UNIVERSITY OF KWAZULU-NATAL

Multivariate Bayesian Modelling of Tick Life-Stage Count Data Incorporating Spatial and Time Variations

Presented by

Thabo Lephoto

A Dissertation Submitted in Fulfilment of the Requirements of the Degree

Doctor of Philosophy

in

Statistics

Under the Supervision of

Professor Henry Mwambi

Doctor Oliver Bodhlyera

and

Professor Holly Gaff

College of Agriculture, Engineering and Science

School of Mathematics, Statistics and Computer Science

Pietermaritzburg Campus

2024

Declaration of authorship

I, Thabo Lephoto, declare that this thesis titled, 'Multivariate Bayesian Modelling of Tick Life-Stage Count Data Incorporating Spatial and Time Variations', contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any degree or diploma in any university or tertiary institution without the prior approval of the University of KwaZulu Natal.

- This work was done wholly and while in candidature for a research degree at this University.
- No part of this thesis has previously been submitted or any other qualification at this University or any institution.
- Where I have consulted the published work of others, this is always attributed.
- Where I have quoted from the work of others, the source is always given.

With such exceptions, this dissertation is entirely my own work.

Signed:  Date: 15/04/2024

Approved by supervisors:

Prof. Henry Mwambi Signed: [redacted] Date: 15/04/2024

Dr Oliver Bodhlyera Signed: [redacted] Date: 15/04/2024

Prof. Holly Gaff Signed: [redacted] Date: 15/04/2024

Acknowledgements

To Almighty God, I am grateful for His guidance and strength throughout the years of good experience working on my PhD dissertation.

This work was funded by NIH grant 1R01AI136035 as part of the joint NIH-NSF-USDA Ecology and Evolution of Infectious Diseases program. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscripts in this dissertation.

I want to thank my supervisors, Professor Henry Mwambi, Doctor Oliver Bodhlyera and Professor Holly Gaff, for their support in making the work of my PhD degree fruitful and enjoyable. The doors to your office were always open whenever I encountered a trouble spot or had a question about my research or writing. You consistently allowed this thesis to be my work but steered me in the right direction whenever you thought I needed it. I will always be grateful to have you by my side. Gratitude to Miss Christel Barnard for her administrative support throughout the years.

I must express my profound gratitude to my late father, Tumelo Lephoto, for his continued guidance and wisdom from when I was a boy. Thanks to my mother, Alinah Lephoto, for being there for me throughout my existence. Thank you to all my siblings, sisters and brothers and the Lephoto family as a whole for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This

accomplishment would not have been possible without them. Thank you.

Lastly, and most importantly, I would like to thank the woman in my life and the mother to our son, Ms Londeka Mbatha. Her unconditional support and love are the most critical drivers to this day.

List of publications

The contents of this dissertation are based on the following papers:

- **Lephoto, T**., Mwambi, H., Bodhlyera, O., and Gaff, H. (2024). Multilevel dynamic multivariate model applied to tick life-stage count data. *Submitted for consideration for publication.*
- **Lephoto, T**., Mwambi, H., Bodhlyera, O., and Gaff, H. (2024). Multivariate Bayesian dynamic tick life stage time series modelling and assessing impacts of sampled sites. *Submitted for consideration for publication.*
- **Lephoto, T**., Mwambi, H., Bodhlyera, O. and Gaff, H. (2021). Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients. *Geospatial Health*, 16(2).

The author has contributed to the following publications during the study period:

- Bansilal, S. and **Lephoto, T**. (2022). Exploring particular learner factors associated with South African mathematics learners' achievement: Gender gap or not. *African Journal of Research in Mathematics, Science and Technology Education*, 26(2), 77-88.
- Bansilal, S., **Lephoto, T**., North, D. and Zewotir, T. (2022). Exploring the association between teacher-related factors and Grade 9 mathematics achievement. *South African Journal of Education*, 42(1).

Dedication

To my son Maila "*Nino*" Philasande Lephoto

Abstract

Increasing tick abundance and tick-borne pathogens constitute a growing threat to public health. Five major ticks species were reported to be active and responsible for transmitting a variety of pathogens of both human and animals during the period of the original research, namely: *Ixodes scapularis, Amblyomma americanum, Dermacentor variability, Amblyomma maculatum*, and *Haeaphysalis longicornis*. This study uses tick life stages monthly data collected from 2009 to 2018 from 12 random sample sites within the edges, grass and woods habitat types in eight southeastern counties in Virginia, United States. The availability of time series and geo-referenced datasets for modelling has necessitated the development and application of dynamic time series and spatio-temporal statistical methods.

