven time series models to the data. Simultaneously taking into account autocorrelation, overdispersion and zero inflation proved to be successful as zero inflated autoregressive models performed better than zero inflated models in most cases.

Apart from contribution to the knowledge of science, predictability of parasite burden will help farmers with effective disease management interventions. Researchers confronted with the task of analysing count data with excess zeroes can use the findings of this illustrative study as a guideline irrespective of their research discipline. Statistical methods from model selection, quantifying of zero inflation through to accounting for serial autocorrelation are described and illustrated.

**Keywords:** Aggregations, autoregressive models, Akaike information criterion, correlation, count data, exponential family, generalised linear models, goats, internal parasites, hosts, negative binomial distribution, overdispersion, Poisson distribution, sheep, time series and zero inflation.

# 1 Introduction

In this section the following are outlined: background, justification, and purpose of the study together with the problem statement. The nature of the data is highlighted together with some inherent issues. We then propose statistical models that solve the problem of overdispersion, zero inflation and serial correlation in count data. Outlining research objectives and noting that the study is solely for illustrative purpose conclude this chapter.

## 1.1 Background

One of the major challenges faced by the agricultural sector today is that of food security. Food security not only relates to the availability and affordability of basic food products but also relates to the self-sufficiency of a nation in producing food and fibre products (McLeod et al., 2008). Both livestock products and produce (field crops and horticulture) contribute to food security. In relation to produce, livestock products make up 47% of the gross agricultural product (Meissner, et al., 2013). According to the abstract of agricultural statistics (Department of Agriculture, Forestry and Fisheries, 2015) there are 24.6 million sheep and 7.0 million goats in South Africa, both in rural and commercial farms. Of the total, 3.1 million sheep and 4.9 million goats are reared by rural farmers and not commercially (see Figure 1). Challenges these rural farmers face include among others, access to market and effective disease management. Diseases in livestock can be both bacterial, viral and also parasitic diseases. In this study the focus is on guiding parasites diseases control measures through understanding of how parasites are distributed among their hosts.

South African rural communities primarily practice subsistence farming. Livestock production as a result is directly linked to the wellbeing of the communities. Outbreak of an infectious livestock disease thus threatens not only the production but also the wellbeing of the farming community. Parasitology studies are characterised by count data, involving presence and absence of faecal eggs of different species, which occur at different time of the year.

Understanding the nature of the distribution of such data and the frequency of occurrence of the species would guide in timely introduction of control interventions. Faecal egg count are generally performed on livestock to monitor the level of parasitism in the herd, determine the effectiveness of certain treatments or determine which animals are resistant to parasite species of interest. In the area of quantitative

parasitology, counts data are commonly used to understand the interaction between parasites and their host. Crofton (1971) suggested the use of a frequency distribution, both in expressing a quantitative relationship between parasites and their hosts and in understanding prevalence and abundance patterns.

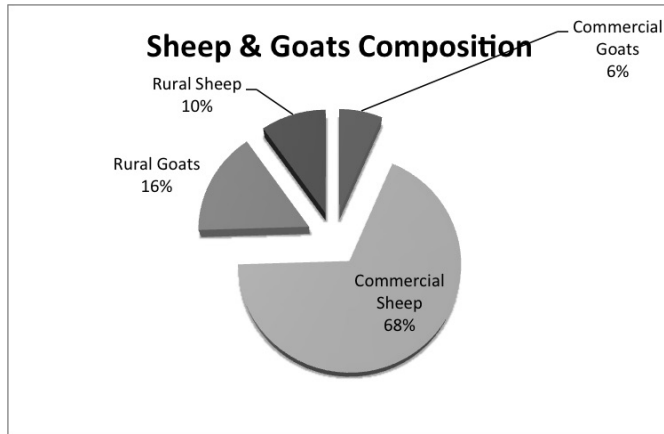


Figure 1: Composition of rural and commercial livestock in South Africa (*Source:* Department of Agriculture, Forestry and Fisheries, 2015)

While prevalence is the proportion of infected hosts among all hosts examined, abundance is the number of parasites found on all hosts. Both prevalence and mean abundance are important measures in indicating the level of parasitism in the herd. They give information on the skewness of the distribution and the proportion of zero counts. Most parasites are aggregated among their hosts, rendering traditional statistical methods ineffective. O'Hara and Kotze (2010) cautioned against the use of log transformation on count data, suggesting the use

of standard count models i.e. Generalised Linear Models (GLMs). The motivation behind this study is its multidisciplinary nature and the opportunity it presents in pulling together distributional theories and applying them to livestock faecal egg count (FEC) data. Understanding the distribution of these FEC will help in identifying when the host are most susceptible and thus the most appropriate time for remedial measures to ensure maximum herd productivity.

## 1.2 Data Description

The data in this study is observational and was collected from January 1998 to February 1999 in three different open grazing regions in the Free State province; Kenstell, QwaQwa and Harrismith. Faeces (dung) from identified animals were examined for parasite species eggs, each animal's dung was examined once. Upon examination, parasite species were identified and the following were recorded: the time of data

collection in months, the age, type (whether it was a sheep or goat) and sex of the ruminant. Blood test were also performed and packed cell volume and infection test results were also recorded, optical density and inhibition percentage were then determined. A total of fifteen different parasite species were identified while 495 animals (sheep and goats) were examined. We use the data on a secondary level and for illustrative purposes, the data was originally collected by Mogaswane K.H.R., Mtsali M.S. and Tsotetsi A. from the University of the North, Qwaqwa campus. Table 1 shows the name and type of parasite species that were identified.

Table 1: Parasite species names and type

Genus name	Known as	Abbreviation	Type
Boophilus decoloratus	B. decoloratus	BDE	Tick
Rhipicephalus evertsi evertsi	R.e. evertsi	REE	Tick
Haemonchus contortus	H. contortus	HAE	Roundworm
Cooperia curticei	C. curticei	COO	Roundworm
Coccidia eimeria	C. eimeria	EIM	Protozoa
Coccidia Isospora	C. Isospora	COO	Protozoa
Fasciola hepatica	F. hepatica	FAS	Fluke worm
Dictyocaulus filarial	D. filarial	DIC	Roundworm
Ostertagia pinnata	O. pinnata	OST	Roundworm
Trichostrongylus axei	T. axei	TRI	Roundworm
Trichuris ovis	T. ovis	TOV	Roundworm
Strongyloides papillosus	S. papillosus	STR	Roundworm
Toxoplasma gondii	T. gondii	TOX	Protozoa
Paramphistomum cervi	P. cervi	PAR	Flatworm
Oesophagostomum columbianum	O. columbianum	OES	Roundworm

The observed sheep and goat's faecal egg count shows that for 14 of the 15 parasite species has more than 50% of zero counts. Parasite counts tends to be overly dispersed (especially if the datasets are zero abundant) with a few hosts harbouring most of the parasite while the majority of the population has low counts. This poses problems in both descriptive and inferential statistics. In descriptive statistics, there

is a question of whether the mean is the best measure of location when the underlying distribution is aggregated. The sample mean tends to be overly influenced by large values in heavily skewed distribution while the geometric mean cannot be determined directly if the data set has zero values.

In inference, most readily used statistical procedures i.e. ordinary least square method in regression analysis, assumes normality of the error terms. When the distribution of faecal egg counts tends to be aggregated among their hosts, alternative methods for analysing the data are considered.

### 1.3 Justification

Count data with numerous zeroes is often heavily skewed and hardly conform to normality assumptions and conventional transformation approach does not resolve the problem. A general linear model addresses this concern. Count data mostly does not follow a normal distribution; log transformation becomes questionable in the event that the data is abundant with zeroes. A conventional technique such as ordinary least squares (OLS) in linear regression results in multiple assumptions being violated. Developed with the aim to address the limiting assumption of normality in linear models, the general linear model is an extension of linear models that allows one to specify the distribution of data via a link function. If parasite encounters with hosts are completely random then the faecal egg count per host are expected to follow a Poisson distribution. The overly dispersed nature of parasitology data again creates a problem in this regard as the variance is usually greater than the mean while a Poisson process has a variance equal to its mean. Given its nature to deal with the issue of over dispersion as compared to the Poisson distribution, the negative binomial distribution (NBD) has historically been widely used in modelling, quantifying or analysing parasite distribution (Gaba, et al., 2005). Distributions that are less restrictive on assumptions in handling of the count data in the presence of many zeroes are considered.

The aim is to package together the different distributions that are considered in quantifying parasite distribution in count data and also count data in other areas in general. Developing models which provide efficient estimates on abundance of disease causing parasites will enable implementation of interventions on the health life of livestock feasible and the livelihood of rural community will subsequently be improved.

The findings will provide a step by step quantitative guideline to researchers in areas of ecology and parasitology interested in modelling occupancy abundance or prevalence patterns. Researchers will be able to apply the simplest count data model (the Poisson distribution) and also move to more complex models that explain aggregation in count data. Problems arising from model misspecification such as inflated standard errors will be made aware to researchers. In a nutshell, results from this study will help researchers dealing with count data to apply a spectrum of models that could be applicable in their studies and provide them with a tool to choose the best model given their particular case.

## **1.4 Purpose**

Livestock keeping in South Africa plays a major role in the livelihood of rural communities. Some of these roles include; livestock keeping as a source of income and cultural activities that relates to both the wellbeing of the community and household. Other secondary roles include livestock keeping to create employment and for companionship purposes. Community grazing adapted by rural society provides ideal ground for parasitological diseases to flourish. Developing control systems through use of good statistical models will lead to timely implementation of the control interventions. This study tends to quantify patterns of different parasite species among their hosts and to investigate any changes in these patterns across time, host gender, host age and the type of livestock in question. Upon completion of this study researchers interested in quantifying count data, either for productivity or cost reduction reasons will use the findings. The researchers will be guided in choosing the best distribution applicable to their count data and steps in conducting the analysis.

## **1.5 Problem statement**

Livestock keeping in a communal grazing setup in certain parts of South Africa is characterised by animal diseases caused by parasitological species. The produce from these livestock is minimal while the mortality is high. Understanding the distribution of disease causing species through analysis of faecal count data characterised by zeroes, guide in formulation of effective interventions.

To understand the distribution of these infectious species, first we focus on the nature of the faecal count data. Some characteristics of the faecal count data require special attention before any analysis methods is identified. Exploring of the data indicated that a majority of the hosts (animal species) recorded low or zero egg counts while

a few recorded very high counts. This results in a phenomenon known as parasite aggregation among hosts. The presence of aggregation creates a problem in terms of applying the simplest count data model, which is the Poisson count model. Extra variation around the mean more than expected by the Poisson model requires a more flexible distribution such as the negative binomial distribution. Another consideration with the FEC data is the presence of too many zero counts. A way to deal with excessive zeroes is to employ zero inflated models which better account for the extra variation caused by the excess zeroes.

Data are available on 15 parasite species and the challenge in this study is investigating which distribution explains each parasite species and if indeed all the parasite species are aggregated among their host population. The procedure involves fitting some discrete probability distributions applicable for count data, namely; the Poisson, quasi-Poisson, Negative Binomial (NB), zero inflated Poisson (ZIP), zero altered Poisson (ZAP) and zero inflated negative binomial (ZINB) and zero altered negative binomial (ZANB) distribution (also called Hurdle Models). The last four address the problem of excessive zeroes.

## 1.6 Research objectives

Keeping in mind that the aim is to illustrate how to handle three main distinctive features (autocorrelation, overdispersion and zero inflation) of count data. The main objective is to review the existing distributions that fit aggregated count data and determine the appropriate model for parasitological data characterised by many zeroes. The following are the specific objectives:

- Quantify aggregation and zero inflation models.
- Characterise distributions applicable to count data.
- Assess the performance of these distributions in the presence of numerous zeroes.
- Determine the significance of covariates in explaining variation in the observed faecal egg counts.
- Check the consistency of the fitted models by simulating random observations from zero inflated distribution.
- Assess the nature of seasonality in the monthly data.



## 2 Literature Review

In this section a thorough review of literature is conducted. The review first looks at application of standard count models in both areas of ecology and parasitology. We then focus on studies that address both issues of zero inflation overdispersion and how they compare with standard Poisson and NB distributions. A summary of the literature review is provided at the end of this chapter.

### 2.1 General approach to modelling count data

The simple and most common initial approach to count data is assuming a Poisson counting process. The Poisson distribution assumes that counts per unit time or space are randomly distributed with the mean equal to the variance. While the Poisson distribution is favoured because of its simplicity, the downside is that it does not take into account overdispersion. This in turn leads to overestimation of standard errors, resulting in significant covariates that would have otherwise not been significant. A more flexible approach to count data is to use a negative binomial distribution. A negative binomial distribution assumes various quadratic mean-variance relationship. Different parameterisations of the negative binomial distribution results in the Negative Binomial1 (NB1), Negative Binomial2 (NB2) and Negative Binomial12 (NB12) models (later explained in section 2.3). The advantage of a negative binomial distribution over the Poisson distribution is ability to account for overdispersion. However, in the presence of excess zeroes the negative binomial distribution may not sufficiently explain the distribution of parasites among hosts. Inability of standard count models to account for zero inflation leads to the need to apply count models for excess zeroes. Generalised Linear Models (GLMs) that account only for zero inflation do not take into account possible correlation between latent variables. Yang, et al., (2015) proposed the use of parameter driven state-space time series models to account for temporal correlation. Zero inflated time series count models will not only account for overdispersion and zero inflation but also for possible correlation between observations (Maiti, et al., 2014).

### 2.2 Lognormal distribution and logistic regression

Using logistic regression and lognormal distribution, Baines, et al., (2015) found the factors; month, year, sex, age and group size to be significant in explaining the occurrence and seasonality of internal parasites elephants. For cases where prevalence was 100%, logistic regression was inappropriate, as counts cannot be reduced to binary

response but a single response. Instead a log transformation was done on counts and a linear regression was used to investigate the occurrence of parasite species. Sileshi (2008) applied logarithmic transformation to stabilise the variance and normalise the counts by taking the  $\log(count + 1)$ , a lognormal distribution was then used to explain the soil organism’s abundance patterns. Normality tests on the transformed counts indicated that, the transformation did not achieve the desired results on most of the data sets. The data were still not normally distributed upon transformation. Even though Sileshi (2008) and Baines, et al., (2015) performed log transformation on counts, we decided not to transform the parasite species egg counts. This is because transformations (either log transformation or square root transformation) always performed poorly compared the Poisson distribution (O’Hara, et al., 2010). O’Hara, et al., (2010) conducted a simulation to compare the performance of the negative binomial distribution (NBD) to that of the lognormal distribution. They found that NB model consistently performed well and effectiveness of transformation decreased with an increase in the number of zeroes. For this reason the lognormal distribution is not included in our study. However, a normality test on the transformed data is included (in the preliminary data exploration section) to support the exclusion of a lognormal distribution.

Apart from count models, Ziadinov, et al., (2010) added logistic regression to their analysis to model prevalence of infections. Prevalence of infection is the proportion of infected host among all host examined. The use of logistic regression is instinctive and largely depends on the aim of individual studies. Logistic regression condenses count data into binary data which potential hides crucial information in the data. With regard to predicting the prevalence pattern (i.e. the percentage of zeroes), logistic regression outperformed count models (Lewin, et al., 2010). While Bailey, Lopez, Camero, Taiquiri, Arhuay and Moore (2013) used logistic regression to investigate the risk factors associated with prevalence of parasitic infection in street children, Ajiferuke and Famoye (2015) used lognormal regression in modelling simulated counts response. Compared to lognormal and Poisson regression, the negative binomial regression was found to be a better fit mostly due to the overdispersed counts.

## 2.3 Common models for count data

The two most widely applied distributions in count data are the Poisson distribution (applicable when internal parasites are randomly distributed among their hosts) and the NB distribution (applicable in the event of overdispersion). When modelling rare

species (parasite affecting a smaller proportion of the herd or organism that are scattered in their environment), researchers historically applied standard count models. The problem with rare species data is the higher proportion of zeroes than expected by the standard count models. Lambert (1992) introduced the use of zero inflated models to handle the problem of excessive zeroes in count data. Ziadinov, et al., (2010) however concluded that excessive zeroes in count data does not necessarily imply the application of zero inflated models. Historically, the negative binomial distribution has been widely used to both quantify aggregation and analyse count data. This has mainly been due to its simplicity, the ability to deal with overdispersion and the availability of software. In this review the focus is on methods and procedures used in analysing the data and research findings as relevant to the objectives of our study. Papers on count data with overdispersion are reviewed together with count data with excessive zeroes. Objectives for count data with overdispersion were primarily centred on comparing the fit of the negative binomial distribution to other distributions with focus on estimating the effects of covariates.

Two of the research objectives are characterising distributions applicable to counts data and quantifying aggregation. Distributions applicable to count data according to Gaba, et al., (2005) include the NBD, log-normal, exponential and Weibull distribution. In addition to fitting the negative binomial model both Linden and Mantyniemi (2011) and Ver Hoef and Boveng (2007) added the Poisson model with a linear mean-variance relationship (Quasi-Poisson) in analysing count data distributions. A common tread throughout the study by Gaba, et al., (2005), Linden and Mantyniemi (2011) and Ver Hoef et al., (2005) is the use of maximum likelihood in parameter estimation and a choice of the negative binomial model as an initial approach in analysing count data. To quantify aggregation, Gaba et al., (2005) and Linden et al., (2011) applied the variance to mean ratio as an aggregation measure. The variance to mean ratio as the name suggests is calculated as the ratio between the variance and the mean. Aggregation occurs when there is more variation around the mean than expected by the Poisson distribution. A value greater than one for the variance-mean ratio is indicative of aggregation while values less than one indicate the absence thereof. Other measures of aggregation available in literature include the index of discrepancy but in most studies only the variance to mean ratio is used, e.g. Marques, et al., (2010) and Alexander (2012) used only the variance to mean ratio as aggregation measure of parasite counts.

While the focus with Gaba, et al., (2005) was to compare the NBD with other distribution, Linden and Mantyniemi (2011) compared only different parameterisa-

tion of the NBD with a wide range of mean-variance relationships. Derivation of the formula below shows how the negative binomial distribution is formulated with different variance to mean relationships. Let

$$\sigma^2 = \omega\mu + \theta\mu^2,$$

where  $\sigma^2$  and  $\mu$  are the variance and mean of the NB distribution, respectively. Both  $\omega$  and  $\theta$  are termed overdispersion parameters, with their different values giving rise to different parameterisation of the negative binomial distribution. We assume the probability mass function (PMF) of the NB distribution takes form:

$$P(Y = y) = \binom{y+r-1}{y} p^r (1-p)^y \quad y = 0, 1, 2, \dots$$

Using the moment generating function of the negative binomial distribution (NBD) in appendix 1, we derived the quadratic mean variance relationship of the NBD. From appendix 1;

$$\mu = \frac{r(1-p)}{p},$$

and

$$\sigma^2 = \frac{r(1-p)}{p^2}.$$

From the  $\mu$  expression:

$$\begin{aligned} \mu p &= r - rp \\ p(\mu + r) &= r \\ p &= \frac{r}{\mu + r}. \end{aligned}$$

It then follows that:

$$\begin{aligned} 1 - p &= 1 - \frac{r}{\mu + r} \\ &= \frac{\mu + r - r}{\mu + r} \\ &= \frac{\mu}{\mu + r}. \end{aligned}$$

Substituting  $(1 - p)$  into the expression for  $\sigma^2$  we obtain:

$$\begin{aligned}
\sigma^2 &= \frac{r(1-p)}{p^2} \\
&= \frac{r \left( \frac{\mu}{\mu+r} \right)}{\left( \frac{r}{\mu+r} \right)^2} \\
&= \frac{\mu r}{\mu+r} \frac{(\mu+r)^2}{r^2} \\
&= \frac{\mu(\mu+r)}{r} \\
&= \frac{\mu^2}{r} + \mu.
\end{aligned}$$

Replacing the aggregation parameter  $\frac{1}{r}$  with the aggregation parameter  $\theta$  and letting  $\omega = 1$  we obtain:

$$\sigma^2 = \omega\mu + \theta\mu^2.$$

Fixing the value of  $\theta$  at zero results in a linear variance mean relationship (Quasi-Poisson model also termed NB1). The value of  $\omega$  can also be fixed at one, resulting in a quadratic mean variance relationship (NB2). Another form of a negative binomial distribution (NB12) is obtained by not fixing either  $\omega$  or  $\theta$  but letting them take specified constant values. Apart from performing a comparative study, Linden and Mantyniemi (2011) also wanted to highlight the relationship between the standard count model (Poisson model) and the variations thereof at different overdispersion level. To decide on the choice of the model, environmental factors such as flocking patterns of the birds were used in conjunction with statistical measures such as the AIC.

The Akaike's Information Criterion (AIC) was used to select the best model that had the minimum AIC. Gaba, et al., (2005) took a further step by calculating AIC weights ( $w_i$ ), the probability that a model is the best one among the set of candidate models for the observed data. Let;

$$w_i = \frac{e^{-0.5\Delta_i}}{\sum_{j=1}^R e^{-0.5\Delta_j}},$$

where  $R$  is the number of models in a set and  $\Delta$  being the AIC difference, calculated as  $AIC_i - AIC_{min}$ , where  $AIC_{min} = \min[AIC_1, \dots, AIC_R]$ . AIC difference can also be used to determine the level of practical support to a model given the best model. Using these AIC differences, Gaba, et al., (2005) found the Weibull distribution to provide a better fit to most of the dataset followed by the NBD. The choice of the best model was not solely based on the AIC but also on a simulation study to determine the bias and consistency of estimators.

## 2.4 Models for excess zeroes in count data

Zero inflated models are introduced. In this section we propose to apply zero inflated models to test for zero inflation both in the presence and absence of overdispersion. In addition to characterising distributions applicable to count data, another objective is to quantify both aggregation and zero inflation. Vidyashankar, et al., (2012) suggested the use of zero inflated Poisson and zero inflated negative binomial distribution in modelling the resistance (pre and post treatment distributions) of equine gastrointestinal nematodes to treatment. Lewin, et al., (2010) studied fish catch data collected at different geographical sites, a large number of zero catches were observed. In addition to the standard count models, zero inflated, zero altered and logistic regression models were added to the analysis. The adequacy of each distribution was tested using the ratio or the deviance and the degrees of freedom (D/DF). Sileshi (2008) first checked the adequacy of standard count models using the ratio (D/DF) in his study on soil animal counts. Ratios greater than one indicate some degree of overdispersion unaccounted for and inadequacy to use the model under concern, whereas ratios around one indicate the appropriateness to apply the model. Both Sileshi (2008) and Vidyashankar, et al., (2012) found the Poisson distribution inadequate with large values of the ratio of deviance and degrees of freedom. Vuong tests (Vuong 1989, cited in Lewin, et al., 2010) were used to compare non-nested models. Vuong test is a likelihood-ratio-based test for model selection based on closeness of model to the true data. The models can be nested or non-nested.

Vuong test favoured models that accounted for excess zeroes each time two models (standard count model and count models for excess zeroes) were compared (Vuong 1987, cited in Lewin., et al, 2010). The null hypothesis that standard count models and count models for excess zeroes are equally close to the observed counts was rejected in all cases, favouring models that account for excess zeroes. To check overall statistical significance of covariates, Lewin, et al., (2010) compared the final regres-

sion models to the null models with only the intercept. All count models with full sets of covariates were better in explaining abundance pattern, however there was low improvement in McFadden  $R^2$  value. McFadden  $R^2$  is a pseudo coefficient of determination that is similar to the  $r^2$  in ordinary least squares (OLS) regression, the low improvement in this value from the full to the null model is an indication that some factors that affect the abundance pattern might not have been recorded Lewin, et al., (2010).

## 2.5 Summary of the reviewed literature

Table 2 provides the summary of the literature review covered. It should be noted that only the purpose and conclusions relevant to this current study have been included in the review summary table. From the review, apart from Lewin et al., (2010), authors either compared two distributions, only used standard count models or zero inflated models in modelling prevalence patterns. In this study we continue with sets of distribution applied by Lewin., et al., (2010) but apply it in areas of parasitology instead of ecology. We use this range of count data models to provide guideline to researchers in characterising aggregation and modelling prevalence and abundance.