In this study, two Bayesian estimation techniques were investigated. The study compares the model overall performance by the Markov Chain Monte Carlo (MCMC) method with the Integrated Nested Laplace Approximation (INLA)technique. Multilevel Bayesian models were introduced using the INLA technique and its flexibility and computation time was demonstrated and compared with the MCMC technique. This dissertation contributed by developing models that enabled incorporation of the association among the components of response vector over time, while modelled as a function of time and space-time covariates. The models developed were successful in revealing environmental and time variations effects on the distribution of tick life stages. Alternative to the frequentist approach, the use of prior distributions in Bayesian modelling was useful for improving the model accuracy. Bayesian models offers flexible specification of complex models through the inclusion of random effects, hyperparameters and time-varying coefficients.

The study investigates the temporal behaviour of tick life cycle stages in each month, taking into consideration the effects of time trends, seasonal, and environmental variations. The association between tick life stages is modelled using

a multivariate Poisson, zero inflated Poisson and/or negative binomial distributions. Similar monthly time effects results were found in chapters 4, 5 and 6. Secondly, variation within the random sample sites was investigated using different prior distributions. This type of model was able to reveal that some areas in York, Portsmouth, Cheasapeake and Northampton counties had a significant higher tick variations while some areas in Portsmouth and Norfolk had significant lower variations.

Lastly, the study employed spatio-temporal models to unpack the effects of space over a period of time. The results showed that tick abundances were influenced by environmental factors and seasonal changes and it was concluded that tick abundances depend on the type of habitat where they are closer to their hosts and the time when their hosts are more likely to be targeted.

Contents

List of Tables **xiii**

List of Figures **xiv**

1 Introduction **1**

1.1 Background 1

1.2 The aim and objectives of this dissertation 4

1.3 Significance of the study 5

1.4 Dissertation layout 5

2 Literature review and preliminary concepts **8**

2.1 Literature review 8

2.1.1 Count data modeling 8

2.1.2 Modelling time series of counts 12

2.2 Preliminary concepts 15

2.2.1 Multivariate Poisson distribution 15

2.2.2 Markov Chain Monte Carlo sampling algorithms 17

2.2.3 Approximate Bayesian inference using INLA 19

3 Exploratory data analysis **23**

3.1 Data description 23

3.2 Exploratory results 24

4 Multilevel dynamic multivariate model applied to tick life-stage count data **27**

4.1 Introduction . . . 27

4.2 Model framework . . . 30

4.3 Model diagnostics . . . 34

4.4 Simulation study . . . 34

4.4.1 Estimation using INLA . . . 36

4.5 Tick life-stage modelling using multilevel model . . . 37

4.5.1 Data description . . . 38

4.5.2 Model framework for tick life stage data . . . 39

4.5.3 Results . . . 44

4.6 Discussion . . . 47

4.7 Conclusion . . . 48

5 Multivariate hierarchical modelling including sample sites as random effects **50**

5.1 Introduction . . . 50

5.2 Data description . . . 52

5.3 Model framework . . . 53

5.4 Simulations . . . 55

5.4.1 Simulation results . . . 57

5.5 Model application on tick life stage data . . . 57

5.6 Model diagnostics . . . 60

5.7 Model application results . . . 61

5.7.1 Sample sites random effects . . . 63

5.7.2 Non-linear effects . . . 64

5.8 Discussion . . . 65

5.9 Conclusion . . . 66

6 Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients **67**

6.1 Introduction 67

6.2 Data description 68

6.3 Model description 69

6.4 Model diagnostics 73

6.5 Fitted models 73

6.6 Model comparisons 75

6.7 Results 75

6.7.1 Space-covariate effects 75

6.7.2 Space-time effects 79

6.7.3 Temporal effects 81

6.8 Discussion 82

6.9 Conclusions 84

7 Summary, conclusion, limitations and future research **85**

7.1 Summary of the main findings 85

7.2 Conclusion 90

7.3 Limitations 90

7.4 Future research 90

References **94**

Appendices **113**

A Simulation R-code for chapter 4 **113**

A.1 R code for simulations in section 4.4 113

A.2 R code for computing MSEs 124

B Simulation R-code for chapter 5 **132**

C Journal articles **140**

List of Tables

3.1 Predictor variables and their type. 24

3.2 Tick counts in the data, 2009-2018. 24

3.3 Summary statistics of the data by habitat type, 2009-2018. 26

4.1 CPU time of model estimation using INLA and MCMC (seconds) 37

4.2 Estimated parameters in simulated models. 37

4.3 Averages of posterior means and MSEs of 500 replicates for $n = 5\,000$ 38

4.4 DIC values for the fitted models 1 to 8. 44

4.5 Estimated coefficients for covariates. 45

5.1 Estimated posterior parameters in the simulated model. 58

5.2 Summary of Poisson and negative binomial model DIC values for different priors. 61

5.3 Posterior estimates of the negative binomial models assuming **M1**: $\Gamma(0.1, 0.0005)$, **M2**: $\Gamma(0.001, 0.001)$ and **M3**: $\Gamma(0.5, 0.0164)$ priors in the precision matrix $\boldsymbol{\Omega}$. 63