Table 2: Summary of the reviewed literature

Citation	Counts	Area	Purpose	Conclusion
Ajiferuke, et al., (2015)	Simulated counts	Statistics	Simulation study	Overdispersed counts were better modelled by the negative binomial distribution.
Baines, et al., (2015)	Internal parasite sites	Parasitology	Modelling occurrence and seasonality of internal parasites in wild elephants.	Age, group size, sex month and year were found to be significant covariates.
Yang, et al., (2015)	Work place injuries	Biostatistics	Developing modelling framework to accommodate three features; overdispersion, zero inflation and autocorrelation	Dynamic zero inflated Poisson model was found to be adequate in modelling all three features.
Maiti, et al., (2014)	Simulated counts	Statistics	Comparing stationarity and autocorrelation structures.	In the presence of excess zeroes, zero inflated Poisson auto-regressive process of order 1 performed better.
Vidyashankar, et al., (2012)	Gastrointestinal nematodes	Veterinary Parasitology	Investigating anthelmintic resistance to control equine gastrointestinal nematodes.	Both pre and post treatment distribution were found to be zero inflated (with the zero inflated Poisson providing a better fit).
Linden, et al., (2011)	Birds Migration	Ecology	Model prevalence using different parametrisation of the NBD.	NBD provide better fit compared to Poisson distribution.
Vaudor, et al., (2011)	Freshwater fish	Ecology	Comparative study.	NBD provided a better fit to 58% of the datasets compared to zero inflated models.
Lewin, et al., (2010)	Fish Catch	Ecology	Modelling counts with excess zeroes.	Hurdle models and zero inflated models performed better than standard count models.

*Continued on next page*



Table 2 – *Continued from previous page*

<b>Citation</b>	<b>Counts</b>	<b>Area</b>	<b>Purpose</b>	<b>Conclusion</b>
O'Hara, et al., (2010)	Simulated Counts	Statistics	Comparative study.	Log transformed counts provides worst fit compared to count models.
Ziadinov, et al., (2010)	Foxes Internal Parasites	Veterinary Parasitology	Modelling prevalence patterns.	Zero inflated models can be dispensable in modelling count data with excess zeroes.
Sileshi (2008)	External Parasites	Parasitology	Comparative study: Modelling counts with excess zeroes.	Despite a high proportion of zeroes in the data, NBD was a better fit to 75% of the datasets compared to excess zero models.
Ver Hoef, et al., (2007)	Harbour Seal	Ecology	Comparative study. Different mean-variance relationships.	Comparative study. Quasi-Poisson provided better fit than NBD.
Gaba, et al., (2005)	Sheep Macroparasites	Parasitology	Characterise Aggregation	Variance to mean ratio, aggregation parameter of the NB distribution and scale parameter of the Weibull distribution used to characterise aggregation.
Shaw, et al., (1998)	Wildlife Macroparasites	Parasitology	Compare performance of NBD with Poisson distribution.	NBD better describes aggregated counts that Poisson distribution.
Lambert (1992)	Defects	Manufacturing	Apply a ZIP model to counts of defected wiring boards.	ZIP predicted defects better than standard count models.
Crofton (1971)	Fresh Water Organisms	Ecology	Comparative study.	NBD truncated at various values provides a better fit.
Crofton (1971)	Fresh Water Organisms	Ecology	Comparative study.	NBD truncated at various values provides a better fit.

### 3 Statistical Models on Count Data

Consider count data as data in which observations take only nonnegative integers and where these integers arise from counting rather than ranking. Conventional approaches in data analyses consider linear regression, which not only assumes normality but also heterogeneity and independence. Even though a normal distribution is continuous, the Gaussian linear regression model is still applicable in analysis of count data; however it is not the best option (Zuur, Leno, Walker, Savelier and Smith, 2007). With some data transformation the lognormal distribution is also an applicable fit to count data. A new response variable ( $Y_{new}$ ), is usually calculated as the original observed count plus one and natural log transformation [ $Ln(Y_{new})$ ] is assumed to be normally distributed. Gaba et al., (2005) fitted the normal distribution and the lognormal distribution to counts of nematodes macroparasites found on sheep. The normal and the lognormal distributions more often provided the worst fits compared to other distributions. O'Hara et al., (2010) cautions against log transformation of count data and argues that generalised linear models (GLMs) are better suited in dealing with count data, especially in the presence of overdispersion. Generalised linear models are an extension of linear models which allows the response variable to follow any member of the exponential family of distributions. In our case two of the distributions of interest are members of the exponential family (Poisson and Negative binomial). GLMs consist of three components;

1. The random component which specifies the conditional distribution of the response variable given a set of explanatory variables,  $Y_i|X_i$ , where  $Y$  is the data vector and  $X$  is the design matrix.
2. The systematic component which specifies the linear function of the explanatory variables (also called the linear predictor) on which the expected value of the response variable,  $\mu_i$  depends on.

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik},$$

where  $\alpha$  and  $\beta_1, \dots, \beta_k$  are unknown regression coefficients.

3. The link function which describes how the expected value of the response variable,  $\mu_i$  is linked to a set of covariates through a linear predictor. A log link function is usually used for counts as it ensures predicted outcomes do not go below zero.

$$g(\mu_i) = \eta_i.$$

The mean can thus be estimated as:

$$\mu_i = e^{\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}}.$$

### 3.1 The exponential family

According to Cox and Hinkley (1979), any discrete or continuous random variable  $Y_i$  that takes on the general form below is classified as a member of the exponential family.

$$f(y_i, \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \quad i = 0, 1, 2, \dots, n,$$

where  $\phi$  is the scale parameter,  $\theta$  is the link parameter and  $a(\cdot)$ ,  $b(\cdot)$ ,  $c(\cdot)$  are functions which determine the distribution when specified and  $n$  is the number of observations. The mean and the variance are given by:

$$\begin{aligned} E(Y) &= b'(\theta) \\ \text{Var}(Y) &= \phi b''(\theta). \end{aligned}$$

The Poisson and the NBD are thus members of the exponential family. Even though the probability mass function (PMF) of the zero-inflated and zero-altered distributions cannot be expressed as exponential family, GLM theory can still be used in applying these distributions. As a result zero inflated distributions will be applied in explaining the distribution of the parasite counts.

### 3.2 The Poisson distribution

We first start by showing that the Poisson distribution is an exponential family member with the mean and the variance of  $\lambda$ . Suppose  $Y \sim \text{Poisson}(\lambda)$ . From appendix A the probability mass function (PMF) of  $Y$  is:

$$\begin{aligned} P(Y = y) &= \frac{e^{-\lambda} \lambda^y}{y!} & y = 0, 1, 2, \dots \\ &= \exp \left[ \log \left( \frac{e^{-\lambda} \lambda^y}{y!} \right) \right] \\ &= \exp[\log e^{-\lambda} + \log \lambda^y - \log y!] \\ &= \exp[-\lambda + y \log \lambda - \log y!]. \end{aligned}$$

Comparing the expression with the general exponential family form it is clear that the Poisson distribution is a member of the exponential family with:

$$\theta = \log \lambda$$

$$b(\theta) = -\lambda$$

$$a(\phi) = 1$$

$$c(y, \phi) = \log y!.$$

From the expression above we can write:

$$\begin{aligned} e^\theta &= e^{\log \lambda} \\ &= \lambda. \end{aligned}$$

For any member of an exponential family we know that  $E(Y) = b'(\theta)$  and  $Var(Y) = b''(\theta)$ .

$$\therefore E(Y) = e^\theta = \lambda$$

and

$$Var(Y) = e^\theta = \lambda.$$

We can see that the mean and the variance of a Poisson distribution are equal and that the canonical link linking  $E(Y)$  and the parameter  $\theta$  is a log link, that is  $g(\mu) = \log \lambda = \theta$ .

## Maximum likelihood estimation

Taking the logarithm of the likelihood function makes it additive. Unknown parameters can then be estimated taking the first-order derivative of the log likelihood function with respect to unknown regression parameters and setting it to zero. Second-order derivatives of the log likelihood function are used in calculating the standard errors of estimates. Unlike linear regression analysis which results in closed form solution for parameter estimates, the Poisson GLM (along with other distributions) results in equation that must be solved iteratively. Using iterative reweighted

least squares such as Fisher scoring or Newton-Raphson unknown parameters can be estimated. Substituting the estimated regression parameters in the mean expression an estimate for  $\mu_i$  can be obtained. For any given  $\mu_i$  the probability of observing a count  $y_i$  can be estimated. The log-likelihood ( $l$ );

$$\begin{aligned} l &= \log(L) \\ &= \sum_{i=1}^n [y_i \log(\mu_i) - \mu_i - \log(y_i!)] \\ &= \sum_{i=1}^n [y_i(x_i'\beta) - e^{x_i'\beta} - \log(y_i!)]. \end{aligned}$$

The partial derivative with respect to  $\beta$ :

$$\begin{aligned} \frac{\delta l}{\delta \beta} &= \sum_{i=1}^n x_i y_i - x_i e^{x_i'\beta} \\ &= \sum_{i=1}^n x_i (y_i - e^{x_i'\beta}) \\ &= \sum_{i=1}^n x_i (y_i - \mu_i). \end{aligned}$$

The estimating equation can then be written as a function of  $\beta$  and is solved iteratively using Newton-Raphson method.

$$\sum_{i=1}^n x_i [y_i - \mu_i(\beta)] = 0.$$

We use the Newton-Raphson method that proceeds iteratively to estimate the parameters, according to Hardin and Hilbe (2012) if  $\beta_{(t)}$  is the starting point and  $t$  takes on any integer, the next value is obtained as:

$$\beta_{(t+1)} = \beta_{(t)} - H_t^{-1} g_{(t)},$$

$g_{(t)}$  is the log-likelihood first partial derivative, evaluated at  $\beta_{(t)}$ .  $H_{(t)}$  is the second partial derivative, also called Hessian (evaluated at  $\beta_{(t)}$ ).

For the Poisson model:

$$g = \sum_{i=1}^n (y_i - \hat{\mu}_i),$$

and

$$H = \sum_{i=1}^n \hat{\mu}_i x_i x_i'.$$

### 3.2.1 Goodness of fit

The deviance is defined as twice the difference between the log likelihood of the model that provides the best fit from the model under study (Zuur, et al., 2007). In generalised linear model theory, the deviance is similar to residual sum of squares in linear models. A deviance decrease constitutes an improvement in model fit, if the model is an exact fit to the data deviance is zero.

$$\begin{aligned} D &= 2[l(y; y) - l(y; \hat{\mu})] \\ &= 2 \sum_{i=1}^n [(y_i \log(y_i/\mu_i) - (y_i - \mu_i))], \end{aligned}$$

where  $D$  is the deviance and  $l$  is the log-likelihood function. A similar measure to  $R^2$  in GLM is called the explained deviance, the portion of the null deviance accounted for by the model. The null deviance is the residual deviance of the intercept only model. Letting  $D_0$  be the null deviance and  $D_1$  be the residual deviance of the model in question, the explained deviance can be calculated as:

$$R^2 \equiv 1 - \frac{D_1}{D_0}.$$

### 3.2.2 Model selection

Model selection is done to ensure important explanatory variables are included in the model. For model selection, the Akaike Information Criterion (AIC) is used. If we have two models  $M_1$  and  $M_2$ , where  $M_2$  is a sub-model of  $M_1$  and  $D_1$  and  $D_2$  are their deviance respectively. The drop1 command in R performs a likelihood ratio test with the difference between the deviances approximated by a Chi-Square distribution.

$$D_2 - D_1 \sim \chi^2(p - q).$$

Where  $p$  and  $q$  are number of parameters for models  $M_1$  and  $M_2$  respectively, with  $q < p$ .  $p$  and  $q$  are the number of parameters in model  $M_1$  and  $M_2$  respectively. The null hypothesis is that the regression parameter  $\beta_i$  (the dropped variable in the sub-model  $M_2$ ) is equal to zero.

### 3.2.3 Overdispersion

When there is evidence that the variance is greater than the mean, then there is overdispersion. In exploring our data informally, the computed variance to mean ratio and index of discrepancy indicated possible overdispersion. Hardin and Hilbe (2008) suggest a regression based test for overdispersion, where under the null hypothesis the variance is equal to the mean. To run the test, first the fitted values and the test statistic,  $Z$ , need to be calculated and then regress the test statistic,  $Z$ , as intercept only model.  $Z$  is standard normal distributed with the mean of 0 and the variance of 1,  $Z \sim (0, 1)$ .

$$Z = \frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i \sqrt{2}}.$$

When dealing with overdispersion in a Poisson regression, a quasi-Poisson GLM can be employed. A quasi-Poisson GLM is a Poisson GLM in all aspect, except that the variance is specified as a linear function of the mean.

$$\text{var}(Y_i) = \phi \mu_i.$$

Adding the dispersion parameter will inflate the standard errors in the effort to adjust for overdispersion. This can be a downside if the model is misspecified, as parameters will become less significant. Similar to Poisson GLM, the `drop1` command in R will be used for model selection. For hypothesis testing however, the new test statistic follows an F distribution (Hilbe and Hardin, 2012).

$$\frac{(D_2 - D_1)/(p - q)}{\hat{\phi}} \sim F(p - q, n - p).$$

### 3.2.4 Model validation

Residual analysis and plots are used in model validation. The variance of the Poisson distribution increases with larger mean values, as a result the standard residual cannot be useful. In GLMs either the Pearson's residuals or the deviance residuals can be employed in model validation (Hilbe, 2014). The Pearson's residuals are just the standardised residuals divided by the square root of the variance of  $Y_i$ .

The standardised residuals are defined as the difference between observed and fitted values,  $y_i - \hat{\mu}_i$ , for  $i = 1, \dots, n$

$$\varepsilon_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\mu_i)}}.$$

Deviance residuals on the other hand can be calculated as;

$$\varepsilon_i^D = \text{sign}(y_i - \mu_i) \sqrt{d_i},$$

where  $d_i$  is each observation's contribution to the deviance. Both the residuals can then be plotted against the fitted values and each explanatory variable to check for any patterns in the plots. If there are any patterns in the plot necessary changes will be implemented, e.g. including some of the explanatory variables that were initially omitted. If the problem persists then a distribution that addresses overdispersion will be fitted.

### 3.3 The negative binomial distribution

We begin by showing the connection between the Poisson and the negative binomial distributions. To achieve this we derive the negative binomial distribution from the first principles. Suppose  $Y_i|\lambda_i$  follows a Poisson distribution with conditional mean  $E(Y_i|\lambda_i) = \mu_i$  and the parameter,  $\lambda_i$ , follows a gamma distribution with the mean  $E(\lambda_i) = \mu_i$  and variance  $\text{var}(\lambda_i) = \mu_i^2 \alpha^{-1}$ .

$$P(Y_i = y_i|\lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad \text{and}$$

$$g(\lambda_i) = \frac{\lambda_i^{\alpha-1} e^{-\lambda_i/\beta}}{\Gamma(\alpha) \beta^\alpha}, \quad \lambda_i > 0.$$

The joint density of  $Y_i|\lambda_i$  is then given by:

$$P(Y_i = y_i|\lambda_i)g(\lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \frac{\lambda_i^{\alpha-1} e^{-\lambda_i/\beta}}{\Gamma(\alpha) \beta^\alpha}.$$



We now show that the marginal distribution of  $Y_i$  follows a negative binomial distribution.

$$\begin{aligned}
P(Y_i = y_i) &= \int_0^\infty P(Y_i = y_i | \lambda_i) g(\lambda_i) d\lambda_i \\
&= \int_0^\infty \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \frac{\lambda_i^{\alpha-1} e^{-\lambda_i/\beta}}{\Gamma(\alpha) \beta^\alpha} d\lambda_i \\
&= \frac{1}{\Gamma(\alpha) \beta^\alpha} \int_0^\infty \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \lambda_i^{\alpha-1} e^{-\lambda_i/\beta} d\lambda_i \\
&= \frac{1}{y_i! \Gamma(\alpha) \beta^\alpha} \int_0^\infty \lambda_i^{\alpha+y_i-1} e^{-\lambda_i(1+1/\beta)} d\lambda_i \\
&= \frac{1}{\Gamma(y_i+1) \Gamma(\alpha) \beta^\alpha} \Gamma(y_i + \alpha) \left( \frac{\beta}{\beta+1} \right)^{y_i+\alpha} \\
&= \binom{y_i + \alpha - 1}{y_i} \left( \frac{1}{\beta+1} \right)^\alpha \left( 1 - \frac{1}{\beta+1} \right)^{y_i}.
\end{aligned}$$

Looking at the expression of the negative binomial distribution in the next paragraph, we can conclude that the marginal distribution of  $Y_i$  is a negative binomial distribution with  $r = \alpha$  and  $p = 1/(\beta+1)$ . The  $E(\lambda_i) = \mu_i = \alpha\beta$  and the  $\text{var}(\lambda_i) = \mu_i^2 \alpha^{-1} = \alpha\beta^2$ . The distribution of  $Y_i$  converges to a Poisson distribution if we let  $\alpha \rightarrow \infty$  while keeping  $\beta = \lambda/\alpha$ , this is because the variance of  $\lambda_i$  goes to 0.

Suppose  $Y \sim NB(r, p)$ . We start by showing that the NB distribution is a member of the natural exponential family. From Appendix A the probability mass function (PMF) of  $Y$  is:

$$\begin{aligned}
P(Y = y) &= \binom{y+r-1}{y} p^r (1-p)^y \quad y = 0, 1, 2, \dots \\
&= \exp \left[ \log \binom{y+r-1}{y} p^r (1-p)^y \right] \\
&= \exp \left[ y \log(1-p) + r \log p + \log \binom{y+r-1}{y} \right].
\end{aligned}$$

$$\begin{aligned}
\text{let } \theta &= \log(1-p) \\
e^\theta &= (1-p) \\
p &= 1 - e^\theta \\
\log p &= \log(1 - e^\theta).
\end{aligned}$$

We substitute  $\log(1 - p)$  and  $\log p$  by  $\theta$  and  $\log(1 - e^\theta)$ :

$$\therefore P(Y = y) = \exp \left\{ \theta y - [-r \log(1 - e^\theta)] + \log \binom{y + r - 1}{y} \right\}.$$

The NB distribution is thus an exponential family member with:

$$\theta = \log(1 - p)$$

$$b(\theta) = -r \log(1 - e^\theta)$$

$$a(\phi) = 1$$

$$c(y, \phi) = \log \binom{y + r - 1}{y}.$$

For any member of an exponential family we know that  $E(Y) = b'(\theta)$  and  $Var(Y) = b''(\theta)$ .

From the  $\theta$  expression we can write:

$$\begin{aligned} e^\theta &= e^{\log(1-p)} \\ &= 1 - p \end{aligned}$$

$$\begin{aligned} E(Y) &= -r \frac{1}{1 - e^\theta} (-e^\theta) \\ &= r \frac{1}{1 - (1 - p)} (1 - p) \\ &= \frac{r(1 - p)}{p}. \end{aligned}$$

From  $E(Y)$  we know that  $b'(\theta) = re^\theta(1 - e^\theta)^{-1}$ .

$$\begin{aligned}
Var(Y) &= b''(\theta) \\
Var(Y) &= (-1)(1 - e^\theta)^{-2}(-e)^\theta re^\theta + re^\theta(1 - e^\theta)^{-1} \\
&= \frac{r(e^\theta)^2}{(1 - e^\theta)^2} + \frac{re^\theta}{1 - e^\theta} \\
&= \frac{r(1 - p)^2}{[1 - (1 - p)]^2} + \frac{r(1 - p)}{1 - (1 - p)} \\
&= \frac{r(1 - p)^2}{p^2} + \frac{r(1 - p)}{p} \\
&= \frac{r(1 - p)^2 + pr(1 - p)}{p^2} \\
&= \frac{r(1 - p)[(1 - p) + p]}{p^2} \\
&= \frac{r(1 - p)}{p^2}.
\end{aligned}$$

The canonical link,  $g(\mu) = \theta$  is thus:

$$\begin{aligned}
\text{let } \Gamma = E(Y) &= \frac{re^\theta}{1 - e^\theta} \\
&= \frac{r}{e^{-\theta} - 1} \\
\Gamma^{-1} &= \frac{e^{-\theta} - 1}{r} \\
r\Gamma^{-1} + 1 &= e^{-\theta} \\
\log e^{-\theta} &= \log(r\Gamma^{-1} + 1) \\
\theta &= \log\left(\frac{1}{r\Gamma^{-1} + 1}\right) \\
&= \log\left(\frac{1}{r/E(Y) + 1}\right) \\
&= \log\left(\frac{E(Y)}{r + E(Y)}\right).
\end{aligned}$$

Thus  $\log\left(\frac{E(Y)}{r+E(Y)}\right)$  is the canonical link of the NBD, which is the function that takes the  $E(Y)$  to  $\theta$ .

A better alternative to a quasi-Poisson in dealing with overdispersion is more often the negative binomial distribution (NBD). The difference between a Poisson and the NBD is that the variance of the NBD is specified as a quadratic function of the mean. Similar to a Poisson regression, to fit a negative binomial GLM we first specify the model in three steps.

If  $Y_i$  Follows a NBD with the parameters  $\mu_i$ , the mean and  $k$ , the inverse aggregation measure.

$$Y_i \sim NB(\mu_i, k).$$

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i + \frac{\mu_i^2}{k}$$

. The systematic part of the model in terms of covariates.

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

. The logarithmic link like in Poisson GLM, which ensures the estimated values are always nonnegative.

$$\log(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

.

Regression parameters can be estimated by first specifying the likelihood function, then take the first and second order derivatives. First we need the PMF for a NBD. Hilbe (2015) expresses the PMF of a NBD as:

$$f(y) = \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left(\frac{k}{\mu_i + k}\right)^k \left(1 - \frac{k}{\mu_i + k}\right)^{y_i}, \quad y_i \geq 0$$

.

From the PMF we derive the log-likelihood function ( $l$ ) which is used to find maximum likelihood estimates.  $k$  is the dispersion parameter and  $\mu_i$  is the mean.

$$\begin{aligned}
L &= \prod_{i=1}^n \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left( \frac{k}{\mu_i + k} \right)^k \left( 1 - \frac{k}{\mu_i + k} \right)^{y_i} \\
l &= \log \left[ \prod_{i=1}^n \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left( \frac{k}{\mu_i + k} \right)^k \left( 1 - \frac{k}{\mu_i + k} \right)^{y_i} \right] \\
&= \sum_{i=1}^n \log \left[ \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left( \frac{k}{\mu_i + k} \right)^k \left( 1 - \frac{k}{\mu_i + k} \right)^{y_i} \right] \\
&= n \log \Gamma(y_i + k) - n \log \Gamma(y_i + 1) - n \log [\Gamma(k)] + nk \log \left( \frac{k}{\mu_i + k} \right) + ny_i \log \left( 1 - \frac{k}{\mu_i + k} \right)^{y_i}.
\end{aligned}$$

Taking the derivative of the log likelihood of the NB distribution does not results in a closed form solution, as a result the Newton-Rhapon method is also used here to obtain parameter estimates.

### 3.4 Excess zeroes

Zero inflation in a count process is when there are far too many zeroes observed than expected by the standard Poisson or Negative binomial distribution. Apart from potentially causing overdispersion, ignoring zero inflation can result in incorrect parameter estimates and also biased standard errors. Zero-inflated models and zero-altered models (hurdle models) are count-response models that can address the issue of zero inflation. Zero-inflated models are mixture models, as the outcomes are modelled as originating from two different (but not separate) statistical processes: a binomial process (indicating exposure or non-exposure to a particular parasite species) and if exposed, a count process (giving rise to either a zero or positive count). Zero-altered models on the other hand are called two parts models; the first part being a binomial distribution determining if the outcome is a zero or nonzero and the second part being a truncated at zero count model. The core difference is in the count process, while the count process of a zero-inflated model can produce zeroes the count process of zero-altered model is zero truncated and as such cannot produce zeroes.

### 3.4.1 Zero inflated models

As stated in the previous section, zero inflated models are mixtures of a binary process and a count process. To fit a zero inflated model, we first need to make an assumption about the distribution of the count process and then get the probability mass function to generate the log likelihood function. If  $\pi_i$  is the probability of a zero outcome from the binary process and  $P(0)$  is the probability of a zero outcome from the count process then:

$$\begin{aligned} P(Y_i = 0) &= \pi_i + (1 - \pi_i)P(0), \\ (Y_i = y_i | y_i > 0) &= (1 - \pi_i)P(y_i). \end{aligned}$$

According to Zuur et al. (2007), if a count process is Poisson distributed, the probability mass function of a zero-inflated Poisson (ZIP) can be written as:

$$\begin{aligned} f(y_i = 0) &= \pi_i + (1 - \pi_i)e^{-\mu_i}, \\ (y_i = y_i | y_i > 0) &= (1 - \pi_i) \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}. \end{aligned}$$

Introducing covariates just like in the Poisson GLM the mean can be modelled as;

$$\mu_i = e^{\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}}.$$

$\pi_i$  can be modeled with an intercept only logistic regression or different sets of covariates. The mean and the variance of a ZIP can be expressed as:

$$\begin{aligned} E(Y_i) &= \mu_i(1 - \pi_i), \\ var(Y_i) &= (1 - \pi_i)(\mu_i + \pi_i + \mu_i^2). \end{aligned}$$

Assuming that the count process is negative binomial distributed, Zuur et al. (2007) express the PMF of a ZINB distribution as:

$$f(y_i = 0) = \pi_i + (1 - \pi_i) \left( \frac{k}{\mu_i + k} \right)^k,$$

$$f(y_i = y_i | y_i > 0) = (1 - \pi_i) \frac{\Gamma(1+k)}{\Gamma(y_i+1)\Gamma(k)} \left( \frac{k}{\mu_i+k} \right)^k \left( 1 - \frac{k}{\mu_i+k} \right)^{y_i}.$$

With the mean and the variance of:

$$E(Y_i) = \mu_i(1 - \pi_i),$$

$$var(Y_i) = (1 - \pi_i) \left( \mu_i + \frac{\mu_i^2}{k} \right) + \mu_i^2(\pi_i^2 + \pi_i).$$

### 3.4.2 Zero-altered models (hurdle models)

Similar to zero-inflated models, zero-altered models were developed to deal with excessive zeroes in count response models (Hardin and Hilbe., 2007). Zero-altered models are separate the data into two groups; the binomial process with a zero inflation probability of  $\pi_i$  and a count process giving rise to only positive counts. Once again an assumption about the distribution of the count process needs to be made. From here a PMF can be obtained and the log likelihood function can be formulated for parameter estimation.

As already explained, the second part of a hurdle model is a truncated at zero count process. A probability distribution truncated at zero is just the very same probability distribution that cannot take the value zero. This can be generally expressed by dividing the probability distribution by one subtract the same probability distribution at the value zero:

$$f(y_i | y_i > 0) = \frac{f(0)}{1 - f(0)}.$$

Assuming that the count process is Poisson distributed, Zuur et al. (2007) expresses the PMF of a ZAP as:

$$f(y_i = 0) = \pi_i,$$

$$f(y_i = y_i | y_i > 0) = (1 - \pi_i) \frac{\mu_i^{y_i} e^{-\mu_i}}{(1 - e^{-\mu_i}) y_i!}.$$

If the count process is assumed to be negative binomial distributed, Zuur et al. (2007) expresses the PMF of a ZANB as:

$$f(y_i = 0) = \pi_i,$$

$$f(y_i = y_i | y_i > 0) = (1 - \pi_i) \frac{\frac{\Gamma(1+k)}{\Gamma(y_i+1)\Gamma(k)} \left(\frac{k}{\mu_i+k}\right)^k \left(1 - \frac{k}{\mu_i+k}\right)^{y_i}}{\left[1 - \left(\frac{k}{\mu_i+k}\right)^k\right]}$$

With all the distributions outlined we now move to the next chapter to compare the performance of each distribution. All two distinctive features of count data will be accounted for (overdispersion and zero inflation). We start by fitting the Poisson distribution, (assuming that egg counts are randomly distributed among their host) we then fit the negative binomial distribution for overdispersion. To account for possible zero inflation we fit ZIP, ZINB, ZAP and ZANB, then conclude by fitting parameter driven time series models to account for serial autocorrelation.



## 4 Fitting the Poisson, NB, ZIP and ZINB distributions

In this section we apply the distributions outlined in Chapter 3 to the faecal egg count data. The purpose in this section is to illustrate the application of both standard and zero inflated count models. Despite fitting the models to all the fifteen parasite species, we only show results for only two of the parasite species (*Cooperia isospora* and *Dictyocaulus filaria*), with the notion that the methodology can be applied to similar datasets. Prior to investigating aggregation / overdispersion and zero inflation patterns of *Cooperia isospora* and *Dictyocaulus filaria* we start with some exploratory analysis.

### 4.1 Preliminary data exploration

As mentioned in the data description section the data is observational and was collected from January 1998 to February 1999 in three different open grazing regions in the Free State province; Kenstell, QwaQwa and Harrismith. Dung from identified animals were examined for parasite species eggs, each animal's dung was examined once. Upon examination, parasite species were identified and the following were recorded: the time of data collection in months, the age, type (whether it was a sheep or goat) and sex of the ruminant. Blood test were also performed and packed cell volume and infection test results were also recorded, optical density and inhibition percentage were then determined. A total of fifteen different parasite species were identified while 495 animals (sheep and goats) were examined. We use the data on a secondary level and for illustrative purposes, the data was originally collected by Mogaswane K.H.R., Mtsali M.S. and Tsotetsi A. from the University of the North, Qwaqwa campus. Prior to application of the specified models to the data we start with some exploratory analysis. This includes looking at the correlation between some key variables, computing key aggregation measures and performing an analysis of variance. Table 3 shows all variable, their types together with their descriptions.

For categorical variables, the number of categories or groups is indicated in the brackets. In total there are 9 possible covariates and 15 identified parasite species each with a sample size of 425 hosts. We start by evaluating the correlation between continuous variables. A weak correlation is observed between packed cell volume and optical density ( $r = 0.007392$ ) and between packed cell volume and percentage

Table 3: Variable description

Variable	Description	Type
FEC	Actual egg count	Discrete
AGE	Age group of the animal	Nominal (2)
ANIMAL	Type of ruminant	Nominal (2)
MONTH	Month of the year	Nominal (12)
PCV	Packed cell volume	Continuous
RESULTS	Whether or not the ruminant tested positive for the actual infection.	Nominal (2)
SEX	Gender	Nominal (2)
SITE	The site where the faecal sample was obtained.	Nominal (3)
OPTDENS	Optical density	Continuous
INHIBIT	Percentage Inhibition	Continuous

inhibition ( $r = -0.00745$ ). Optical density and percentage inhibition have a moderately high negative correlation ( $r = -0.63573$ ). The high correlation is indicative of possible multicollinearity, which we account for in the analysis.

#### 4.1.1 The presence of excessive zeroes in the data

Figure 2 shows the frequency distributions of all fifteen parasite species. For all these frequency distributions, on the horizontal axis we have the faecal egg counts (FEC) and on the vertical axis we have the percentage of zero counts for different species. For all parasite species, most hosts carry a few parasites or no parasites at all while only a few hosts harbour most of the parasites (indicating the possibility of parasites being aggregated among their hosts). For example more than 80% of the hosts recorded zero counts for *B. decoloratus* species (Figure 3) while only 3% of hosts are heavily infected. There is a possibility of zero inflation due to a high proportion of zeroes in the datasets. Except for the *H. hepatica* species (46% zero counts), all parasite species have more than 50% zero counts. A Poisson distribution with the mean of one (will have the highest possible number of zeroes) is expected to

have below 40% zero counts. The dataset clearly has more zeroes than expected by standard count models, indicating the possibility of zero inflated distribution.

#### 4.1.2 Variability of counts

Figure 3 shows boxplots for two parasite species, highlighting the egg count spread across some explanatory variables. Both the boxplot of *C. eimeria* and *O. pinnata* are shown for varying factors (AGE, ANIMAL, SEX and SITE). Due to the high percentage of zeroes in data the boxplots were constructed using median intensity rather than prevalence. This means that in calculating the median for the boxplot, zeroes of uninfected hosts were excluded.

*Coccidia eimeria* within factor comparison indicates that angora goats and merino sheep do not differ much in terms of their median faecal egg counts and that site OBW has the highest median counts compared to site OA and OHS. For *Ostertagia pinnata*, differences are observed in terms of both age and sex median egg counts. Generally the 75th percentile of all boxplots is low (averaging around 3 egg counts), indicating that the majority of the hosts have low egg counts. This is indicative of parasite aggregation among their hosts.

Figure 4 shows boxplots for all parasite species. The boxplots were constructed using the mean prevalence, ignoring uninfected hosts with zero egg counts. Ignoring zero counts, the means and the variation of most species are similar except for; BDE (*Boophilus decoloratus*), REE (*Rhipicephalus evertsi evertsi*), HAE (*Haemonchus contortus*), CO (*Coccidia Isospora*) and EIM (*Coccidia Eimeria*). REE and HAE are shown to have the highest means. In addition HAE has the most varying egg counts, shown in Figure 6 by the length of the boxplot being the longest. HAE is also the most abundant species (host are heavily infected by HAE compared to other species). The top extending whiskers of the boxplot shows that all egg counts are negatively skewed. The mean of these counts is far less than the median as a result a higher frequency of low counts can be expected. This is the case with most parasite counts as indicated in the previous sections. The large outliers shows the skewness of the parasite species distribution, indicating once again that a most of the animal have zero or low egg counts while a few animal have high egg counts.

#### 4.1.3 Characterising aggregation

Aggregation was characterised using two measures of aggregation, the variance to mean ratio and the index of discrepancy. The commonly used measure of aggregation

in literature is the variance to mean ratio, the index of discrepancy is added to validate and check the harmony between the two measures. In calculating the index of discrepancy ( $D$ ), hosts in a sample are ranked from least to most infected and  $D$  is then calculated using the formulae:

$$D = 1 - \frac{2 \sum_{i=1}^N \left( \sum_{j=1}^i x_j \right)}{\bar{x}N(N+1)},$$

where  $N$  is the total number of hosts in a sample and  $x$  is the egg counts on host  $j$ . The variance to mean ratio was simply calculated by taking the ratio between the sample variance and the sample mean. Table 4 shows the two calculated measures of aggregation for all parasite species.

Table 4 shows all parasite species to have a variance to mean ratio considerably greater than one. *T. Gondii* has the highest variance to mean ratio of 5.5465 and while *O. Columbianum* has the lowest variance to mean ratio of 2.3104. This indicates that the variance is far greater than the mean and violates the Poisson assumption of equal mean and variance. As a result potential overdispersion will be factored in when formulating distributional assumptions of the parasite species. Both aggregation measures show all fifteen parasites are aggregated among their hosts. The index of discrepancy shows values close to one while the variance to mean ratio shows values greater than two. We investigate the nature of this aggregation given the covariates (AGE, ANIMAL, MONTH, SEX and SITE where the livestock were kept), by looking for any patterns of seasonality and difference in aggregation given each covariate. Besides the negative binomial distribution we demonstrate other distributions that provide a good fit to over dispersed count data.

Table 4: Parasite species index of discrepancy and variance to mean ratio

Parasite Species	Sample Mean	Sample Variance	Index of discrepancy	Variance to mean ratio
<i>B. Decoloratus</i>	0.3576	0.9677	0.8843	2.7062
<i>C. Curticei</i>	0.3576	0.9302	0.8875	2.6013
<i>C. Eimeria</i>	0.8067	2.0938	0.7806	2.5957
<i>C. Isospora</i>	1.3867	3.8960	0.684	2.8096
<i>D. Filaria</i>	0.5717	1.5120	0.8374	2.6447
<i>F. Hepatica</i>	0.5863	1.4514	0.8209	2.4756
<i>H. Contortus</i>	1.9875	7.1415	0.6413	3.5932
<i>O. Columbianum</i>	0.4761	1.0999	0.5801	2.3104
<i>O. Pinnata</i>	0.5821	1.5938	0.8346	2.7379
<i>P. Cervi</i>	0.5156	1.3128	0.8419	2.5462
<i>R. Evertsi</i>	1.2495	3.5460	0.7016	2.8380
<i>S. Papillosus</i>	0.7318	2.0925	0.8044	2.8594
<i>T. Axei</i>	0.7422	2.0376	0.8	2.7453
<i>T. Gondii</i>	0.2287	1.2684	0.9365	5.5465
<i>T. Ovis</i>	0.4969	1.7922	0.8864	3.6069

#### 4.1.4 Nonparametric tests for differences between factors

In addition to the preliminary data exploration, non parametric tests are conducted to test for differences between factor medians. Parametric tests like the ANOVA assume normality, non-parametric test are based on fewer assumption as they do not assume normality. For this reason we use the Kruskal-Wallis test, which test for difference in medians across multiples groups. The Kruskal-Wallis test statistic is

denoted as,  $H$ , and is calculated as;

$$H = \left( \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} \right) - 3(N+1),$$

where  $N$  is the total sample size,  $k$  is the number levels in a factor,  $n_j$  is the sample size in the  $j^{th}$  factor and  $R_j$  is the sum of ranks in the  $j^{th}$  factor.  $H$  is then compared with the  $\chi^2$  critical value to make a conclusion about the median comparison. Factors include AGE, ANIMAL, MONTH, SEX, SITE and SPECIES. First boxplots of egg counts of all 15 parasite species is presented, showing how the mean and spread of each species. Results from the Kruskal-Wallis test are then provided, giving an idea of how the median egg counts differ across factors.

The Kruskal-Wallis test results are presented in Table 5. Factors AGE, ANIMAL, MONTH, SEX and SITE are given with their degrees of freedom in brackets together with their  $\chi^2$  critical value and the corresponding p-value. Values that bolded indicate significance at a 5% level. We also tested the overall medial differences across all fifteen parasite species (results shown at the bottom of Table 5). With a p-value  $< 0.0001$  there is a high significant difference across the medians of all parasite species. ANIMAL and MONTH are key factors with significant difference across eleven out of the fifteen parasite species, followed by SITE (significantly different for ten parasite species) the AGE (significantly different for 7 parasite species). SEX is significantly different for the least number of parasite species.

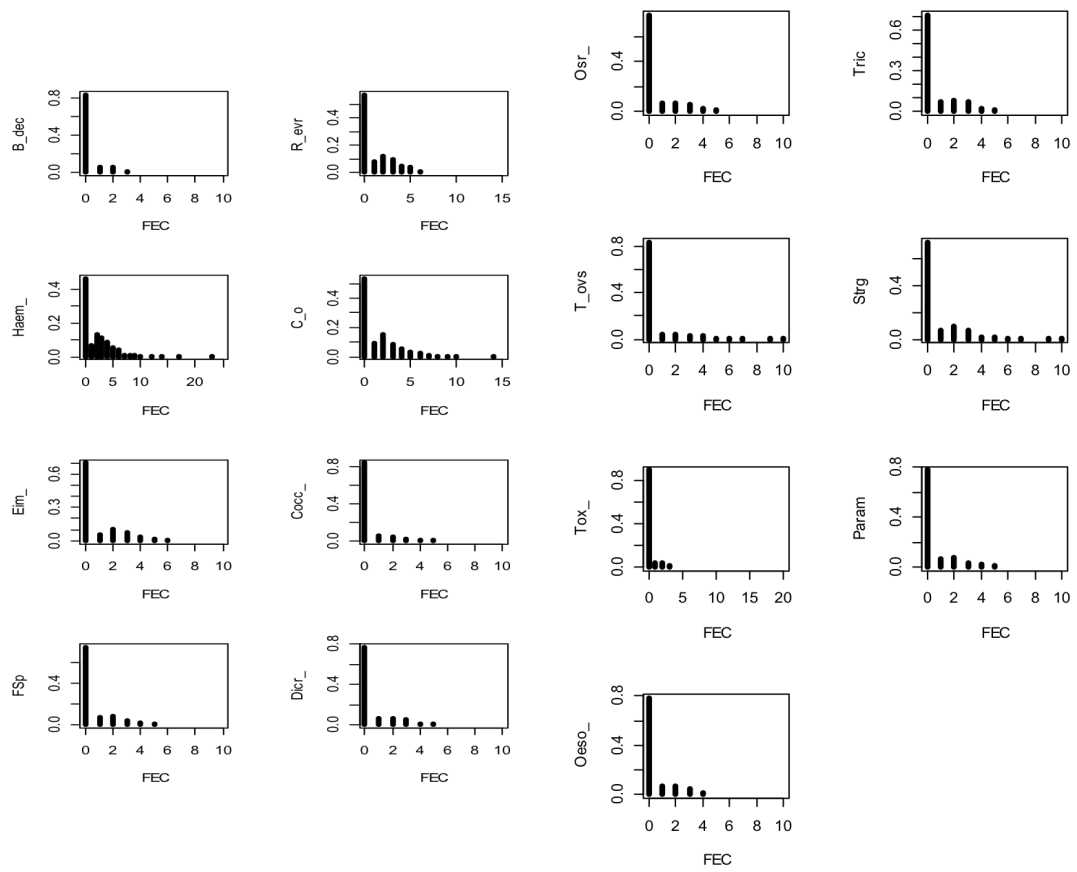


Figure 2: Parasite species frequency distribution

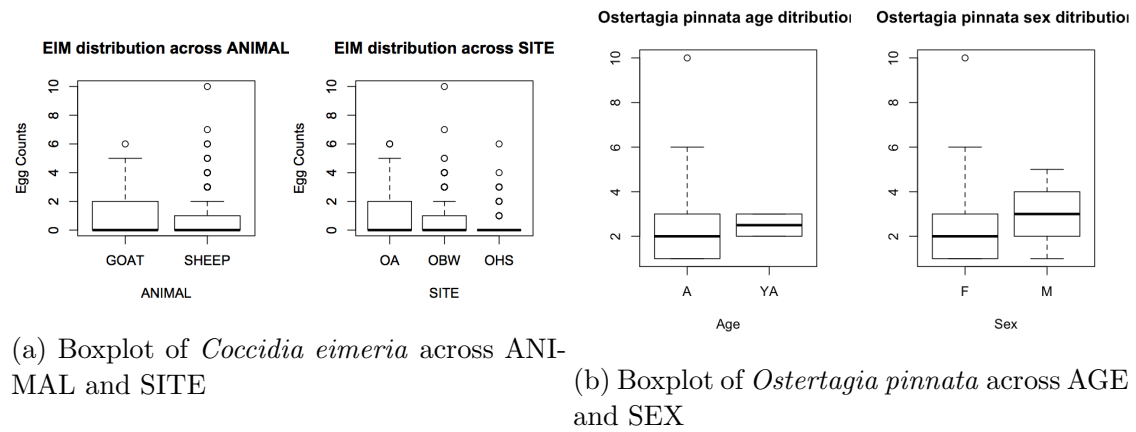


Figure 3: *Coccidia eimeria* and *Ostertagia pinnata* boxplots

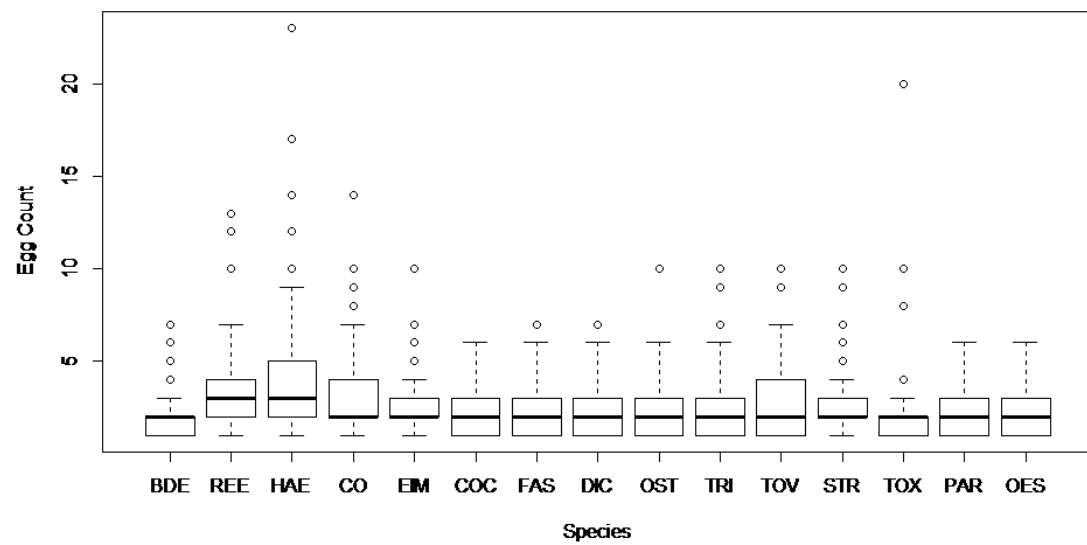


Figure 4: Boxplots of all parasite species



Table 5: Kruskal-Wallis Test

	AGE (1)		ANIMAL (1)		MONTH (11)		SEX (1)		SITE (2)	
Parasite Species	$\chi^2$	p-value	$\chi^2$	p-value	$\chi^2$	p-value	$\chi^2$	p-value	$\chi^2$	p-value
<i>B. Decoloratus</i>	0.72	0.3955	0.43	0.5108	<b>72.35</b>	< 0.0001	<b>6.04</b>	<b>0.0140</b>	<b>16.13</b>	<b>0.0003</b>
<i>C. Curticei</i>	0.92	0.3370	0.03	0.8641	<b>32.97</b>	<b>0.0005</b>	0.08	0.7815	<b>18.43</b>	<b>0.0001</b>
<i>C. Isospora</i>	3.07	0.0797	<b>5.74</b>	<b>0.0166</b>	<b>37.97</b>	<b>0.0001</b>	3.49	0.0617	3.82	0.1479
<i>D. Filaria</i>	<b>4.75</b>	<b>0.0294</b>	<b>115.71</b>	< 0.0001	<b>23.68</b>	<b>0.0142</b>	1.33	0.2493	<b>18.03</b>	<b>0.0001</b>
<i>C. Eimeria</i>	<b>4.28</b>	<b>0.0386</b>	<b>9.03</b>	<b>0.0027</b>	<b>26.26</b>	<b>0.0059</b>	3.68	0.0549	<b>8.33</b>	<b>0.0156</b>
<i>F. Hepatica</i>	0.63	0.4281	<b>14.21</b>	<b>0.0002</b>	14.41	0.2111	1.53	0.2164	1.84	0.3989
<i>H. Contortus</i>	<b>16.85</b>	< 0.0001	<b>54.92</b>	< 0.0001	<b>42.63</b>	< 0.0001	1.57	0.2096	<b>32.08</b>	< 0.0001
<i>O. Columbianum</i>	<b>7.96</b>	<b>0.0048</b>	<b>138.63</b>	< 0.0001	8.90	0.6313	1.35	0.2450	2.41	0.3000
<i>O. Pinnata</i>	<b>11.08</b>	<b>0.0009</b>	<b>44.12</b>	< 0.0001	<b>25.40</b>	<b>0.0080</b>	1.60	0.2061	<b>12.45</b>	<b>0.0020</b>
<i>P. Cervi</i>	1.73	0.1883	<b>68.72</b>	< 0.0001	19.36	0.0549	<b>5.96</b>	<b>0.0146</b>	<b>20.61</b>	< 0.0001
<i>R. Evertsi</i>	0.51	0.4738	<b>4.74</b>	<b>0.0296</b>	<b>62.65</b>	< 0.0001	0.20	0.6555	<b>17.32</b>	<b>0.0002</b>
<i>S. Papillosus</i>	1.04	0.3079	<b>93.54</b>	< 0.0001	<b>42.83</b>	< 0.0001	<b>10.27</b>	<b>0.0014</b>	<b>12.88</b>	<b>0.0016</b>
<i>T. Ovis</i>	<b>6.62</b>	<b>0.0101</b>	2.17	0.1404	<b>66.28</b>	< 0.0001	1.50	0.2213	0.90	0.6384
<i>T. Gondii</i>	0.86	0.3530	0.01	0.9857	19.53	0.0522	2.28	0.1309	1.16	0.5597
<i>T. Axi</i>	<b>9.62</b>	<b>0.0019</b>	<b>86.08</b>	< 0.0001	<b>33.12</b>	<b>0.0005</b>	<b>6.34</b>	<b>0.0118</b>	<b>33.86</b>	< 0.0001

$\chi^2_{df=14} = 574.41$  for SPECIES with a p-value < 0.0001.

## 4.2 Analysis of *Cooperia isospora* egg counts

In this section we investigate abundance patterns of *Cooperia isospora* egg counts on both sheep and goats.

### 4.2.1 The Poisson model

Despite the high proportion of zeroes in the data, standard count models were able to fit some data well Sileshi et al (2007). For this reason, first standard count models were fitted to the data despite the presence of high proportion of zeroes. We applied model selection procedure on all the 15 datasets. However, due to the large number only one dataset (*Cooperia isospora*) was chosen to illustrate the employed model selection technique. For nested models, model selection was done using either the  $z$  statistic (testing whether or not individual parameter estimates were significantly different from zero) or the deviance test (using the Chi-square statistic). For the Poisson model, the Chi-square deviance test was applied using the R drop1 command. The R codes implementing the fitted models are provided in the appendix. Table 6 indicates the Poisson model selection procedure employed. Both the AIC and the p-value are provided at each step a covariate is dropped. The numbers 1, 2, ..., 5 indicate sequential steps in model selection.

In every step the insignificant covariate to be deleted is indicated in bold. The final model after all insignificant covariates are removed and/or included back in the model with reason is depicted in Step 5. Table 7 shows parameter estimates and their standard errors for *C. isospora*. Poisson GLM. For all the categorical variables, one level within a factor is used as a reference. For RESULTS, the level RESULTSNEG is used as a reference, for SITE, the level SITEOA is used as a reference while for MONTH, the level MONTHJAN is used as a reference. All variable are significant at a 10% level of significance.

Despite OPTDENS and INHIBIT being highly correlated, both are significant in the Poisson model. The determinants OPTDENS and INHIBIT indicate that optical density and percentage inhibition both have a negative relationship with (*Cooperia isospora*) egg counts. The covariate RESULTS shows that animals that tested positive for an infection have lower egg counts than those that tested negative, indicating that *C. isospora* adults have poor egg laying capacity and could cause severe clinical symptoms before a large number of eggs are present in the faeces. Looking at SITE, OHS has lower *C. isospora* egg counts followed by OBW and then site OA. The variable MONTH indicates that egg counts expected to be lowest in Novem-

ber while highest egg counts are expected in February. Compared to January, *C. isospora* egg counts expected to be 185% [ $\exp(1.0482)=2.85$ ] higher in February and 36% [ $\exp(-0.4437)=0.64$ ] lower in November.

Table 6: Poisson model selection

DETERMINANTS	1		2		3		4		5	
	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value
NONE	1775		1773		1772		1772		1772	
AGE	1773	0.411	<b>1772</b>	<b>0.4576</b>						
ANIMAL	1776	0.0584	1774	0.0702	1773	0.0578	<b>1772</b>	<b>0.1100</b>		
MONTH	1870	<0.0001	1869	<0.0001	1869	<0.0001	1781	<0.0001	1875	<0.0001
PCV	<b>1773</b>	<b>0.4688</b>								
RESULTS	1778	0.0183	1777	0.0182	1775	0.0203	1775	0.0181	1774	0.0412
SEX	1775	0.1442	1773	0.1472	<b>1772</b>	<b>0.1437</b>				
SITE	1779	0.0127	1778	0.0114	1777	0.0108	1776	0.0158	1780	0.0022
OPTDENS	1777	0.0371	1776	0.0343	1774	0.0369	1774	0.0340	1775	0.0312
INHIBIT	1784	0.0009	1782	0.0008	1781	0.0008	1781	0.0008	1783	0.0004

Table 7: Poisson model parameter estimates

DETERMINANTS	Estimates	Std. Error	p-value
INTERCEPT	1.1374	0.2006	<0.0001
MONTH02FEB	1.0482	0.1699	<0.0001
MONTH03MAR	0.4920	0.1796	0.0062
MONTH04APR	0.1873	0.1872	0.3170
MONTH05MAY	0.3589	0.1856	0.0531
MONTH06JUN	0.1129	0.1942	0.5610
MONTH07JUL	0.4314	0.1783	0.0155
MONTH08AUG	-0.2996	0.2203	0.1738
MONTH09SEP	-0.3003	0.2099	0.1525
MONTH10OCT	-0.3725	0.2511	0.1379
MONTH11NOV	-0.4437	0.2649	0.0939
MONTH12DEC	0.6192	0.1912	0.0012
RESULTSPoS	-0.3140	0.1534	0.0406
SITEOBW	-0.1796	0.0983	0.0677
SITEOHS	-0.3744	0.1107	0.0007
OPTDENS	-0.1515	0.0710	0.0327
INHIBIT	-0.0082	0.0023	0.0004

Parameter estimates from the fitted model in Table 7 were used to calculate the fitted values. Figure 8 shows a plot of observed against fitted values for all distributions namely the Poisson, negative binomial (NB) zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB).

With focus on the Poisson model, there are a large number of values that have been misfit (especially the zeroes) which have been severely under fitted. Misfit by the Poisson model prompts some diagnostic plot. As a result, we perform model validation prior to completely ruling out the Poisson model as an appropriate fit for the distribution of *C. isospora*. Model validation was done by plotting the residual

against each of the explanatory variables as well as against the fitted values. Figure 5 represents the residual plots for the Poisson model. The residual plots against both OPTDENS and INHIBIT show no clear pattern, however that of the fitted values shows a clear pattern, indicating that the residual are somewhat dependent on the fitted values. The presence of systematic feature in residual plot indicates violation of one or more of the assumptions. For the Poisson model we take a closer look at the assumption of equidispersion.

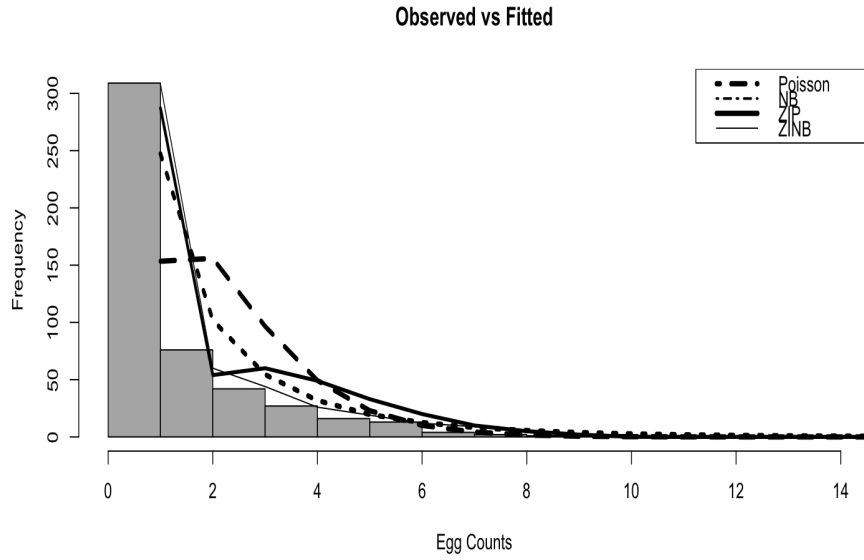


Figure 5: Observed and fitted values for *C. isospora*

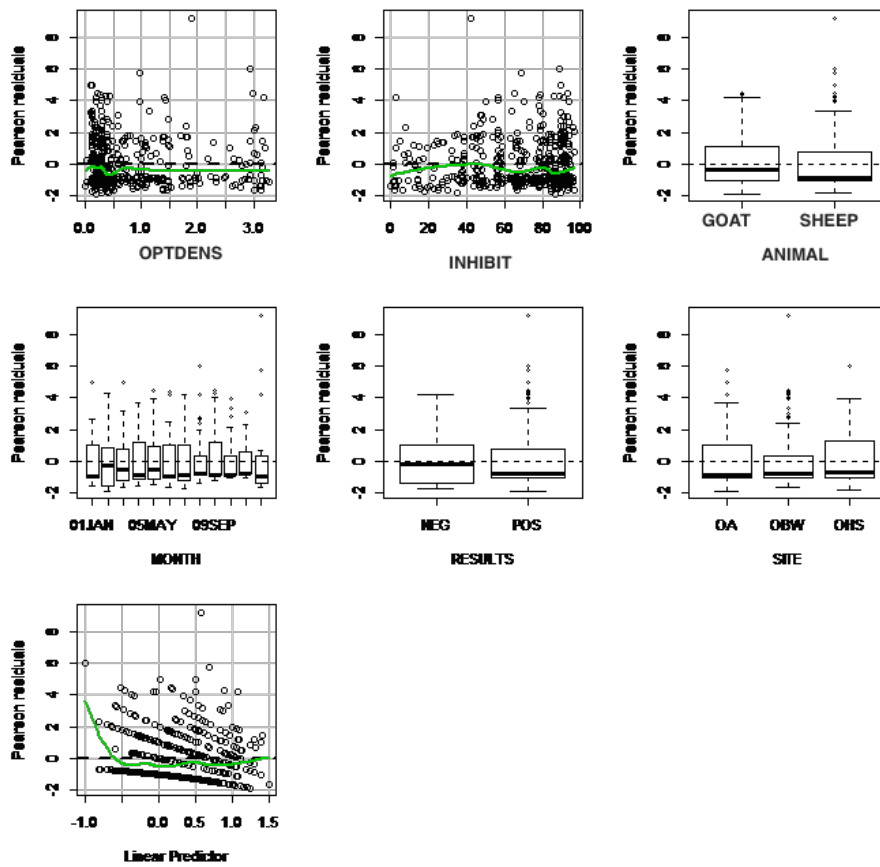


Figure 6: Poisson residual plot

#### 4.2.2 The overdispersion test

To test the Poisson assumption of equidispersion, a dispersion test was run in R software. Table 8 shows results from the R output (overdispersion test) for fifteen of the parasite species.

Table 8: Overdispersion test

Parasite species	DF	$z$	p-value
<i>B. decoloratus</i>	15	7.56	<0.0001
<i>C. curticei</i>	13	5.55	<0.0001
<i>C. eimeria</i>	15	11.67	<0.0001
<i>C. isospora</i>	13	12.64	<0.0001
<i>D. filaria</i>	13	4.99	<0.0001
<i>F. hepatica</i>	16	5.88	<0.0001
<i>H. contortus</i>	13	7.13	<0.0001
<i>O. columbianum</i>	15	6.16	<0.0001
<i>O. pinnata</i>	15	5.26	<0.0001
<i>P. cervi</i>	16	8.21	<0.0001
<i>R.e. evertsi</i>	16	6.55	<0.0001
<i>S. papillosus</i>	15	7.03	<0.0001
<i>T. axei</i>	16	7.92	<0.0001
<i>T. gondii</i>	16	10.44	<0.0001
<i>T. ovis</i>	13	8.54	<0.0001

Looking at *C. isospora* in particular, the null hypothesis that the variance is equal to the mean is rejected because of the small p-value of  $< 0.0001$ . Looking at all other cases the dispersion parameters are highly significant, all with p-values were very close to zero ( $p < 0.0001$ ), we reject the null hypothesis that the variance is equal to the mean and conclude that there is substantial amount of overdispersion in the data. Overdispersion in one is an indication of the inadequacy of the Poisson models, prompting the fitting of a negative binomial models which accounts for overdispersion.



### 4.2.3 The negative binomial model

Table 9 shows the NB model selection procedure employed. Comparing the model selection procedure of the Poisson model to that of the NB model, there are more significant covariates in the Poisson model compared to the NB model. Due to overdispersion not accounted for in the Poisson model, standard errors tend to be large, resulting in large confidence intervals of parameter estimates. This in turn causes significance of covariates, which would otherwise not have been significant if overdispersion was accounted for. 1, 2, ..., 7 indicate sequential model selection steps.

The NB model selection in Table 9 is obtained from the `drop1` command in R. Step 1 in Table 10 shows that PCV is the most insignificant covariate with the highest p-value of 0.8393. The null hypothesis that  $\beta_{PCV} = 0$  is first tested against the alternative hypothesis that  $\beta_{PCV} \neq 0$ . The difference between the deviance of the two models ( $D_2 - D_1 = 0.685$ ) is approximately chi-square distributed with one degree of freedom (single term deleted) giving a p-value of 0.8393. Comparing this p-value with a 10% level of significance, the null hypothesis is not rejected, concluding that the parameter estimate  $\beta_{PCV}$  is not significantly different from zero. The covariate packed cell volume is then dropped from the model and the `drop1` command is run again, taking us to Step 2. From Step 2 the covariate in bold is removed from the model, stepwise deletion of insignificant variable continues all the way through to the final model in Step 7. The negative binomial final model has less significant terms than the Poisson counterpart.

Table 9: Negative binomial model selection

	1		2		3		4		5		6		7	
DETERMINANTS	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value	AIC	p-value
NONE	1547		1545		1544		1543		1542		1542		1543	
AGE	1546	0.287	1545	0.2945	1543	0.3073								
ANIMAL	1551	0.0145	1549	0.0146	1548	0.0136	1548	0.0087	1546	0.0142	1546	0.0147	1547	0.0124
MONTH	1568	<0.0001	1566	<0.0001	1567	<0.0001	1566	<0.0001	1567	<0.0001	1566	<0.0001	1568	<0.0001
PCV	1545	0.8393												
RESULTS	1549	0.0696	1547	0.0709	1545	0.0906	1544	0.1005	1543	0.1023	1544	0.0494	1557	0.0001
SEX	1547	0.2802	1545	0.2768	1543	0.2682	1542	0.2759						
SITE	1545	0.3698	1543	0.3682	1544	0.1409	1543	0.1421	1542	0.1621				
OPTDENS	1546	0.4079	1544	0.4004										
INHIBIT	1548	0.1174	1546	0.1136	1544	0.1782	1543	0.1738	1542	0.1767	1541	0.2801		

Table 10 shows the final NB model with parameter estimates and standard errors. For all the categorical variables, one level within a factor is used as a reference. For ANIMAL, the level ANIMALGOAT is used as a reference, for RESULTS, the level RESULTSNEG is used as a reference, for SITE, the level SITEOA is used as a reference while for MONTH, the level MONTHJAN is used as a reference.

Table 10: Negative binomial model parameter estimates

DETERMINANTS	Estimates	Std. Error	p-value
INTERCEPT	1.0052	0.3166	0.0015
ANIMALSHEEP	0.0039	0.1572	0.0131
MONTH02FEB	1.2031	0.3153	0.0001
MONTH03MAR	0.6509	0.3159	0.0393
MONTH04APR	0.2295	0.3241	0.4789
MONTH05MAY	0.4133	0.3232	0.2011
MONTH06JUN	0.2350	0.2350	0.4636
MONTH07JUL	0.4874	0.3108	0.1168
MONTH08AUG	-0.2411	0.3418	0.4806
MONTH09SEP	-0.1958	0.3264	0.5485
MONTH10OCT	-0.2830	0.3773	0.4533
MONTH11NOV	-0.4585	0.3918	0.2420
MONTH12DEC	0.6468	0.3444	0.0604
RESULTSPOS	-0.5394	0.2813	0.0552
THETA	1.6531	0.0846	0.0007

ANIMAL indicates that sheep are expected to have 4% [ $\exp(0.0039)=1.04$ ] more *C. isospora* egg counts compared to goats. As in the Poisson model the predictor RESULTS indicate low egg lying capacity of adults parasite. MONTH indicates that lowest *C. isospora* egg counts are expected in November while highest egg counts are expected in February. Compared to January, *C. isospora* egg counts expected to be 233% [ $\exp(1.2031)=3.33$ ] higher in February and 37% [ $\exp(-0.4585)=0.63$ ] lower in November. We return to figure 8 which shows observed and fitted values for the

Poisson, NB, ZIP and ZINB models. Despite the NB model providing a better fit than the Poisson model, zero counts are still largely under fitted for the NB model. Similar to the Poisson case, we plot the fitted values for the NB model against the residuals. Figure 7 shows the negative binomial model residual plot. The residual plot against predictors somewhat shows dependence while those against the fitted values are not as severe as those in the Poisson model. In addition the zeroes in the NB binomial model were under fitted, rendering zero inflated models as possible alternatives.

The residual plot against predictors somewhat shows dependence while those against the fitted values are not as severe as those in the Poisson model. In addition the zeroes in the NB binomial model were under fitted, rendering zero inflated models as possible alternatives.

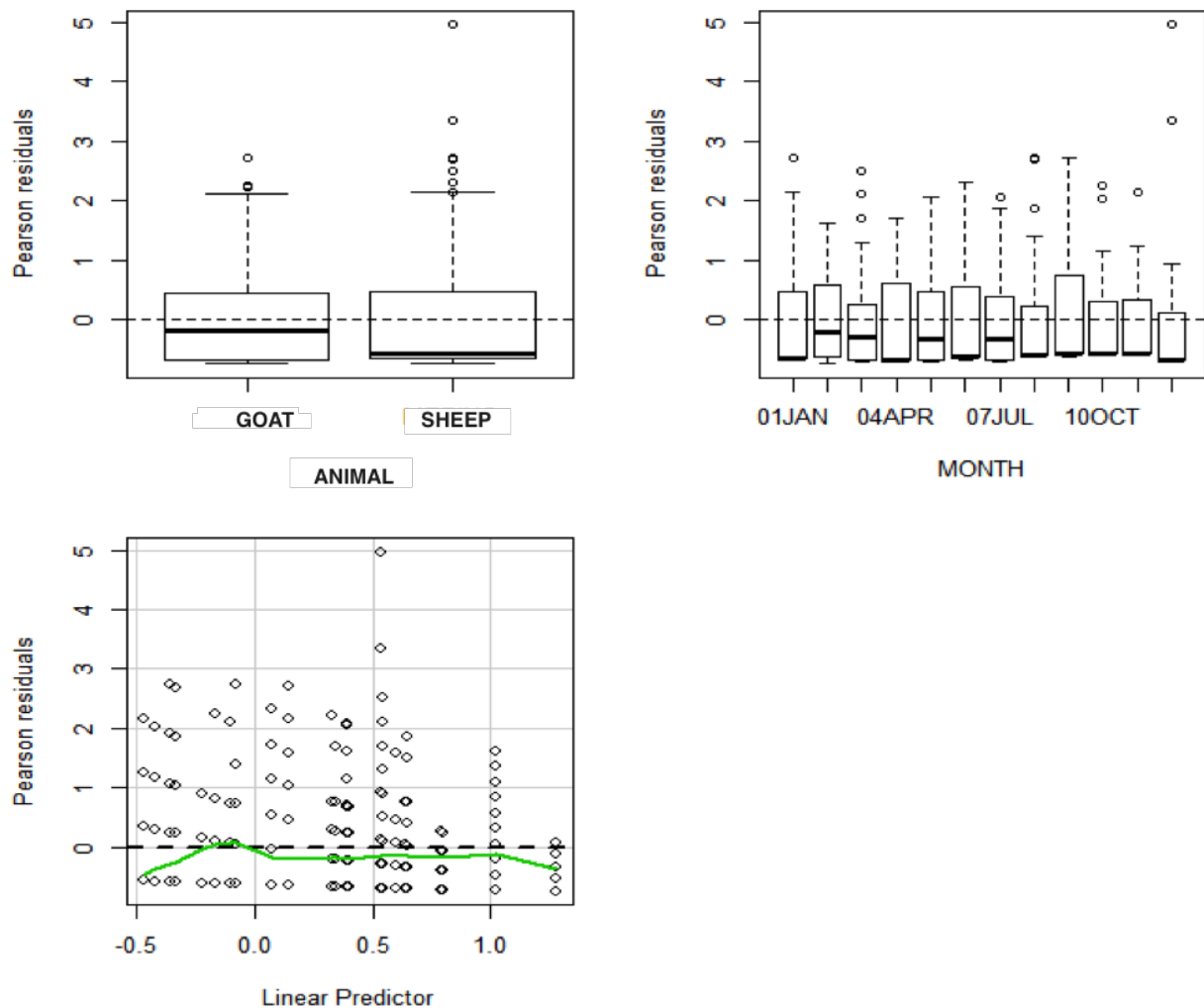


Figure 7: Negative binomial residual plot

#### 4.2.4 Zero inflated models

Zero inflated and zero altered models provided quite similar results; for this reason, results from only the best models (either zero inflated or zero altered) that explained the distribution of *C. isospora* is presented. *C. isospora* was best explained by the ZINB distribution, subsequently only results from zero inflated models are presented

in this section. Despite overdispersion being accounted for by the negative binomial model, zero inflation was still an apparent problem (zeroes in the NB model were under fitted and residual plots had some degree of systematic patterns). We addressed this problem by fitting both zero inflated and zero altered models.

#### 4.2.5 Zero inflated Poisson model (ZIP)

Table 11 shows the zero inflated Poisson full model. The first Step of model selection had OPTDENS removed first from the model. For the nested zero inflated models, the  $z$  statistic was used in model selection.

Table 11: ZIP model selection

DETERMINANTS	Poisson Model		Binomial model	
	Estimate	p-value	Estimate	p-value
INTERCEPT	1.5493	0.0224	-0.9281	0.6069
AGEYA	0.2780	0.0596	0.7960	0.0318
ANIMALSHEEP	0.5202	0.0001	1.8477	<0.0001
MONTH	0.1524	0.4428	-2.4538	0.0006
PCV	-2.1390	0.3466	-7.4502	0.2235
RESULTSPOS	0.3366	0.0558	1.6925	0.0028
SEXM	-0.1644	0.1121	-0.6272	0.0463
SITEOBW	-0.0704	0.5829	0.1670	0.6074
SITEOHS	-0.6805	<0.0001	-0.9535	0.0231
OPTDENS	0.0167	0.8586	0.4888	0.0686
INHIBIT	-0.0052	0.0742	0.0105	0.2320

First the null hypothesis that  $\beta_{OPTDENS} = 0$  was first tested against the alternative hypothesis that  $\beta_{OPTDENS} \neq 0$ . Comparing the p-value of 0.8586 to the 10% level of significance, the null hypothesis is rejected and we conclude that the parameter estimate  $\beta_{OPTDENS}$  is not significantly different from zero. Similar to the NB model and due to its high correlation with percentage inhibition, optical density is first

removed from the count part of the ZIP model. Insignificant predictors are then removed from either the count or the binomial part of the model stepwise until getting to the final model in Table 12 (ZIP Model Parameter Estimates). Table 12 shows the ZIP final model. The ZIP model allows us to test if the process is indeed zero inflated.

Table 12: ZIP model parameter estimates

Count Model	Estimates	Std. Error	p-value
INTERCEPT	1.0597	0.1889	<0.0001
ANIMALSHEEP	0.3717	0.1130	0.0010
MONTH02FEB	0.2858	0.1854	0.1232
MONTH03MAR	-0.0092	0.2038	0.9639
MONTH04APR	-0.0386	0.2071	0.8524
MONTH05MAY	0.0258	0.2012	0.8980
MONTH06JUN	-0.3161	0.2285	0.1667
MONTH07JUL	0.0985	0.1952	0.6141
MONTH08AUG	-0.0680	0.2515	0.7868
MONTH09SEP	-0.0471	0.2367	0.8423
MONTH10OCT	-0.4664	0.3073	0.1291
MONTH11NOV	-0.7753	0.3262	0.0175
MONTH12DEC	0.2599	0.2045	0.2038
SITEOBW	-0.0897	0.0986	0.3629
SITEOHS	-0.4239	0.1255	0.0007
INHIBIT	-0.0032	0.0017	0.0565
Binomial Model	Estimates	Std. Error	p-value
INTERCEPT	0.0461	0.5206	0.3827
ANIMALSHEEP	1.5276	0.3457	<0.0001
MONTH02FEB	-2.1314	0.6138	0.0005
MONTH03MAR	-1.1118	0.5461	0.0418
MONTH04APR	-0.3413	0.5121	0.5051
MONTH05MAY	-0.8156	0.5093	0.1093
MONTH06JUN	-1.1022	0.5489	0.0446
MONTH07JUL	-0.8967	0.4906	0.0676
MONTH08AUG	0.1369	0.5266	0.7949
MONTH09SEP	-0.0010	0.4834	0.9984
MONTH10OCT	-0.5372	0.6093	0.3780

*Continued on next page*

Table 12 – *Continued from previous page*

Count Model	Estimates	Std. Error	p-value
MONTH11NOV	-0.7375	0.7029	0.2941
MONTH12DEC	-0.9621	0.5494	0.0799
RESULTSPOS	1.1643	0.3628	0.0013
SEXM	-0.4905	0.2714	0.0707

From the zero inflated part of the model (binomial model), we conclude that the zero inflated Poisson model is preferred to the Poisson model and that the process is indeed zero inflated. This is because zero inflation intercept is positive and not statistically significant (intercept = 0.0461 with a p-value= 0.3827). For all the categorical variables, one level within a factor is used as a reference. For ANIMAL, the level ANIMALGOAT is used as a reference, for RESULTS, the level RESULT-SNEG is used as a reference, for SITE, the level SITEOA is used as a reference while for MONTH, the level MONTHJAN is used as a reference. Higher percentage of inhibition results in lower the egg counts while sheep generally have higher *C. isospora* egg counts than goats. Sheep are expected to have 62% [ $\exp(0.4837)=1.62$ ] *C. isospora* egg counts compared to goats. Much lower egg counts are expected from hosts located in site OHS followed by site OBW with highest egg counts expected in site OA. Monthly variation is present with January, December and February having the highest egg counts respectively while lowest counts are observed in the months of November and October, respectively. From the binomial part of the model we conclude that even though age, results and gender do not have an influence on the actual egg count, they do however have an influence on the absence or presence of egg counts in host faecal sample. Adults seem to have more faecal eggs present compared to their yearlings counterparts. Males have fewer eggs present compared to females in their faecal samples.

From the estimated parameter values, fitted values for the ZIP model are then calculated. From Figure 8 the ZIP model has a better fit to zero counts than the standard Poisson and negative binomial. Just like with the previous two models, a residual plot of quantitative predictor(s) included in the model is shown in Figure 11 together with that of the fitted values.

From both plots, no clear systematic pattern in the residuals is observed, suggesting the ZIP model is a better fit to the distribution of *C. isospora* faecal egg counts.



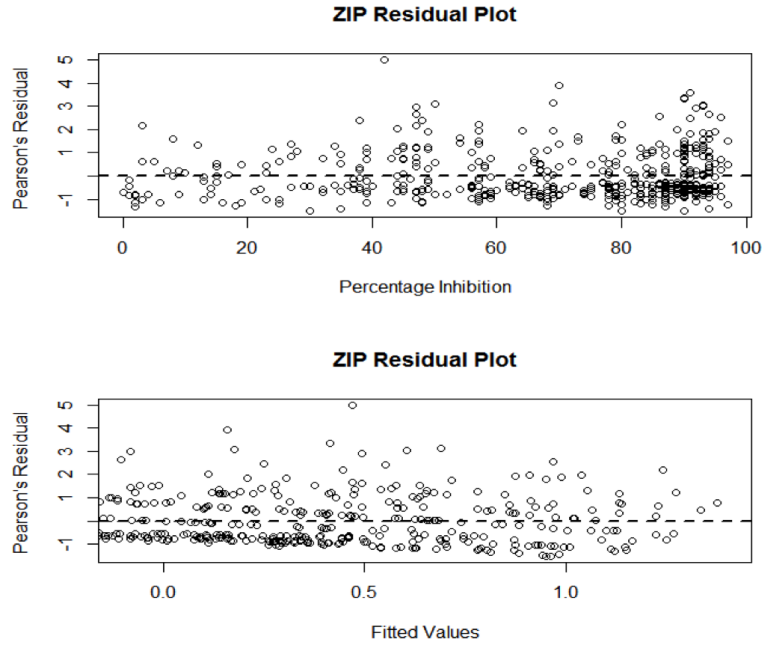


Figure 8: ZIP model residual plot

#### 4.2.6 Zero inflated negative binomial (ZINB) model

Table 13 shows the ZINB full model and only the first Step of model selection, with the variable in bold being removed first from the model. Similar to the ZIP model, optical density is the first covariate to be removed from the full model (indicated in bold in Table 13). Variables are removed stepwise from both the count and binomial part of the model until the significant covariates are left in the model.

Table 14 shows the ZINB final model after model section (ZINB parameter estimates). For all the categorical variables one group is used as a reference. For results, the negative group is the reference and for site, site OA is the reference.

Table 13: ZINB model selection

DETERMINANTS	Poisson Model		Binomial model	
	Estimate	p-value	Estimate	p-value
INTERCEPT	1.7801	0.0379	-0.5066	0.8390
AGEYA	0.3797	0.0572	1.2057	0.0762
ANIMALSHEEP	0.6416	<0.0001	2.9873	0.0155
MONTH	0.1279	0.6026	-3.4047	0.0015
PCV	-2.7743	0.3307	-11.0415	0.2277
RESULTSPOS	0.3573	0.0944	2.1735	0.0131
SEXM	-0.1689	0.1751	-0.8988	0.0516
SITEOBW	0.0015	0.9916	0.4408	0.3227
SITEOHS	-0.7959	<0.0001	-1.6080	0.0612
<b>OPTDENS</b>	<b>-0.0786</b>	<b>0.5121</b>	<b>0.2930</b>	<b>0.4304</b>
INHIBIT	-0.0078	0.0279	0.0031	0.8107

Table 14: ZINB model parameter estimates

Count Model	Estimates	Std. Error	p-value
INTERCEPT	0.9553	0.2345	<0.0001
ANIMALSHEEP	0.4837	0.1386	0.0005
MONTH02FEB	0.2471	0.2232	0.2681
MONTH03MAR	-0.1294	0.2618	0.6210
MONTH04APR	-0.1256	0.2670	0.6382
MONTH05MAY	0.0265	0.2423	0.9129
MONTH06JUN	-0.3878	0.2709	0.1523
MONTH07JUL	0.0823	0.2399	0.7317
MONTH08AUG	-0.0528	0.2964	0.8586
MONTH09SEP	-0.0020	0.2839	0.9944

*Continued on next page*

Table 14 – *Continued from previous page*

Count Model	Estimates	Std. Error	p-value
MONTH10OCT	-0.3575	0.3369	0.2887
MONTH11NOV	-0.7933	0.3532	0.0247
MONTH12DEC	0.1819	0.2537	0.4733
SITEOBW	-0.0147	0.1205	0.9029
SITEOHS	-0.5745	0.1946	0.0032
INHIBIT	-0.0036	0.0021	0.0859
LOG(THETA)	1.9642	0.5854	0.0008
Binomial Model	Estimates	Std. Error	p-value
INTERCEPT	2.4463	0.8181	0.4837
ANIMALSHEEP	-2.3709	0.6812	0.0005
MONTH02FEB	-2.9411	0.9029	0.0011
MONTH03MAR	-1.7988	0.9722	0.0643
MONTH04APR	-0.6572	0.7262	0.3655
MONTH05MAY	-1.0157	0.6056	0.0935
MONTH06JUN	-1.6415	0.7575	0.0302
MONTH07JUL	-1.1520	0.6211	0.0636
MONTH08AUG	-0.1030	0.6155	0.8671
MONTH09SEP	-0.1638	0.5600	0.7699
MONTH10OCT	-0.6108	0.6632	0.3570
MONTH11NOV	-0.9728	0.8153	0.2328
MONTH12DEC	-1.3515	0.6918	0.0508
RESULTSPOS	1.4969	0.4640	0.0013
SEXM	-0.5717	0.3339	0.0869
SITEOBW	0.5932	0.3280	0.0706
SITEOHS	-0.7358	0.5299	0.1649

Table 14 confirms the result of the overdispersion test in Table 8, which concluded that egg counts of *C. isospora* are overly dispersed. This conclusion is reached by looking at the dispersion parameter in Table 15 from the count part of the model. The p-value associated with the dispersion parameter, LOG(THETA) is 0.008. As a result we reject the Poisson assumption and conclude that the variance of the process is far greater than the mean. Another process to test from table 15 is the zero inflation test between zero inflated negative binomial and the standard negative binomial model. Looking at the intercept from the binomial part of the model, it is both positive and not statistically significant (intercept=2.4463 with a p-value of 0.4837). From these findings a conclusion is made that, a ZINB model is preferred

to the standard NB model and that the process is indeed zero inflated. This is in harmony with the Poisson model zero inflation test.

From both the overdispersion and the zero inflation tests, one conclude that the distribution of *C. isospora* is best described by the zero inflated negative binomial model. Figure 9 and Figure 10 further supports these findings by showing hanging rootogram of the four distributions. A hanging rootogram align all deviations along the horizontal line. Bars are drawn from  $\sqrt{fitted}$  to  $\sqrt{fitted} - \sqrt{observed}$ . The hanging rootogram provides some insight into the pros and cons of each model. Apart from highlighting aspect of model fit such as deviation from the model, it also highlights potential overdispersion and excess zeroes problem.

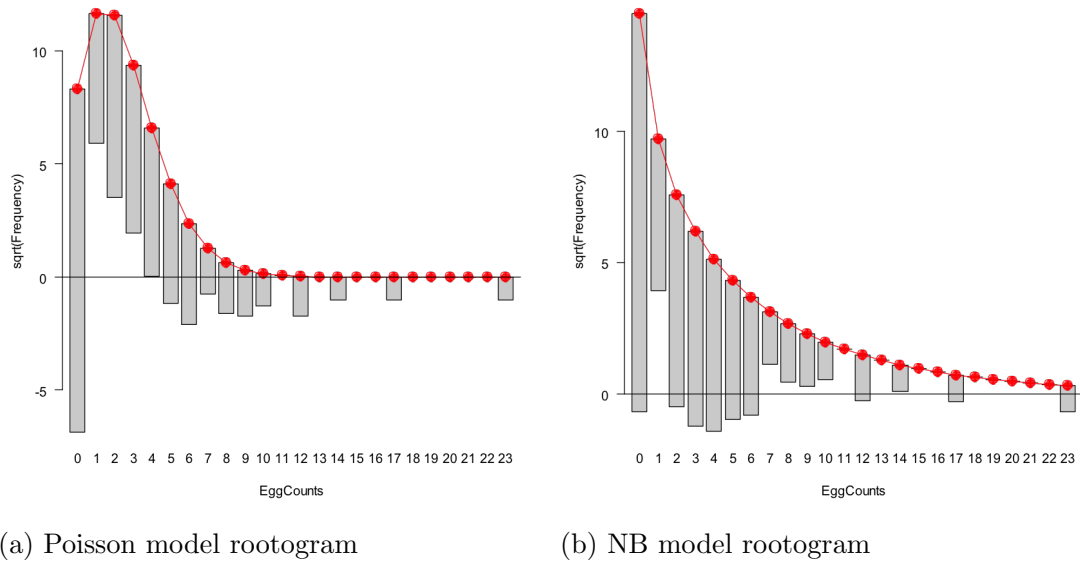
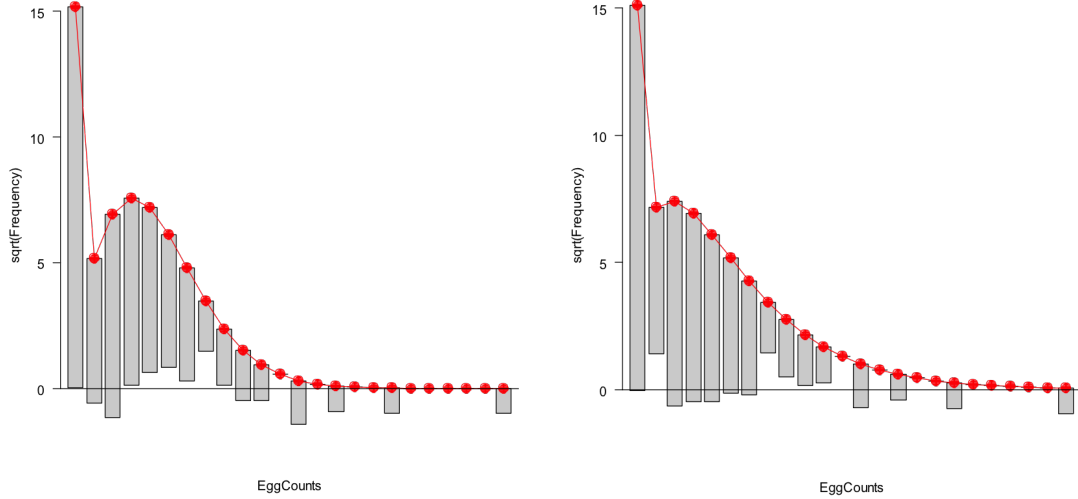


Figure 9: *Cooperia isospora* rootogram (Poisson and NB model)

The strong wave-like pattern of the bars indicates a high level of overdispersion. The Poisson rootogram shows that zero counts are highly under fitted while counts 1 to 4 are over fitted, indicating a problem of excess zeroes. The NB rootogram shows far fewer misfits compared to the Poisson and the wave-like pattern is not as severe as in the Poisson rootogram. This indicates that the NB model provides a better fit than the Poisson model and account for a great deal of overdispersion. Although no clear distinction is made between the ZIP and the ZINB rootogram fitting the



(a) ZIP model rootogram

(b) ZINB model rootogram

Figure 10: *Cooperia isospora* rootogram (ZIP and ZINB model)

distribution of *C. isospora* with standard count models provides worst fits compared to zero inflated models.

Interpreting Table 4.13 count model coefficients: *C. Isospora* egg counts are expected to be higher in February and December and at the lowest in November. Compared to January (which is the reference in the model) and holding other variables constant, *C. isospora* egg counts are expected to be 28% higher in February [ $\exp(0.2471) = 1.28$ ]. Compared to goats and keeping other variables constant, sheep are expected to have 62% more *C. isospora* egg counts than goats [ $\exp(0.4837) = 1.62$ ], this is statistically significant with a p-value of 0.0005. Looking at the covariate site, with site OA as the reference, higher counts of *C. isospora* eggs are expected for animals reared in site OA, while lowest counts are expected for animals in site OHS. Compared to site OA and keeping other variables constant, animals reared in site OHS are expected to have 78% more egg counts.

#### 4.2.7 Logistic regression model

Faecal egg counts are an indication of worm burden for some parasite species while for other species this is not the case. A logistic model can provide an indication of how well egg counts relates to worm burden. A logistic model was fitted to the data first to check for any relationship between egg counts and species abundance, secondly to compare the predicted abundance to that of zero inflated models. Table 15 shows results from the fitted logistic regression model. While egg count abundance models depended mostly on AGE, MONTH, optical density and SITE, logistic regression results in Table 15 indicates that the probability of a non-infection depends on ANIMAL, percentage inhibition, MONTH, SITE and *C. Isospora* egg counts (COO) as significant predictors.

Table 15: Logistic regression parameter estimates

DETERMINANTS	Estimates	Std. Error	p-value
INTERCEPT	2.8132	0.7206	0.0001
ANIMALSHEEP	-2.3928	0.6233	0.0001
COO	-0.2294	0.0752	0.0023
MONTH	1.8427	0.5838	0.0016
SITEOBW	-0.1195	0.3463	0.7300
SITEOHS	1.3291	0.5024	0.0082
INHIBIT	0.1736	0.0276	<0.0001

Similar to the previous models, individual groups are reference to individual factors. For ANIMAL, merino is the reference and for SITE, SITEOA is the reference. The covariate month is included in the model as a single effect to test the general effect across all months. Table 16 shows that that holding other variables constant, the odds of testing for negative for an infection for merino sheep is 91% [ $\exp(-2.3928)=0.092$ ] lower compared to the odds of testing negative for angora goats. The odds of testing negative for an infection is much higher in the month of February and December and also in site OHS. Percentage inhibition (OPTDENS) coefficient indicates that a 1% increase in percentage inhibition, increases the odds of testing negative for an infection by 19% [ $\exp(0.1736)=1.190$ ], with other variables held constant. Looking at the COO coefficient, an increase of *C. Isospora* egg counts by 1 decreases the odds of testing negative for an infection by 20% [ $\exp(-0.2294)=0.795$ ]. This

indicates a negative relation relationship between egg counts and the probability of testing negative for an infection. It is concluded that *C. Isospora* has a good egg laying capacity since the presence of egg means a lower chance of testing negative for an infection.

Figure 11 highlights the results in Table 16 and shows changes in the probability of testing negative with increasing *C. isospora* egg counts for some significant categorical predictors in the logistic model (ANIMAL and SITE).

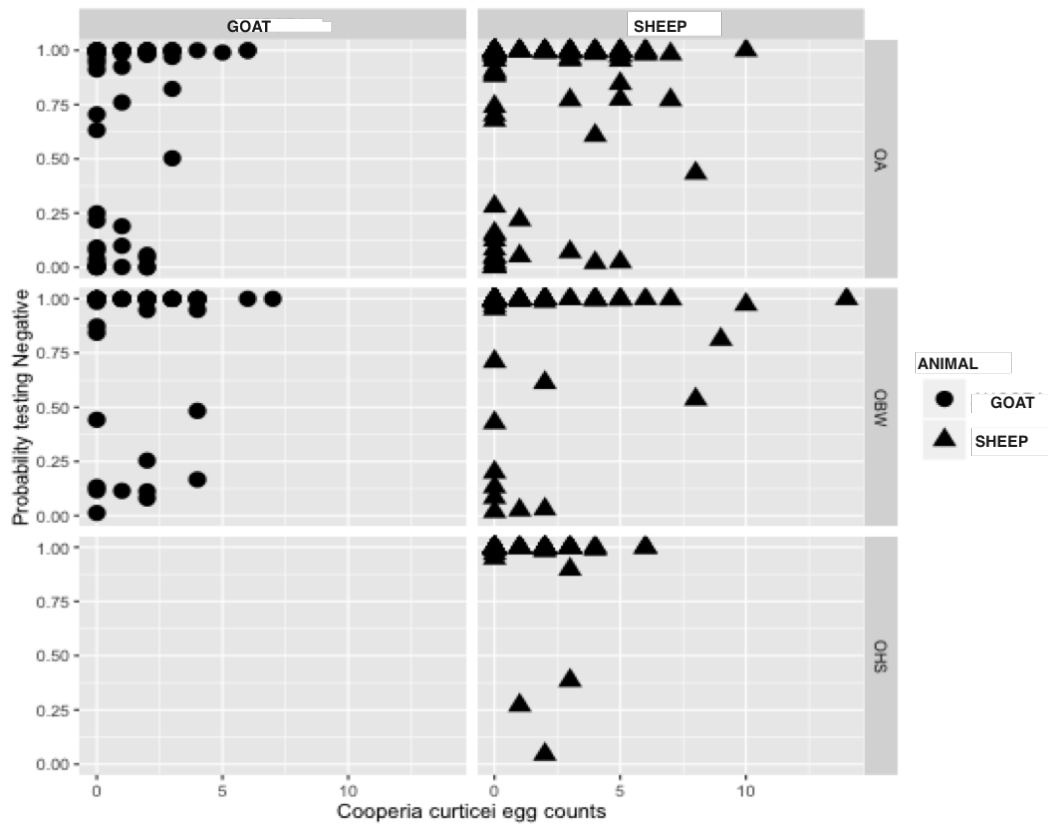


Figure 11: Predicted probabilities of testing negative for an infection

All animals reared in site OHS have lower egg counts of *C. isospora* and thus a higher probability of testing negative for an infection irrespective of whether they are sheep or goats. sheep reared in site OBW somewhat have a lower probability of testing negative for an infection compared to goats in site OBW. The same cannot be said about site OA since the patterns is not very clear. Most animals were reared

in site OA compared to other sites. We find that site OA with lower egg counts has higher probability of testing negative for an infection. Table 16 shows results (only predicted percentage of zeroes) of a Logistic regression model done on the actual egg counts and not results (testing negative or positive), after reducing actual counts to binary outcomes (absence or presence of parasite eggs).

Results from all data sets showed that zero inflated models predicted the number of zeroes better than the logistic regression models. Table 16 shows the observed percentage of zeroes in each dataset and those predicted by zero inflated models and those predicted by logistics regression models. For the zero inflated model, the best model between the ZIP, ZAP, ZINB and ZANB was used to calculate the predicted percentage of zeroes i.e. ZINB model was used for *C. Isospora*. For each of the parasite species, logistic regression consistently overestimated the number of zeroes while zero inflated models estimates were closer to the observed zeroes. Looking at *C. curticei* observed percentage of zeroes (53%) for example, the zero inflated negative binomial model predicted 48% of zero counts while the logistic regression model predicted 59% of zero counts. Sileshi et al (2009) however concluded that even though zero inflated models better explains the distribution of rare species, logistic regression often predict the number of zero counts in the data better.

### 4.3 Analysis of *Dictyocaulus filaria* egg counts

In the previous section egg counts of *C. isospora* were analysed, in this section a similar analysis is conducted on *D. filaria* egg counts to highlight some of the findings. Both Table 17 and Table 18 represent parameter estimates of the Poisson, NB, ZIP and ZINB models.

An overdispersion test (to test if the Poisson model is the appropriate model) can be conducted by looking at the dispersion parameter (THETA) from the NB model. The p-value of the dispersion parameter is 0.0045, the null hypothesis that the variance is equal to the mean is rejected. We concluded that the distribution of *D. filaria* is overly dispersed and the Poisson model does not account for such overdispersion (between the Poisson model and the NB model, the NB model is preferred). Zero inflation is tested using either ZIP model parameter estimates or ZINB model parameter estimates (using results from Table 19). Since a conclusion has been made that the negative binomial model is preferred to the Poisson model, zero inflation is tested using ZINB model and not the ZIP model.



Table 16: Predicted percentage of zeroes

Parasite species	Observed zeroes	Zero inflated Poisson	Zero inflated negative binomial	Logistic regression model
<i>B. decoloratus</i>	0.83	0.80	0.82	0.90
<i>R.e. evertysi</i>	0.57	0.56	0.54	0.62
<i>H. coturtus</i>	0.46	0.41	0.46	0.50
<i>C. curticei</i>	0.53	0.46	0.48	0.59
<i>C. eimeria</i>	0.70	0.63	0.69	0.88
<i>C. isospora</i>	0.84	0.80	0.84	0.98
<i>F. hepatica</i>	0.75	0.74	0.74	1.00
<i>D. filaria</i>	0.77	0.71	0.75	0.83
<i>O. pinnata</i>	0.77	0.71	0.76	0.90
<i>T. axei</i>	0.71	0.67	0.69	0.80
<i>T. ovis</i>	0.83	0.74	0.82	0.95
<i>S. Papillosus</i>	0.72	0.62	0.69	0.83
<i>T. gondii</i>	0.90	0.74	0.88	1.00
<i>P. cervi</i>	0.78	0.72	0.78	0.87
<i>O. colubianum</i>	0.79	0.68	0.78	0.76
$\chi^2$		143.75 (0.06883)	135 (0.1652)	145.69 (0.0481)

Table 17: Poisson and negative binomial parameter estimates

DETERMINANTS	Poisson Model			Negative binomial model		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
INTERCEPT	-5.5286	0.8941	<0.0001	-4.9425	1.4484	0.0006
ANIMALSHEEP	-1.9349	0.1578	<0.0001	-2.5942	0.2383	<0.0001
MONTH02FEB	0.3841	0.3345	0.2509	0.8662	0.5046	0.0861
MONTH03MAR	0.6536	0.3087	0.0342	0.8174	0.4888	0.0945
MONTH04APR	0.9218	0.2984	0.0020	0.8973	0.4818	0.0626
MONTH05MAY	0.2817	0.3491	0.4197	0.2426	0.5506	0.6595
MONTH06JUN	-0.0646	0.3690	0.8611	-0.1661	0.5645	0.7685
MONTH07JUL	0.6912	0.3009	0.0216	1.0060	0.4797	0.0360
MONTH08AUG	1.1087	0.2970	0.0002	1.7627	0.4676	0.0002
MONTH09SEP	0.2002	0.3487	0.5658	0.0906	0.5404	0.8668
MONTH10OCT	-0.8905	0.5622	0.1132	-0.8275	0.7351	0.2603
MONTH11NOV	-0.2355	0.4053	0.5612	-0.1758	0.5977	0.7687
MONTH12DEC	0.8145	0.3298	0.0135	1.5917	0.5166	0.0021
PCV	21.6028	3.1634	<0.0001	20.5747	5.0830	0.0001
RESULTSPOS	-	-	-	-0.8603	0.3269	0.0085
SEXM	0.6574	0.1654	0.0001	0.5984	0.2641	0.0234
SITEOBW	-0.3001	0.1278	0.0189	-	-	-
SITEOHS	-0.7816	0.2747	0.0044	-	-	-
INHIBIT	-0.0061	0.0025	0.0150	-	-	-
THETA				1.5620	0.1060	0.0045

Table 18: ZIP and ZINB model parameter estimates

Count model	ZIP Model			ZINB Model		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
INTERCEPT	-3.2290	0.8343	0.0001	-3.1117	1.0667	0.0035
PCV	1.4093	0.2901	0.0000	-	-	-
SITEOBW	-	-	-	-0.3425	0.1587	0.0310
SITEOHS	-	-	-	-0.3684	0.3887	0.3432
OPTDENS	-0.1239	0.0883	0.1604			
INHIBIT	-	-	-	-0.0009	0.0032	0.7689
LOG(THETA)	-	-	-	1.4289	0.5874	0.0050
Binomial model	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
INTERCEPT	3.2661	1.4001	0.0197	1.2872	0.6207	0.8357
ANIMALSHEEP	5.7309	1.3215	0.0000	-	-	-
MONTH02FEB	-2.4815	1.0422	0.0173	-1.9033	0.9126	0.0370
MONTH03MAR	-2.1053	0.9679	0.0296	-1.4497	0.8808	0.0998
MONTH04APR	-1.4613	0.9440	0.1216	-0.8535	0.8786	0.3313
MONTH05MAY	0.2066	1.2034	0.8637	0.9240	0.5847	0.8744
MONTH06JUN	0.9239	1.2988	0.4769	1.3198	0.6209	0.8317
MONTH07JUL	-2.4530	1.0144	0.0156	-1.2677	0.9100	0.1636
MONTH08AUG	-2.9254	1.0344	0.0047	-1.7542	0.8796	0.0461
MONTH09SEP	0.1248	1.1686	0.9150	0.9138	0.5692	0.8725
MONTH10OCT	2.3044	1.4991	0.1242	1.5700	0.6215	0.8006
MONTH11NOV	0.9929	1.3595	0.4652	1.4140	0.6213	0.8200
MONTH12DEC	-3.8715	1.0810	0.0003	-2.4754	0.9507	0.0092
RESULTSPOS	1.8994	0.6098	0.0018	1.6753	0.6290	0.0077
SEXM	-1.0887	0.4522	0.0161	-	-	-

The intercept from the zero inflation part of the model is positive and not statistically significant (p-value=0.8357), indicating that the ZINB model is preferred to the NB model. From the overdispersion test and the zero inflation test we conclude that the zero inflated negative binomial model describes the distribution of *D. filaria* egg counts model best.

Since *D. filaria* egg counts are best explained by the ZINB model, we interpret only coefficients from the ZINB model is Table 18. For all the categorical variables one group is used as a reference. For ANIMAL, GOAT is the reference, for RESULTS, the NEGATIVE group is the reference, for SEX, FEMALE is the reference and for SITE, OA is the reference. Unlike with *C. isospora*, month is not significant in explaining *D. filaria* egg counts, meaning there is no linear time trend. However month does have influence on the absence and presence of *D. filaria* egg counts (month was not significant in the count part of the model but is significant in the binomial part of the model). Compared to site OA and holding other variables constant, *D. Filaria* egg counts are expected to be 29% lower for animals reared in site OBW [ $\exp(0.3425)=0.7100$ ], this is statistically significant (p-value=0.0310).

Interpreting the zero inflated part of the model: holding other variables constant, the odds of having zero *D. Filaria* egg counts in February are 0.78 [ $\exp(-2.4815)=0.78$ ] times lower compared to January. The covariate RESULTS indicate that, the odds of having zero *D. Filaria* egg counts for animals that tested positive for an infection are 5.68 [ $\exp(1.8994)=5.68$ ] times higher compared to those that tested negative.

## 5 Simulation Study

In this section a simulation study is carried out with the aim of studying some properties of the zero inflated models and how they compare with the standard Poisson and negative binomial models using Markov Chain Monte Carlo (MCMC) simulation methods. These properties are then generalised to similar datasets. We also check the consistency of parameters and evaluate their predictive performance of each model. We begin the chapter by giving an overview of MCMC methods. In the simulation example, counts are simulated from a ZINB distribution, we fit four different models to the simulated data (Poisson, NB, ZIP and ZINB) and evaluate their predictive performance.

So far what has been done in the previous chapters was to model the probability that host  $y_i$  is found to have a particular number of egg counts, given a number of attributes  $x_i, x_{(i+1)}, \dots, x_p$ . That is we modelled the probability  $p(y_i|x_i)$  using a generalised linear model technique with a logistic link. All the datasets in the study were found to be zero inflated with a majority of them following a zero inflated negative binomial distribution due to both excess zeroes and overdispersion. Since this process is defined by a probability distribution, MCMC methods can help with simulating the behaviour of how egg counts are spread among hosts. Prior to explaining how MCMC methods can generate independent samples, we start by a brief definition of what a Markov Chain is. Stated loosely a Markov Chain is a stochastic process where future events are independent of past events given the current event. If the parameter  $\theta$  is drawn from the parameter space  $\Theta$ , with  $\theta^{(t)}$  being the current draw and  $\theta^{(t+1)}$  being the next draw then the Markov chain property is written as:

$$p[\theta^{(t+1)}|\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(t)}] = p[\theta^{(t+1)}|\theta^{(t)}].$$

The next draw  $\theta^{(t+1)}$  depends only on the current draw  $\theta^{(t)}$  through a transitional kernel (the probability of moving to the next state within the parameter space  $\Theta$  given the current state  $\theta^{(t)}$ ). MCMC simulation involves simulation of independent, random values from the posterior distribution. The empirical distribution of the simulated values will then provide an approximation to the posterior distribution, with the approximation improving as the sample size increases. One of the general MCMC algorithms is the Gibbs Sampler which generates random samples from a joint distribution when the distribution is unknown or difficult to calculate, using the conditional distribution as an approximation.

## 5.1 Gibbs sampler

The Gibbs sampler is a technique for generating random variables from a marginal distribution indirectly, without having to calculate the density (Casella and George, 1992). Using Gibbs Sampler one can sample from the joint distribution if the full conditional distribution of each parameter is known (Famoye and Singh, 2006). This is the case because according to Hammersley-Clifford theorem, the joint distribution  $f(x, y)$  is written in terms of the individual conditional densities  $f(x|y)$  and  $f(y|x)$ . The Hammersley-Clifford theorem states that a probability distribution with a positive density satisfies one of the Markov Chain properties (Besag, 1974). Suppose we are interested in sampling from a posterior distribution  $p(\theta|y)$ , with  $\theta$  being a vector of parameters:  $\theta_1, \theta_2, \dots, \theta_p$ .

- Choose a vector of starting values  $\theta^{(0)}$ .
- In no particular order draw:
  - $\theta_1^{(1)}$  from the full conditional distribution  $p(\theta_1|\theta_2^{(0)}, \dots, \theta_p^{(0)}, y)$
  - $\theta_2^{(1)}$  from the full conditional distribution  $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_p^{(0)}, y) \dots$
  - $\theta_p^{(1)}$  from the full conditional distribution  $p(\theta_p|\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_p^{(0)}, y)$ .
- Draw  $\theta^{(2)}$  using  $\theta^{(1)}$  continually using most updated values.

This sampling method is integrated in both SAS and R and random independent samples are generated using the following function RANDGEN in SAS or RZINB in R. To illustrate how the simulated empirical distribution approximates properties of the posterior distribution, a zero inflated negative binomial distribution with a mean of 0.8 and a variance of 3.36 is simulated. We then calculate both the mean and the variance, at different sample sizes. Table 19 shows the calculated means and variances under different sample sizes.

Table 19: Mean and Variance changes as sample size (N) increase

N	10	100	1000	
Mean	1.500	0.960	0.810	0.802
Variance	4.500	6.039	3.368	3.360

We can see that as N increases both the mean and the variance of the empirical distribution gets closer to their respective true values.

## 5.2 Simulation example

To validate findings from chapter 4, we generate data which are very similar (skewed counts with a high proportion of zeroes) to the frequency of the observed faecal egg counts. Samples are generated from a zero inflated negative binomial distribution with sample sizes of 100 and 500. Simulating from a ZINB distribution involved a two-step process of generating values from both the binomial (for zeroes in the data) and the negative binomial distribution (for both zero and positive counts). This is done to investigate both consistency of estimators for each distribution and also the effect of changes in sample sizes. We have six different cases with varying number of simulations and sample sizes.

Upon generating the zero inflated datasets, an intercept only Poisson, NB, ZIP and ZINB models were fitted to the data. The parameters:  $\mu$  (mean),  $k$  (dispersion parameter) and  $\pi$  (zero-inflation probability) were determined as applicable for each model. Based on the simulation study by Denwood, Stear, Matthews, Reid, Toft and Innocent (2008) to investigate the distribution of pathogenic nematodes in lambs, we chose the following values for the initial values of the parameters:  $\mu = 1.5$ ,  $k = 4.5$  and  $\pi = 0.65$ . The simulation results are represented in Table 20. The MSE for any given parameter  $\Theta$  (with  $\theta$  as the estimate of  $\Theta$ ) is calculated as the sum of the square of the standard error and the square of the bias.

$$\begin{aligned} MSE_{\Theta} &= E[(\Theta - \theta)^2] \\ &= (standard\ error)^2 + (bias)^2. \end{aligned}$$

In the first scenario (100 simulations each with a sample size of 100) the estimated mean from the zero inflated Poisson model is closer to the true value of the mean compared to the other models. The dispersion parameter of the negative binomial distribution has a minimum MSE compared to that of its zero inflated counterparts, indicating that the dispersion parameter  $k$  of the NB distribution is a consistent estimator compared to other models (Same applies to the zero inflated probability  $\pi$ , of the zero inflated Poisson model). In all cases none of the parameters from the Poisson distribution are closer to their true values as compared to other models, supporting the finding that standard Poisson models consistently provided the worst fit to overdispersed count data with excess zeroes. As number of simulations and sample size increase, parameters from the ZINB model tend to be more and more consistent. In the last scenario (10000 simulations each with a sample size of 500) all parameter from the ZINB model have minimum MSE, meaning they more consistent estimators. In the last scenario however, the ZIP model also a minimum

Table 20: Simulation results with MSE of parameters

Sim	N	Measure	Poisson	NB		ZIP		ZINB		
			$\mu$	$\mu$	$k$	$\mu$	$\pi$	$\mu$	$k$	$\pi$
100	100	Estimate	6.3010	0.9949	1.8700	1.1700	0.6169	1.2209	2.1600	0.4041
		Std.Error	1.3586	0.3893	0.5716	0.3158	0.2322	0.1357	0.3988	0.1798
		MSE	3.1772	1.2242	9.2736	0.9489	0.1929	0.9439	0.1691	0.2065
100	500	Estimate	3.3724	1.1534	1.3098	1.1181	0.5754	0.9931	1.3100	0.3699
		Std.Error	1.9890	0.3342	0.5668	0.2016	0.2010	0.1286	0.3991	0.1504
		MSE	2.0237	1.9748	0.1456	0.9322	0.1929	0.9124	0.1443	0.1998
1000	100	Estimate	3.1987	1.1489	1.2596	1.0012	0.5757	0.7743	1.1100	0.3561
		Std.Error	1.9770	0.3331	0.5799	0.1573	0.2212	0.1173	0.3902	0.1538
		MSE	1.8989	1.5212	0.1452	0.9310	0.1721	0.9001	0.1209	0.1853
1000	500	Estimate	2.5438	0.9688	0.8976	0.9661	0.5495	0.7901	0.7549	0.3343
		Std.Error	1.8950	0.3185	0.5431	0.1567	0.2121	0.1067	0.3883	0.1427
		MSE	2.8123	1.0865	0.1401	0.9287	0.1751	0.8974	0.1001	0.1745
10000	100	Estimate	2.4544	1.1578	0.8641	0.8174	0.5545	0.7883	0.7978	0.3259
		Std.Error	1.8850	0.2999	0.5433	0.1442	0.2314	0.1002	0.3837	0.1392
		MSE	0.9213	0.9961	0.1321	0.9141	0.1679	0.8921	0.0991	0.1453
10000	500	Estimate	2.4412	1.1574	0.7627	0.8014	0.5112	0.7811	0.7967	0.3217
		Std.Error	1.7540	0.2981	0.5429	0.1134	0.1615	0.1006	0.3761	0.1301
		MSE	0.9203	0.9866	0.1301	0.9012	0.1668	0.8912	0.0987	0.1403

$N_{sim}$  – number of simulations,  $n$  – sample size

MSE. MCMC algorithm estimated the empirical distribution better as the number of simulation increase, which is the case in last scenario with 10000 simulations each with a sample size of 500.

### 5.3 Evaluation of predictive performance

In this subsection each model's performance is evaluated in terms of its predictive capacity. To further support the findings that zero inflated models provide a better fit to count data with excess zeroes, we resampled 100 samples with replacement



from the original dataset using the boot function in R. For each of the resampled dataset, we fit the Poisson, quasi-Poisson (QP), NB, ZIP, ZINB, ZAP and ZANB models. Both the Pearson and spearman's rank correlation between observed and fitted values are computed. Table 212 shows the Pearson's correlation and the Pseudo  $R^2$  for all fifteen of the egg counts parasites species. While the Pearson's correlation indicates how close observed and predicted values are in relative terms, the Pseudo  $R^2$  value gives the percentage of variation explained by the fitted model. A high correlation between observed and predicted values supports the model as a better one compared to the other models. For each species the Pearson's correlation is in the first row and the Pseudo  $R^2$  value in the second row.

Table 21: Pearson's correlation ( $\rho_p$ ) and pseudo  $R^2$

Species	Poisson	QP	NB	ZIP	ZINB	ZAP	ZANB
<i>B. decoloratus</i> ( $\rho_p$ )	0.4524	0.4630	0.5220	0.5681	0.5623	0.5252	0.5252
<i>B. decoloratus</i> (pseudo $R^2$ )	0.6402	0.6403	0.7199	0.7341	0.7621	0.7327	0.7568
<i>R. evertsi</i> ( $\rho_p$ )	0.3972	0.3980	0.4741	0.4662	0.4831	0.4410	0.4464
<i>R. evertsi</i> (pseudo $R^2$ )	0.4361	0.4360	0.4508	0.4467	0.4482	0.4446	0.4426
<i>H. contortus</i> ( $\rho_p$ )	0.324	0.324	0.376	0.480	0.472	0.468	0.463
<i>H. contortus</i> (pseudo $R^2$ )	0.3474	0.3496	0.3510	0.3552	0.3553	0.3557	0.3553
<i>C. curticei</i> ( $\rho_p$ )	0.227	0.285	0.385	0.405	0.414	0.406	0.405
<i>C. curticei</i> (pseudo $R^2$ )	0.5144	0.5361	0.5445	0.5561	0.5982	0.5622	0.5631
<i>C. eimeria</i> ( $\rho_p$ )	0.250	0.255	0.277	0.317	0.288	0.287	0.287
<i>C. eimeria</i> (pseudo $R^2$ )	0.7674	0.7513	0.7596	0.7937	0.7977	0.7983	0.7982
<i>C. isospora</i> ( $\rho_p$ )	0.218	0.274	0.314	0.337	0.337	0.323	0.323
<i>C. isospora</i> (pseudo $R^2$ )	0.3591	0.3623	0.3713	0.3701	0.3726	0.3712	0.3731
<i>F. hepatica</i> ( $\rho_p$ )	0.174	0.175	0.178	0.180	0.175	0.254	0.225
<i>F. hepatica</i> (pseudo $R^2$ )	0.6281	0.6280	0.6291	0.6871	0.7001	0.7264	0.7178
<i>D. filaria</i> ( $\rho_p$ )	0.368	0.389	0.518	0.529	0.539	0.513	0.525
<i>D. filaria</i> (pseudo $R^2$ )	0.7120	0.7490	0.7537	0.7552	0.7517	0.7598	0.7522
<i>O. pinnata</i> ( $\rho_p$ )	0.250	0.340	0.378	0.389	0.346	0.339	0.339
<i>O. pinnata</i> (pseudo $R^2$ )	0.6212	0.6334	0.6502	0.6526	0.6528	0.6510	0.6506
<i>T. axei</i> ( $\rho_p$ )	0.383	0.333	0.421	0.484	0.495	0.462	0.457
<i>T. axei</i> (pseudo $R^2$ )	0.6614	0.6619	0.6704	0.6753	0.6765	0.6725	0.6726
<i>T. ovis</i> ( $\rho_p$ )	0.383	0.294	0.254	0.389	0.371	0.419	0.390
<i>T. ovis</i> (pseudo $R^2$ )	0.6326	0.6324	0.6355	0.6368	0.6366	0.6369	0.6361
<i>S. papillosus</i> ( $\rho_p$ )	0.350	0.351	0.353	0.485	0.503	0.443	0.444
<i>S. papillosus</i> (pseudo $R^2$ )	0.6443	0.6724	0.6650	0.6945	0.7071	0.6921	0.6934

Continued on next page

Table 21 – *Continued from previous page*

<b>Species</b>	<b>Poisson</b>	<b>QP</b>	<b>NB</b>	<b>ZIP</b>	<b>ZINB</b>	<b>ZAP</b>	<b>ZANB</b>
<i>T. gondii</i> ( $\rho_p$ )	0.096	0.105	0.155	0.244	0.108	0.128	0.142
<i>T. gondii</i> (pseudo $R^2$ )	0.6360	0.6601	0.7464	0.7725	0.7130	0.7084	0.7047
<i>P. cervi</i> ( $\rho_p$ )	0.495	0.502	0.503	0.515	0.507	0.510	0.505
<i>P. cervi</i> (pseudo $R^2$ )	0.7320	0.7424	0.7536	0.7554	0.7541	0.7540	0.7539
<i>O. columbianum</i> ( $\rho_p$ )	0.661	0.607	0.618	0.654	0.654	0.654	0.654
<i>O. columbianum</i> (pseudo $R^2$ )	0.410	0.355	0.432	0.424	0.424	0.429	0.429

Looking at both measures, the NB model consistently outperformed both the Poisson and the quasi-Poisson model, with higher  $\rho_p$  and pseudo  $R^2$  values. For all other parasite species zero inflated and zero altered models always recorded a higher Pearson's correlation and pseudo  $R^2$  compared to standard count models. This indicates that in addition to proving better understanding of aggregation patterns in parasites counts, zero inflated models are also better in terms of their predictive accuracy.

## 6 Zero Inflated Time Series Counts

As seen in the previous chapters, none of the standard count models could explain the distribution of internal parasite egg counts. Mean abundance was instead explained by either zero inflated or zero altered models, addressing the issue of both overdispersion and excess zeroes. What was not addressed in the previous chapters is possible temporal correlation between observations collected over time. In this chapter a further step is taken to account for autocorrelation between adjacent observations. It has been noted that failure to account for overdispersion and zero inflation results in misleading inference and false association between variables, the same applies when autocorrelation is unaccounted for. To account for autocorrelation, we fit observation-driven models for zero inflated time series to the data.

According to Yang, Cavanaugh and Zamba (2015) there are generally two types of time series models (which differ in the way they account for autocorrelation): observation-driven models and parameter-driven models. In observation-driven models, serial autocorrelation is modelled directly through a function of past response. In parameter-driven models, an unobserved latent process is employed to account for serial autocorrelation. In this study the focus is on observation-driven models due to their computational ease in statistical software.

### 6.1 Observation-driven models

In time series, autoregressive moving average (ARIMA) models are used to describe normally distributed processes. Winkelmann (2008) suggests the use of partial likelihood framework when the response time series is non-normal. It is assumed that the conditional distribution of the response time series  $\{Y_t\}$  belongs to an exponential family if it can be written in the form:

$$f(y_t|\zeta_{t-1}) = \exp \left[ \frac{y_t\theta_t - b(\theta_t)}{a_t(\phi)} + c(y_t, \phi) \right],$$

where  $\phi$  is the dispersion parameter and  $\zeta_{t-1}$  represent all known information at time  $t - 1$ . Similar to generalised linear models in chapter 3, first the conditional mean and variance are defined.

$$\mu_t = E(Y_t|\zeta_{t-1}) = b'(\theta_t)$$

and

$$\sigma_t^2 = Var(Y_t|\zeta_{t-1}) = b''(\theta_t).$$

The conditional mean is modelled through a systematic component just as outlined with GLM in chapter 3. For count data with excess zeroes, we specify both the ZIP autoregression and the ZINB autoregression and how they will be applied.

### 6.1.1 ZIP Autoregression (ZIP-AR)

According to Kedeo and Fokianos (2002), the probability mass function (PMF) of a ZIP autoregression is derived based on the zero inflated Poisson distribution. If  $\{Y_t\}$  is the response count series conditionally distributed as  $\text{ZIP}(\lambda_t, \pi_t)$ , then the PMF is written as follows:

$$f(y_t|\zeta_{t-1}) = \begin{cases} \pi_t + (1 - \pi_t)e^{-\lambda_t} & \text{for } y_t = 0 \\ (1 - \pi_t)\frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!} & \text{for } y_t = 1, 2, \dots \end{cases}$$

Similar to the ZIP distribution, the mean and the variance of the ZIP autoregressive is expressed as:

$$\begin{aligned} E(Y_t|\zeta_{t-1}) &= E[E(Y_t|\mu_t, \zeta_{t-1})] \\ &= E[(1 - \mu_t)\lambda_t|\zeta_{t-1}] \\ &= \lambda_t(1 - \pi_t). \end{aligned}$$

and

$$\begin{aligned} \text{Var}(Y_t|\zeta_{t-1}) &= E[\text{Var}(Y_t|\mu_t, \zeta_{t-1})] + \text{Var}[E(Y_t|\mu_t, \zeta_{t-1})] \\ &= E[(1 - \mu_t)\lambda_t|\zeta_{t-1}] + \text{Var}[(1 - \mu_t)\lambda_t|\zeta_{t-1}] \\ &= \lambda_t(1 - \pi_t) + \lambda_t^2\pi_t(1 - \pi_t). \end{aligned}$$

The variance to mean ratio is simply the ratio between the two.

$$\begin{aligned} \frac{\text{Var}(Y_t|\zeta_{t-1})}{E(Y_t|\zeta_{t-1})} &= \frac{\lambda_t(1 - \pi_t) + \lambda_t^2\pi_t(1 - \pi_t)}{\lambda_t(1 - \pi_t)} \\ &= 1 + \lambda_t\pi_t. \end{aligned}$$

The variance-to-mean ratio in this case is greater or equal to one, indicating that excessive zeroes in the data also result in overdispersion. For the ZIP autoregressive model overdispersion can only be accounted for by the zero inflation parameter  $\pi_t$ , resulting in the estimate of  $\pi_t$  being higher for the ZIP-AR model as compared to the ZINB-AR model.

### 6.1.2 ZINB Autoregression (ZINB-AR)

To simultaneously account for autocorrelation, overdispersion and zero inflation, Kedem and Fokianos (2002) extended the ZIP autoregression to a more conventional ZINB autoregression. Once again if  $\{Y_t\}$  is the response count series following a ZINB distribution, then the PMF is written as:

$$f(y_t|\zeta_{t-1}) = \begin{cases} \pi_t + (1 - \pi_t) \left( \frac{k_t}{k_t + \lambda_t} \right)^{k_t} & \text{for } y_t = 0 \\ (1 - \pi_t) \frac{\Gamma(k_t + y_t)}{\Gamma(k_t) y_t!} \left( \frac{k_t}{k_t + \lambda_t} \right)^{k_t} \left( \frac{\lambda_t}{k_t + \lambda_t} \right)^{y_t} & \text{for } y_t = 1, 2, \dots \end{cases}$$

The mean and the variance are written as follows:

$$\begin{aligned} E(Y_t|\zeta_{t-1}) &= E[E(Y_t|\mu_t, \zeta_{t-1})] \\ &= E[(1 - \mu_t)\lambda_t|\zeta_{t-1}] \\ &= \lambda_t(1 - \pi_t). \end{aligned}$$

and

$$\begin{aligned} Var(Y_t|\zeta_{t-1}) &= E[Var(Y_t|\mu_t, \zeta_{t-1})] + Var[E(Y_t|\mu_t, \zeta_{t-1})] \\ &= E[(1 - \mu_t)\lambda_t + (1 - \mu_t)^2\lambda_t/k_t|\zeta_{t-1}] + Var[(1 - \mu_t)\lambda_t|\zeta_{t-1}] \\ &= \lambda_t(1 - \pi_t) + \frac{\lambda_t^2(1 - \pi_t)}{k_t} + \lambda_t^2\pi_t(1 - \pi_t) \\ &= \lambda_t(1 - \pi_t)(1 + \lambda_t\pi_t + \lambda_t/k_t). \end{aligned}$$

The variance to mean ratio is simply the ratio between the two.

$$\begin{aligned} \frac{Var(Y_t|\zeta_{t-1})}{E(Y_t|\zeta_{t-1})} &= \frac{\lambda_t(1 - \pi_t)(1 + \lambda_t\pi_t + \lambda_t/k_t)}{\lambda_t(1 - \pi_t)} \\ &= 1 + \lambda_t\pi_t + \lambda_t/k_t. \end{aligned}$$

The variance-to-mean ratio in this case is greater or equal to one, indicating that overdispersion. It is also a function of both the zero inflation parameter  $\pi_t$  and the dispersion parameter  $k_t$ . This shows that for the ZINB-AR model, overdispersion is accounted for by both the zero inflation and dispersion parameter.

Table 22 shows different scenarios common in count data (autocorrelation, overdispersion and zero inflation) leading to different models applied throughout the study.

Table 22: Observation-drive time series models for overdispersion and zero inflation

Autocorrelation	Overdispersion	Zero inflation	Model
No	No	No	Poisson Regression
Yes	No	No	Poisson Autoregression
No	Yes	No	NB Regression
Yes	Yes	No	NB Autoregression
No	No	Yes	ZIP Regression
Yes	No	Yes	ZIP Autoregression
No	Yes	Yes	ZINB Regression
Yes	Yes	Yes	ZINB Autoregression

Prior to applying the specified model to the data, descriptions of some forecasting measures are given. These measures are also useful in model comparison and model accuracy. We start by assuming the following a series:

$$Z_1, \dots, Z_N, \dots, Z_{N+M},$$

of size  $N+M$ . The series is first divided into two parts, the first  $N$  observation used for parameter estimation and the remaining  $M$  observations used to describe measures of forecasting accuracy. If  $Z_{t+h}$  is the predicted mean of the  $h$ -step forecasting distribution, then predictive root mean square error (PRMSE) is defined as:

$$PRMSE(h) = \sqrt{\frac{1}{M-h+1} \sum_{t=N}^{N+M-h} (Z_{t+h} - \hat{Z}_{t+h})^2}.$$

The predictive mean absolute deviation is defined as:

$$PMAD = \frac{1}{M-h+1} \sum_{t=N}^{N+M-h} |Z_{t+h} - \hat{Z}_{t+h}|.$$

## 6.2 Illustrative examples

The purpose in this section is to illustrate the application of zero inflated autoregressive models. Despite applying the methodology to all the fifteen parasite species, we only show results for two of the parasite species (*Fasciola hepatica* and *Haemonchus contortus*), with the notion that the methodology can be applied to similar datasets. To address overdispersion, zero inflation and serial correlation we use the glarma and the ZIM package (Dunsmuir and Scott., 2015) to fit standard count models and their zero inflated counterpart respectively.

### 6.2.1 Application to *Haemonchus contortus* egg counts

*H. contortus* is a gastrointestinal nematode in both sheep and goats. Signs of infestation by the parasite include among others: anaemia, diarrhoea, dehydration and low packed cell volume. Just like other parasite species, infested livestock have marked lower growth rates and reduced reproductive performance. Figure 12 represents a time series plot of *H. contortus*

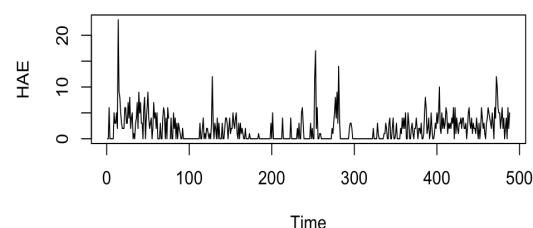


Figure 12: *Haemonchus contortus* time series plot (time in months)

As noted in earlier chapters, observed egg counts are generally lower, with a significant portion being zero counts. At some time periods higher egg counts are observed, indicated by spikes in the time series plot. There is also no clear upward or downward trend in the *H. contortus* series, indicating that there might be no need to have a trend variable in the time series model. To identify the autoregressive (AR) structure of the series, both plots of the autocorrelation function (ACFs) and the partial autocorrelation function are given in Figure

13.

The ACF plot indicate no correlation from lag 5 onwards, the partial ACF however shows that after taking out the effect in between correlation is zero from lag 2 onwards. This suggests an AR structure of order two for *H. contortus* series. However, Yang et al., (2015) pointed out that in the presence of zero inflation, identifying AR order with ACFs could be misleading. As a result we use R to identify the number of lags, by fitting the maximum number of lags and removing lags that were not significant.

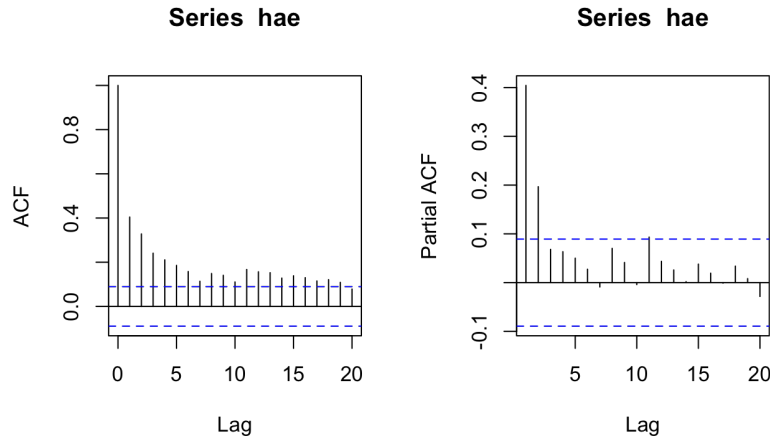


Figure 13: ACF and Partial ACF of *H. contortus*

The partial ACF in suggested an AR(2) model. Table 23 shows parameter estimates from the fitted ZINB autoregressive model of order two. The initial model fitted included a third autoregressive term, a trend variable and other covariates (age, packed cell volume and sex), which were later dropped because of their insignificance. Results from an overdispersion test are also provided, testing the fit of the ZIP autoregressive model against that of the ZINB autoregressive model. The p-value for the test is zero, indicating rejection of the null hypothesis that the dispersion parameter is equal to one. The ZINB autoregressive model thus provides a better fit compared to the ZIP autoregressive model.

From the fitted model we calculated the expected values for both the ZIP-AR and the ZINB-AR. Table 24 shows frequency distribution of expected and observed egg counts for the two of these autoregressive models together with their AIC. Using the minimum AIC criterion, the ZINB autoregressive model is favoured with a minimum AIC of 1667. The expected values for the ZINB model are also much closer to the observed *H. contortus* egg counts. To further investigate how the observed compare with the expected frequencies of each model we carried out the  $\chi^2$  goodness of fit test. From the  $\chi^2$  p-value for both the ZIP-AR and ZINB-AR (p-values=0.2243 and p-values=0.2289 respectively) models, we do not reject the null hypothesis that there is no significant difference between observed and fitted values. We can then conclude that the expected values for both models are closer to the observed frequencies. Figure 14 shows a plot of the observed and fitted values.



Table 23: ZINB autoregression parameter estimates

NB Model	Estimates	Std. Error	p-value
INTERCEPT	0.9595	0.0982	0.0001
ANIMALSHEEP	0.1521	0.0711	0.0323
AR1	0.0412	0.0097	<0.0001
AR2	0.0291	0.0103	0.0047
Binomial Model	Estimates	Std. Error	p-value
INTERCEPT	-1.1450	1.5402	0.0089
ANIMALSHEEP	2.1282	0.4361	<0.0001
SITEOBW	0.6534	0.3100	0.0351
SITEOHS	0.6303	0.3306	0.0566
AR1	-0.3273	0.0644	<0.0001
AR2	-0.1589	0.0569	0.0053

Test for overdispersion

SCORE TEST	10.6480		
P-VALUE	<0.0001		
AIC	1622		
BIC	1669		
TIC	1627		

From the fitted values we calculated the h-step predictive measures of accuracy (PRMSE and PMAD). Table 25 shows the estimates values together with their standard errors for two models, the AIC and the h-step predictive measures of forecasting accuracy. The h-step tells us the number of period ahead we can predict which PRMSE gives the error associated with the prediction. Thus between any two periods, a higher h-step is desirable given that it has a minimum PRMSE.

Table 24: Frequency distribution of observed and expected *H. contortus* egg counts

Counts	Observed	ZIP-AR	ZINB-AR
0	228	206	229
1	65	112	90
2	54	71	66
3	43	39	48
4	32	43	34
5	28	14	11
6	20	4	7
$\geq 7$	21	2	6
AIC		1667	1622
$\chi^2$		30 (0.2243)	56 (0.2289)

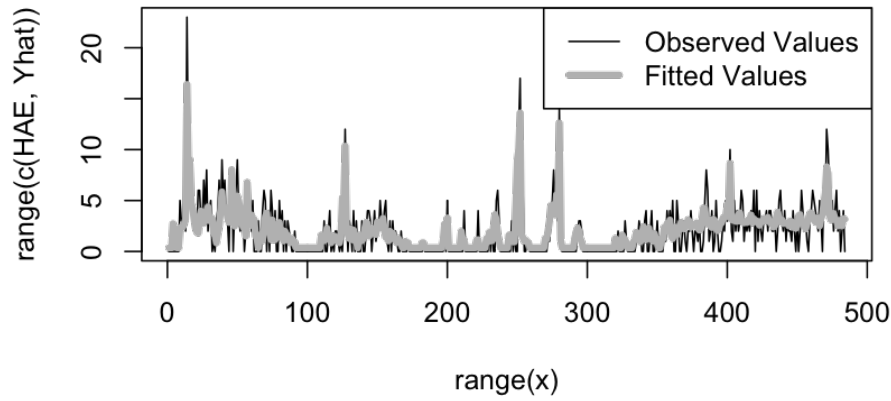


Figure 14: *H. contortus* observed and fitted egg counts

The ZINB autoregressive model has a dispersion parameter of 4.602, indicating a great deal of overdispersion. Both the ZINB and the ZIP autoregressive have the

Table 25: 5-step forecasting measures of accuracy

Model	Estimate	Std. Error	AIC	h-step	PRMSE	PMAD
<b>ZINB-AR(2)</b>	$\hat{k} = 4.602$	0.001	1622	1	0.589	0.337
	$\hat{\mu} = 2.610$	0.098		2	0.596	0.340
	$\hat{\pi} = 0.318$	0.438		3	0.605	0.344
				4	0.611	0.346
<b>ZIP-AR(2)</b>	$\hat{\mu} = 2.611$	0.069	1667	1	0.602	0.337
	$\hat{\pi} = 0.363$	1.540		2	0.602	0.340
				3	0.604	0.341
				4	0.609	0.346
				5	0.614	0.349

same estimated mean. As it is mostly the case, the estimated zero inflation probability  $\hat{\pi}$ , is larger for the ZIP model. This indicates that excessive zeroes in the data also results in overdispersion.

### 6.2.2 Application to *Fasciola hepatica* egg counts

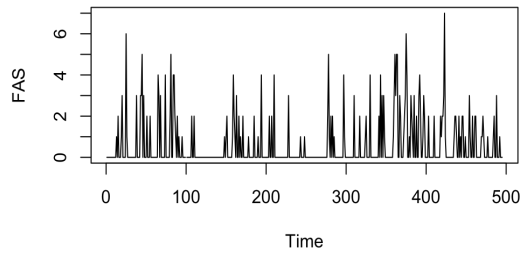


Figure 15: *Fasciola hepatica* time series plot (time in months)

*Fasciola hepatica* (also known as a sheep liver fluke) is a common internal parasite in grazing ruminants. Figure 15 shows a time series plot for *Fasciola hepatica* (*F. hepatica*).

Similar to *H. contortus*, there is also no clear upward or downward trend in the series. To identify the autoregressive (AR) structure of *F. Hepatica* series, both plots of the autocorrelation function (ACFs) and the partial autocorrelation function are given in Figure 16. The ACF plot indicates no correlation from lag 2 onwards. From both the ACF and the partial ACF plot, an AR(1) model

could be appropriate for this series. We also confirm the autoregressive structure using R.

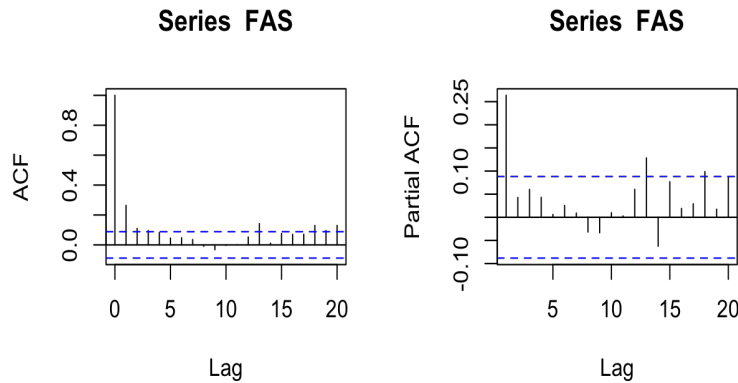


Figure 16: ACF and Partial ACF of *F. hepatica*

Table 26 shows parameter estimates from the fitted ZIP autoregressive model of order one. Results from an overdispersion test are also provided, testing the fit of the ZIP autoregressive model against that of the ZINB autoregressive model.

Table 26: ZIP-AR parameter estimates

NB Model	Estimates	Std. Error	p-value
INTERCEPT	0.8563	0.0977	<0.0001
SITEOBW	-0.5707	0.1668	0.0006
SITEOHS	0.0298	0.1632	0.8552
Binomial Model	Estimates	Std. Error	p-value
INTERCEPT	0.4530	0.2356	0.0546
ANIMALSHEEP	0.9319	0.2605	0.0003
AR1	-0.4004	0.0944	<0.0001
Test for overdispersion			
SCORE TEST	-0.5912		
P-VALUE	0.7154		
AIC	894		
BIC	920		
TIC	894		

The p-value for the test is 0.71538, indicating non-rejection of the null hypothesis. The ZIP autoregressive model thus provides a better fit compared to the ZINB autoregressive model.

## 7 Discussion and conclusion

### 7.1 Discussion

In this section we discuss what the results means in relation to the research objectives, with some support from literature. This chapter is concluded by stating which model generally worked well in understanding aggregation patterns in count data and what their shortcomings are. We start by discussing results in relation to the first objective of quantifying aggregation and zero inflation.

#### 7.1.1 Quantifying aggregation and zero inflation

Similar to our findings, Sileshi et al. (2009) found the ZINB model to generally have lower estimates for both  $k$  and  $\pi$  as compared to the negative binomial and the zero inflated Poisson model respectively. This is because for in the ZIP excess zeroes can only be accounted for by the zero inflation probability while for the ZINB model they can be accounted for by both the zero inflation probability and the aggregation parameter  $k$ .

Assuming the same fixed-effects structure, the reduction in the value of the zero inflation probability,  $\pi$ , from the ZIP to the ZINB model indicates how greatly distributional assumption and fixed-effects structure affect the zero inflation probability. This also indicates the significance of the particular covariate in explaining the distribution of parasites counts. In our case ANIMAL is the covariate that caused the most changes in both aggregation and zero inflation. When there are no covariates to explain the variation in each model, estimates for both the dispersion parameter and the zero inflation probability tend to be very high, indicating once again the importance of fixed-effects structure in quantifying aggregation and zero inflation.

#### 7.1.2 Characterise distributions applicable to count data and assessing their performance under zero-inflation

Despite the fact that Ziadinov et al. (2010) found zero inflated models to not necessarily provide in better fit in the presence of too many zeroes, we found that all the fifteen parasite egg counts were better explained by models for excess zeroes (zero inflated or zero altered models). Table 27 shows the AIC together with the AIC weight (for six of the models) of the 15 internal parasite species. For each species AIC is in the top row and AIC weight in the second row. A minimum AIC criterion is used to select the best model among a set of candidate models.

Table 27: Parasite species AIC and AIC weights

Species	Poisson	NB	ZIP	ZINB	ZAP	ZANB
<i>B. decoloratus</i>	666.65 0.00	616.64 0.00	<b>602.80</b> <b>0.67</b>	604.20 0.33	619.60 0.00	616.60 0.00
<i>C. curticei</i>	1357.70 0.00	1143.80 0.00	<b>1071.80</b> <b>0.96</b>	1079.60 0.02	1079.80 0.02	1081.80 0.01
<i>C. eimeria</i>	770.37 0.00	646.30 0.00	<b>622.40</b> <b>0.72</b>	624.40 0.27	631.60 0.01	635.20 0.00
<i>C. isospora</i>	1772.20 0.00	1543.10 0.00	1481.20 0.07	<b>1472.20</b> <b>0.88</b>	1478.60 0.01	1476.80 0.04
<i>D. filaria</i>	910.23 0.00	820.64 0.00	741.20 0.29	<b>739.40</b> <b>0.71</b>	754.00 0.00	754.80 0.00
<i>F. hepatica</i>	1161.50 0.00	963.71 0.00	911.60 0.34	913.40 0.14	<b>911.40</b> <b>0.38</b>	913.40 0.14
<i>H. contortus</i>	2118.30 0.00	1779.80 0.00	1676.20 0.00	<b>1630.20</b> <b>1.00</b>	1692.40 0.00	1651.20 0.00
<i>O. Columbianum</i>	749.00 0.00	704.72 0.00	656.20 0.30	657.60 0.15	<b>655.60</b> <b>0.40</b>	657.60 0.15
<i>O. pinnata</i>	1062.00 0.00	893.87 0.00	<b>841.60</b> <b>0.59</b>	845.00 0.11	843.60 0.22	845.60 0.08
<i>P. cervi</i>	925.54 0.00	815.38 0.00	<b>765.60</b> <b>0.49</b>	767.60 0.18	767.00 0.24	769.00 0.09
<i>R.e. evertsi</i>	1608.10 0.00	1441.80 0.00	1347.20 0.21	<b>1344.60</b> <b>0.79</b>	1360.00 0.00	1361.40 0.00
<i>S. papillosus</i>	1145.00 0.00	990.23 0.00	925.40 0.13	<b>921.80</b> <b>0.81</b>	929.00 0.02	928.00 0.04
<i>T. gondii</i>	713.73 0.00	495.61 0.00	<b>486.00</b> <b>0.59</b>	486.80 0.39	494.60 0.01	495.80 0.00
<i>T. axei</i>	1129.20 0.00	998.04 0.00	920.80 0.02	<b>913.40</b> <b>0.98</b>	942.00 0.00	944.00 0.00
<i>T. ovis</i>	978.62 0.00	744.75 0.00	713.80 0.05	712.20 0.10	<b>708.60</b> <b>0.62</b>	710.60 0.23

Only 20% (three out of the fifteen) of the datasets were best explained by the zero altered Poisson distribution, the remaining 80% (twelve out of the fifteen) was best fitted by either the zero inflated Poisson or negative binomial distribution (all datasets are best explained by excess zero models). Even though Sileshi G. (2008) found that standard count models can still provide best fit to data with excess zeroes, none of the fifteen datasets in this study were best fit by count models for excess zeroes. Vaudor, Lamouroux and Olivier (2011) found zero inflated models best fit data with a high proportion of zeroes while standard count models fit data with low proportion of zeroes. From table 28 it is evident that the negative binomial distribution, even though it does not provide the best fit, it consistently outperformed the Poisson distribution. To reinforce this finding, Andreas and Samu (2011) investigated overdispersion in bird migration and flocking behaviour and found the negative binomial distribution to be a better fit compared to the Poisson distribution.

The distribution of *H. contortus* for example; Poisson (AIC = 2118.3), NB (AIC = 1779.8), ZIP (AIC = 1676.2), ZINB (AIC = 1630.2), ZAP (AIC = 1692.4) while for ZANB (AIC = 1651.2). Using the minimum AIC criteria e.g. the distribution of *H. contortus* is best explained by the zero inflated negative binomial models. The AIC weight (the probability that a given model is the best among a set of candidate models) for *H. contortus* also indicate that the ZINB model ( $AIC_w = 1$ ) has the highest chance of being the best model among all six models.

### 7.1.3 Assessing the nature of seasonality

Table 29 shows all 15 parasite species in this study together with their type (whether they are flatworm, lungworm, roundworm, tapeworm, protozoa or ticks). Table 28 also shows the best zero inflated autoregressive model, the best zero inflated model from Chapter 4 together with their AIC. The numbers in the braces represent significant covariate in the models, which are explained at the bottom of Table 7.2. Models taking into account serial autocorrelation between observations performed better than those that did not. Using the minimum AIC criterion, it is clear from Table 28 that 87% (thirteen out of the fifteen) of the datasets are best explained by zero inflated autoregressive models (with the exception of only *B. decoloratus* and *T. gondii*).

Accounting for autocorrelation to some extent had an impact on distributional assumptions. For all parasite species (protozoa, ticks, flatworms, roundworms and tapeworms), accounting for autocorrelation did not change the distributional assumption. Prior to accounting for autocorrelation *C. Curticei* is best described by a ZINB model, after accounting for autocorrelation the distribution of *C. Curticei* is still described by a ZINB autoregressive models of order one ZINB-AR(1). However, for the only lungworm (*D. Filaria*) in the study, accounting for serial autocorrelation does affect the distributional assumption. The fixed-effects structure also depends vary with the type of parasite. Looking at all the six roundworms in Table 28, co-



variates shared among all these species are age (1), site (5) and optical density (7). Similarly all protozoa have ANIMAL and SITE as common covariates while all ticks have ANIMAL and percentage inhibition as common covariates.

Table 28: Model comparison for 15 parasite species: AIC

<i>B. decoloratus</i>		<i>C. curticei</i>		<i>C. eimeria</i>		<i>C. isospora</i>		<i>D. filaria</i>	
<b>T (2)(6)(7)</b>		<b>RW (1)(2)(3)(4)(5)(6)(7)</b>		<b>P (2)(5)</b>		<b>P (2)(3)(4)(5)(7)</b>		<b>LW (2)(3)(8)</b>	
ZIP AR(2)	ZIP	ZINB AR(1)	ZINB	ZIP AR(1)	ZIP	ZIP AR(3)	ZIP	ZIP AR(3)	ZINB
620	603	1293	1472	1059	1072	603	622	718	739.4
<i>F hepatica</i>		<i>H. contortus</i>		<i>O. columbianum</i>		<i>O. pinnata</i>		<i>P. cervi</i>	
<b>FW (2)(5)(6)</b>		<b>RW (1)(2)(5)</b>		<b>RW (2)(5)</b>		<b>RW (1)(2)(3)(5)(6)(7)</b>		<b>FW (2)(5)(8)</b>	
ZIP AR(1)	ZAP	ZINB AR(2)	ZINB	ZIP AR(1)	ZAP	ZIP AR(1)	ZIP	ZIP AR(3)	ZIP
890	911	1639	1642	649	656	794	841	748	757
<i>R.e. evertsi</i>		<i>S. papillosus</i>		<i>T. axei</i>		<i>T. gondii</i>		<i>T. ovis</i>	
<b>T(2)(6)(8)</b>		<b>RW (1)(2)(5)</b>		<b>RW (1)(2)(5)(7)(8)</b>		<b>P (2)(3)(5)</b>		<b>TW (1)(4)(5)(7)(8)</b>	
ZINB AR(3)	ZINB	ZINB AR(3)	ZINB	ZIP AR(1)	ZINB	ZIP AR(3)	ZIP	ZIP AR(3)	ZAP
1274	1345	913	921	907	913	671	486	671	709

(1) – AGE, (2) – ANIMAL, (3) – RESULTS, (4) – SEX, (5) – SITE, (6) – Percentage inhibition, (7) – Optical density, (8) – packed cell volume.

FW – Flatworm, LW – Lungworm, P – Prptozoa, RW – Roundworm, TP – Tapeworm, T – Tick.

#### 7.1.4 Significance of covariates in explaining FEC

Table 29 shows the percentage of explained deviance (egg count variation) for each covariate and parasite species. The last row in Table 29 is the average explained deviance across all fifteen parasite species. ANIMAL was the most important covariate in explaining egg counts variation, ANIMAL has an average explained deviance of 7%. This indicates that most of the species employed in this study are host specific. When applying treatment measures it should be noted that sheep and goats will require different attention. Other covariates of importance are MONTH, SITE and AGE with average explained deviance of 5%, 4% and 3% respectively. While some species thrive in temperate regions, some prefer humid regions while others reproduce better in periods of high rainfall. Three sites were available: OA, OBW and OHS. It is not clear how these sites differ. We can however conclude that; regions that differ in terms altitude, rainfall, temperature and wind will differ in parasite species abundance patterns. Livestock producers should take this into consideration.

## 7.2 Conclusion

Limitations of the Poisson distribution as highlighted include the assumption of randomness and a variance to mean ratio equal to one. The NBD solves this problem at the cost of precision of the model. The extra parameter  $k$  (dispersion parameter), results in decrease in precision. The dispersion parameter of the NB distribution depends on the sample size, as a results  $k$  is an unreliable measure of aggregation for different and smaller sample sizes. Only models for excess zeroes provided a better fit to the sheep and goats faecal egg counts. In our application we assumed each species is independent of each other, and thus require the use of univariate models. *H. Contortus* and *F. Hepatica* were found to be time dependent, multivariate modelling is also required to be done for future work to investigate any dependency among species.

Excess zeroes in count data affect the patterns of aggregation in the data. Based on their predictive performance and fit, zero inflated models are better at explaining the distribution of sheep and goats faecal egg counts, due to both excess zeroes present in the data and their nature of aggregation. For count data collected over time, it is important to account for serial autocorrelation between observations. Accurate prediction of seasonal variation in egg counts is important as it allows for timeous treatment measures. Autoregressive models capture these seasonal variations and also provide forecasting measures for them.

Table 29: Percentage of explaine deviance

SPECIES	AGE	ANIMAL	INHIBIT	MONTH	OPTDENS	PCV	RESULTS	SEX	SITE
<i>B. Decoloratus</i>	0.05	0.12	0.00	0.10	0.00	0.00	0.08	0.00	0.07
<i>C. Curticei</i>	0.00	0.06	0.00	0.05	0.00	0.00	0.00	0.01	0.03
<i>C. Isospora</i>	0.02	0.03	0.02	0.04	0.00	0.00	0.00	0.01	0.00
<i>D. Filaria</i>	0.06	0.08	0.00	0.08	0.02	0.00	0.00	0.00	0.06
<i>C. Eimeria</i>	0.06	0.19	0.00	0.13	0.05	0.00	0.00	0.00	0.05
<i>F. Hepatica</i>	0.00	0.08	0.00	0.06	0.00	0.00	0.04	0.00	0.02
<i>H. Contortus</i>	0.00	0.03	0.01	0.02	0.00	0.00	0.00	0.00	0.02
<i>O. Columbianum</i>	0.08	0.18	0.05	0.00	0.00	0.00	0.00	0.08	0.00
<i>O. Pinnata</i>	0.00	0.06	0.00	0.07	0.00	0.01	0.00	0.01	0.05
<i>P. Cervi</i>	0.06	0.10	0.04	0.00	0.00	0.04	0.05	0.00	0.00
<i>R. Evertsi</i>	0.00	0.06	0.00	0.04	0.00	0.00	0.00	0.00	0.03
<i>S. Papillosus</i>	0.00	0.05	0.00	0.04	0.00	0.00	0.00	0.00	0.09
<i>T. Ovis</i>	0.05	0.08	0.00	0.06	0.03	0.00	0.04	0.00	0.00
<i>T. Gondii</i>	0.07	0.00	0.05	0.00	0.00	0.09	0.05	0.00	0.00
<i>T. Axi</i>	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.04	0.18
Average	0.03	0.07	0.01	0.05	0.01	0.01	0.02	0.01	0.04

# Appendices

## A Moment Generating Function (MGF) of the Negative Binomial Distribution

To derive the moment generating function of the negative binomial distribution (NBD), we write PMF as:

$$P(Y = y) = \binom{y+r-1}{y} p^r (1-p)^y \quad y = 0, 1, 2, \dots$$

To derive the moment generating function of the NBD, the following identity is first proven:  $\binom{-r}{y} = (-1)^r \binom{r+y-1}{y}$ .

$$\begin{aligned} \binom{-r}{y} &= \frac{(-r)(-r-1)\dots(-r-y-1)}{y!} \\ &= (-1)^y \frac{(r+y-1)\dots(r+1)r}{y!} \\ &= (-1)^y \binom{r+y-1}{y}. \end{aligned}$$

We now derive the moment generating function of the NBD:

$$\begin{aligned} M_Y(t) &= \sum_{y=0}^{\infty} e^{tY} \binom{y+r-1}{y} p^r (1-p)^y \\ &= p^r \sum_{y=0}^{\infty} \binom{y+r-1}{y} [(1-p)e^t]^y \\ &= p^r \sum_{y=0}^{\infty} (-1)^r \binom{-r}{y} [(1-p)e^t]^y \\ &= p^r [1 - (1-p)e^t]^{-r} \\ &= \frac{p^r}{[1 - (1-p)e^t]^r} \end{aligned}$$

To derive expressions for the mean and variance, we find the MGF first and the second order derivative and evaluate them at zero.

$$\begin{aligned} M'_Y(t) &= -rp^r [1 - (1-p)e^t]^{-(r+1)} (-1)(1-p)e^t \\ &= \frac{rp^r (1-p)e^t}{[1 - (1-p)e^t]^{r+1}} \end{aligned}$$

and

$$\begin{aligned}
M_Y''(t) &= (1-p)e^t(-1)(r+1)(rp^r)[1-(1-p)e^t]^{-(r+2)}(-1)1-p)e^t \\
&\quad + (rp^r)[1-(1-p)e^t]^{-(r+1)}(1-p)e^t \\
&= (rp^r)(r+1)[1-(1-p)e^t]^2[1-(1-p)e^t]^{(r+2)} + \frac{(rp^r)(1-p)e^t}{[1-(1-p)e^t]^{r+1}}
\end{aligned}$$

The mean:

$$\begin{aligned}
M_Y'(0) &= \frac{rp^r(1-p)e^{(0)}}{[1-(1-p)e^{(0)}]^{r+1}} \\
&= \frac{rp^r(1-p)}{p^{r+1}} \\
&= \frac{r(1-p)}{p}
\end{aligned}$$

The variance:

$$\begin{aligned}
M_Y''(t) &= (rp^r)(r+1)[1-(1-p)e^{(0)}]^2[1-(1-p)e^{(0)}]^{(r+2)} + \frac{(rp^r)(1-p)e^{(0)}}{[1-(1-p)e^{(0)}]^{r+1}} \\
&= \frac{(rp^r)(r+1)(1-p)^2}{p^{r+2}} + \frac{(rp^r)(1-p)}{p^{r+1}} \\
&= \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p} \\
&= \frac{r(1-p)[1+(1-p)]}{p^2}
\end{aligned}$$

$$\begin{aligned}
\therefore \sigma^2 &= M_Y''(0) - [M_Y'(0)]^2 \\
&= \frac{r(1-p)[1+(1-p)]}{p^2} - \left[ \frac{r(1-p)}{p} \right]^2 \\
&= \frac{r(1-p)[1+(1-p)]}{p^2} - \frac{r^2(1-p)^2}{p^2} \\
&= \frac{r(1-p)[1+(1-p) - r(1-p)]}{p^2} \\
&= \frac{r(1-p)}{p^2}
\end{aligned}$$

## B R Codes

```
m <- read.delim("m.txt", header=T)
attach(m)
p15 <- mm[,15]

x1 <- m[,22]
x2 <- m[,24]

summary(p115 <- glm(p15 ~ AGE + BREED +
RESULTS + SEX + SITE , family = "poisson", data=mm))
drop1 (p115, test= "Chi")

dispersiontest(p115, trafo = 1)

res <- residuals(zip115, type="pearson")
plot(log(predict(zip115)), res)
abline(h=0, lty=2)
qqnorm(res)
qqline(res)

summary(p215 <- glm(p15 ~PCV + x1 + x2 + AGE + BREED
+ RESULTS + SEX, family = "poisson"))
drop1 (p215, test= "Chi")
summary(p315 <- glm(p15 ~PCV + x1 + x2 + AGE
+ RESULTS + SEX, family = "poisson"))
drop1 (p315, test= "Chi")
summary(p415 <- glm(p15 ~PCV + x1 + x2
+ RESULTS + SEX, family = "poisson"))
drop1 (p415, test= "Chi")
summary(p515 <- glm(p15 ~PCV + x1 + x2 + SEX, family =
"poisson"))
drop1 (p515, test= "Chi")

summary(qp115 <- glm(p15 ~ PCV + x1 + x2 + AGE + BREED
+ RESULTS + SEX + SITE, family = "quasipoisson", data=mm))
drop1 (qp115, test= "F")
summary(qp215 <- glm(p15 ~ PCV + x1 + x2 + AGE + BREED
+ RESULTS + SEX, family = "quasipoisson"))
drop1 (qp215, test= "F")
summary(qp315 <- glm(p15 ~ PCV + x1 + x2 + AGE
+ RESULTS + SEX, family = "quasipoisson"))
drop1 (qp315, test= "F")
summary(qp415 <- glm(p15 ~ PCV + x1 + x2
+ RESULTS + SEX, family = "quasipoisson"))
drop1 (qp415, test= "F")
```

```

library(MASS)
summary(nb115 <- glm.nb(p15 ~ BREED +
RESULTS + SEX + SITE, link = "log", data=mm))
drop1 (nb115, test= "Chi")

summary(nb215 <- glm.nb(p15 ~PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, link = "log", data=m))
drop1 (nb215, test= "Chi")
summary(nb315 <- glm.nb(p15 ~PCV + x1 + x2 +
RESULTS + SEX + SITE, link = "log", data=m))
drop1 (nb315, test= "Chi")

library(pscl)

summary(zip115 <- zeroinfl(p15 ~ 1 | x2 + BREED + RESULTS +
SITE,
dist = "poisson", data = mm))

summary(zip215 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX + SITE | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "poisson", data = m))
summary(zip315 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "poisson", data = m))
summary(zip415 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE +
RESULTS + SEX | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "poisson", data = m))
summary(zip515 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE +
RESULTS + SEX | PCV + x1 + x2 +
RESULTS + SEX + SITE, dist = "poisson", data = m))
summary(zip615 <- zeroinfl(p15 ~ PCV + x1 + x2 +
RESULTS + SEX | PCV + x1 + x2 +
RESULTS + SEX + SITE, dist = "poisson", data = m))

```



```
summary(zinb215 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX + SITE | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "negbin", data = m))
summary(zinb315 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "negbin", data = m))
summary(zinb415 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE +
RESULTS + SEX | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "negbin", data = m))
summary(zinb515 <- zeroinfl(p15 ~ PCV + x1 + x2 + AGE +
RESULTS + SEX | PCV + x1 + x2 +
RESULTS + SEX + SITE, dist = "negbin", data = m))
summary(zinb615 <- zeroinfl(p15 ~ PCV + x1 + x2 +
RESULTS + SEX | PCV + x1 + x2 +
RESULTS + SEX + SITE, dist = "negbin", data = m))
```

```
summary(zap115 <- hurdle(p15 ~ x1 | BREED +
SITE, dist = "poisson", data = mm))
```

```
summary(zap215 <- hurdle(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX + SITE | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "poisson", data = m))
summary(zap315 <- hurdle(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX + SITE | PCV + x1 + x2 +
RESULTS + SEX + SITE, dist = "poisson", data = m))
```

```
summary(zanb115 <- hurdle(p15 ~ x1 | BREED + SITE, dist =
"negbin", data = mm))
```

```
summary(zanb215 <- hurdle(p15 ~ PCV + x1 + x2 + AGE + BREED +
RESULTS + SEX + SITE | PCV + x1 + x2 + AGE +
RESULTS + SEX + SITE, dist = "negbin", data = m))
```

```

library(grid)
library(vcd)

xp <- goodfit(p15, type = c("poisson"))
rootogram(xp, xlab = "EggCounts", name = "rootogram")

xn <- goodfit(p15, type = c("nbinomial"))
rootogram(xn, xlab = "EggCounts", name = "Negative Binomial")

phat.zip <- predprob(zip615)
phat.zip.mn <- apply(phat.zip, 2, mean)
prd <- phat.zip.mn*495
obs <- c(389,37,34,22,9,2,2)
rootogram(obs,prd, xlab = "EggCounts")

phat.zinb <- predprob(zinb615)
phat.zinb.mn <- apply(phat.zinb, 2, mean)
prdd <- phat.zinb.mn*495
rootogram(obs,prdd, xlab = "EggCounts")

f.p <- fitted(p115)
f.qp <- fitted(qp115)
f.nb <- fitted(nb115)
f.zip <- fitted(zip115)
f.zinb <- fitted(zinb115)
f.zap <- fitted(zap115)
f.zanb <- fitted(zanb115)

c(cor(p15,f.p), cor(p15,f.qp), cor(p15,f.nb), cor(p15,f.zip),
cor(p15,f.zinb), cor(p15,f.zap), cor(p15,f.zanb))

c(cor(p15,f.p, method = "spearman"), cor(p15,f.qp, method =
"spearman"), cor(p15,f.nb, method = "spearman"), cor(p15,f.zip,
method = "spearman"), cor(p15,f.zinb, method = "spearman"),
cor(p15,f.zap, method = "spearman"), cor(p15,f.zanb, method =
"spearman"))

```

## References

- [1] Abdybekova A.M. and Torgerson P.R., 2012: “Frequency distribution of *Echinococcus multilocularis* and other helminths of foxes in Kyrgyzstan”. *Veterinary Parasitology*, 184(2-4), 348–351.
- [2] Ajiferuke I. and Famoye F., 2012: “Modelling count response variables in informetric studies: Comparing among count, linear and lognormal regression models”. *Journal of Informetrics*, 56(9), 499–513.
- [3] Alexander N., 2012: “Review: analysis of parasite and other skewed counts”. *Tropical Medicine and International Health*, 17(6), 684–693.
- [4] Andreas L. and Samu M., 2011: “Using the negative binomial distribution to model overdispersion in ecological count data”. *Ecology*, 92(7), 1414–1421.
- [5] Bailey C., Lopez S., Camero A., Taiquiri C., Arhauay Y. and Moore D.A.J., 2013: “Factors associated with parasitic infection among street children in orphanages across Lima, Peru”. *Pathogens and Global Health*, 107(2), 57–70.
- [6] Baines L., Morgan E.R., Ofthile M. and Evans K., 2015: “Occurrence and seasonality of internal parasite infection in elephants, *Loxodonta Africana*, in the Okavango Dekta, Botswana”. *Parasitology: Parasites and Wildlife*, 4(1), 43–48.
- [7] Beck M.A., Goater C.P., Colwell D.D. and Van Paridon B.J., 2014: “Fluke abundance versus host age for an invasive trematode of sympatric elk and beef cattle in southern Alberta”. *Parasitology: Parasites and Wildlife*, 3(3), 236–268.
- [8] Berk R. and MacDonald J.M., 2008: “Overdispersion and Poisson regression”. *Journal of Quantitative Criminology*, 23(1), 269–284.
- [9] Besag J., 1974: “Spatial Interaction and the Statistical Analysis of Lattice Systems”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 192–236
- [10] Casella G., and George E., 1992: “Explaining the Gibbs Sampler”. *The American Statistician*, 46 (3), 167 – 174.
- [11] Cox D.R. and Hinkley D.V., 1979: “Theoretical Statistics”. *Chapman and Hall, Florida*.
- [12] Crofton H.D., 1971: “A quantitative approach to parasitism”. *Parasitology*, 62, 179–193.
- [13] Directorate Statistics and Economic Analysis., 2015: “Abstract of agricultural statistics”. South African *Department of Agriculture, Forestry and Fisheries*.

- [14] Denwood M.J., Stear M.J., Matthews L., Reid S.W., Toft N. and Innocent G.T., 2008: “The distribution of the pathogenic nematode *Nematodirus battus* in lambs is zero-inflated”. *Parasitology*, 135(10), 1225–1235.
- [15] Diggle P.J., Heagerty P.J., Liang k-Y, and Zeger S.L., 2002: “Analysis of Longitudinal Data. 2d ed.”. *New York: Oxford University Press*.
- [16] Dobson A., 2001: “An Introduction to Generalized Linear Models. 2d ed.” *London: Chapman & Hall/CRC*.
- [17] Dunsmuir W.T.M. and Scott D.J., 2015: “The glarma Package for Observation-Driven Time Series Regression of Counts”. *Journal of Statistical Software*, 67(7).
- [18] Famoye F. and Singh K.P., 2006: “Zero-Inflated Generalized Poisson Regression with an Application to Domestic Violence Data”. *Journal of Data Science*, 2006(4), 117–130.
- [19] Fitzmaurice, G.M., Laird, N.M., and Ware, J.H., 2004: “Applied Longitudinal Analysis.” *New York: John Wiley & Sons*.
- [20] Gaba S., Ginot V. and Cabaret J., 2005: “Modelling macroparasite aggregation using a nematode-sheep system: the Weibull distribution as an alternative to the Negative Binomial distribution?”. *Parasitology*, 131(Pt 3), 393–401.
- [21] Gill J., 2001: “Generalized Linear Models: A Unified Approach, 07-134.” *Thousand Oaks, CA: Sage Publications*.
- [22] Hardin J.W. and Hilbe J.M., 2012: “Generalized Estimation Equations, Second Edition”. *Chapman and Hall, Florida*.
- [23] Kedem B. and Fokianos K., 2002: “Regression models for time series analysis”. *Wiley Interscience, New Jersey*.
- [24] Lewin W., Freyhof J., Huckstorf V., Mehner T. and Wolter C., 2010: “When no catches matter: Coping with zeroes in environmental assessment”. *Ecological Indicators*, 10(1), 572–583.
- [25] Linden A. and Mantyniemi S., 2011: “Using the negative binomial distribution to model overdispersion in ecological count data”. *Ecology*, 92(7), 1414–1421.
- [26] Lindsey J.K., 1997: “Applying Generalized Linear Models.” *New York: Springer-Verlag*.
- [27] MacLeod N.D., MacDonald C.K. and Van Oudtshoorn F.P., 2008: “Challenges for emerging livestock farmers in Limpopo province, South Africa”. *African Journal of Range & Forage Science*, 25(2), 71–77.

- [28] Maiti R., Biswas A., Guha A. and Ong S.H., 2014: “Modelling and coherent forecasting of zero-inflated count time series”. *Statistical Modelling*, 14(5), 375–398.
- [29] Marques J.F., Santos M.J. and Cabral H.N., 2010: “Aggregation patterns of macroendoparasites in phylogenetically related fish hosts”. *Parasitology*, 137(11), 1671–1680.
- [30] Meissner H.H., Scholtz M.M. and Palmer A.R., 2013: “Sustainability of South African livestock sector towards 2050”. *South African Journal of Animal Science*, 43(3), 282–297.
- [31] Minami M., Lennert-Cody C.E., Gao W. and Román-Verdesoto M., 2007: “Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing”. *Fisheries Research*, 84(2), 210–221.
- [32] Molenberghs, G., and G. Verbeke., 2005: “Models for Discrete Longitudinal Data.” *New York: Springer*.
- [33] Morel, J. G., and Neerchal, N. K., 2012: “Overdispersion Models in SAS.” *Cary, NC: SAS Institute Inc.*
- [34] O’Hara R.B. and Kotze D.J., 2010: “Do not log-transform count data”. *Methods in Ecology and Evolution*, 1, 118–122.
- [35] Potts J.M. and Elith J., 2006: “Comparing species abundance models”. *Ecological Modelling*, 199(15), 153–163.
- [36] Poulin R., 1993: “The disparity between observed and uniform distributions: A new look at parasite aggregation”. *International Journal for Parasitology*, 23(7), 937–944.
- [37] Shaw D.J. and Dobson A.P., 1995: “Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review”. *Parasitology*, 111(S1), S111–S113.
- [38] Sileshi G., Hailu G. and Nyadzi G.I., 2009: “Traditional occupancyabundance models are inadequate for zero-inflated ecological count data”. *Ecological Modelling*, 220(15), 1764–1775.
- [39] Stokes, M. E., C. S. Davis, and G. G. Koch., 2012: “Categorical Data Analysis Using the SAS System, Third Edition.” *Cary, NC: SAS Institute Inc.*
- [40] Ver Hoef J.M. and Boveng P.L., 2007: “Quasi-Poisson vs Negative Binomial regression: How should we model overdispersed count data?”. *Ecology*, 88(11), 2766–2772.

- [41] Vidyashankar A.N., Hanlon B.M. and Kaplan R.M., 2012: “Statistical and biological considerations in evaluating drug efficacy in equine strongyle parasites using fecal egg count data”. *Veterinary Parasitology*, 185(1), 45–56.
- [42] Vuong Q.H., 1989: “Likelihood ratio test for model selection and non-nested hypothesis”. *Econometrica*, 57(2), 307–333.
- [43] Wilson K., Grenfell B.T. and Shaw D.J., 2007: “Analysis of Aggregated Parasite Distributions: A Comparison of Methods”. *Functional Ecology*, 10(5), 592–601.
- [44] Winkelmann R., 2008: “Econometric Analysis of Count Data”. *Springer-Verlag, Berlin*.
- [45] Yang M., Cavanaugh J.E. and Zamba G.K.D., 2015: “State-space models for count time series with excess zeroes”. *Statistical Modelling*, 15(1), 70–90.
- [46] Yau K.K.W., Lee A.H. and Carrivick P.J.W., 2004: “Modelling zero-inflated count series with application to occupational health”. *Computer Methods and Programs in Biomedicine*, 74(1), 47–52.
- [47] Zeileis A., Kleiber C. and Jackman S., 2008: “Regression Models for count data R”. *Journal of Statistical Software*, 27(8), 111–136.
- [48] Zelterman, D., 2002: “Advanced Log-Linear Models Using SAS.” *Cary, NC: SAS Institute Inc.*
- [49] Ziadinov R., Deplazes P., Mathis A., Mutunova B., Abdykerimov K., Nurgaziev R. and Torgerson P.R., 2010: “Frequency distribution of *Echinococcus multilocularis* and other helminths of foxes in Kyrgyzstan”. *Veterinary Parasitology*, 171(3-4), 286–292.
- [50] Zuur A.F., Ieno E.N., Walker N.J., Saveliev A.A. and Smith G.M., 2009: “Mixed Effect Models and Extension in Ecology with R”. *Springer Science + Business Media, New York*.