6.1 DIC values for spatio-temporal Poisson and NB models. 76

List of Figures

3.1 Tick counts of larvae, nymph and adult species in Virginia, 2009-2018. 25

3.2 Plots showing time series of tick life stage counts in each habitat type, 2009 through 2018. 25

4.1 Model estimated non-linear month effects in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types. . . . 46

4.2 Estimated model trends in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types. 47

5.1 Sensitivity analysis on the priors of precision parameter of fitted negative binomial model. The black, red and green curves are for $\Gamma(1, 0.0005)$, $\Gamma(0.001, 0.001)$ and $\Gamma(0.5, 0.0164)$, respectively. . . . 62

5.2 Caterpillar plots for predictors of the random effects of sampling locations with posterior means (dots) and 95% credible intervals in Virgnia, USA. 64

5.3 Estimated shared month and year time trends for larvae, nymph and adult tick life stages. 65

6.1 Map of Virginia with names of counties where ticks were sampled. 69

6.2 Choropleth maps showing the space-covariate effects on larvae.[1] . 77

6.3 Choropleth maps showing the space-covariate effects on nymphs.[1] 78

6.4 Choropleth maps showing the space-covariate effects on adults.[1] . 78

6.5 Non-linear effects of month variable on larvae, nymphs and adults, 2009-2018.[1] 79
6.6 Choropleth maps showing residual spatial effect of larvae tick abundance, 2009-2018.[1] 80
6.7 Choropleth maps showing residual spatial effect of nymphs tick abundance, 2009-2018.[1] 80
6.8 Choropleth maps showing residual spatial effect of adult tick abundance, 2009-2018.[1] 81
6.9 The temporal year random effect for the cumulative best fitting model.[1 2] 82

Chapter 1

Introduction

1.1 Background

There is a vast amount of time series including geo-referenced data in many fields of study including ecological studies. Geo-referencing is usually by point referencing; that is, latitudes and longitudes or by areal referencing, which includes districts, counties, states, provinces and other administrative units, while time series data may include counts or measurements over a period of time. The availability of large geo-referenced and time series datasets for modelling has necessitated the development and application of spatial and time series dynamic statistical methods. This dissertation focuses on spatiotemporal and multivariate dynamic time series modelling of *Amblyomma americanum* tick counts as a function of time dependent and environmental covariates. The introductory part of this dissertation highlights the lack of dynamic time series models for modelling tick species and the importance of developing available spatial, spatiotemporal and multivariate dynamic time series models.

Spatial varying coefficients models exploring the abundance of tick counts remain limited. Efficient assessment of spatial and temporal patterns in species distribu-

tions, for example, to draw species distribution maps and to estimate population trends, has been a challenging but essential endeavour for ecologists and wildlife managers for a long time (Barker and Sauer, 1992; Inger et al., 2015). Species distribution or abundance models that specify population trends help understand basic ecological questions (Sattler et al., 2007). Reliable distribution and trend estimates are essential for a smart resource allocation in conservation (Rodrigues et al., 2006) and species management (Sutherland et al., 2004).

Models in literature for tick modelling involve GLMs, maximum entropy (MaxEnt), classification and regression tree (CART), species distribution modelling (SDM), ecological niche factor analysis (ENFA), and Bayesian hierarchical models among many others. Generalized linear models have been employed to investigate tick aggregation at different spatial ranges, provide robust estimates of tick densities between landscapes, and to determine and disaggregate the drivers of tick density and presence probabilities in an area (De Clercq et al., 2015; Kaizer et al., 2015; Estrada-Peña et al., 2016; Requena-García et al., 2017; Sagurova et al., 2019; Fan et al., 2023). Maximum entropy methods have been used to explore the limits of potential distribution by extrapolating the environmental requirements of ticks. These methods were also utilised to analyse the possible spatial range of tick species, and to explore how climate change can shape their distribution (de Oliveira et al., 2017; Raghavan et al., 2019; Ma et al., 2021; Lippi et al., 2021; Rochlin et al., 2023). Studies using SDM techniques lead to precise boundaries of the range of tick presence based on computational map modelling and demonstrate the way in which local populations of ticks differ in abundance towards the boundaries of the range (Qviller et al., 2016; Estrada-Peña et al., 2016; Lado et al., 2020; Kopsco et al., 2022; Rochlin et al., 2022, 2023). The CART technique has been used to review data on tick distribution and prevalence of tick-borne disease (TBD) for a national TBD management approach using the ecological and epidemiological information of ticks and the related diseases they transmit (Ly-

nen et al., 2008; De Clercq et al., 2015; Kjær et al., 2019). The ENFA technique has been used to measure the extent to which the preferences of a given species deviate from average conditions and the extent to which the species is selective over the range of environmental conditions available in a country (Estrada-Pena and Venzal, 2007; Lynen et al., 2008). Bayesian hierarchical models have been used in modelling prevalence of tick-borne pathogens and spatio-temporal distributions (Wimberly et al., 2008; Simpson et al., 2019; Lephoto et al., 2021; Clark and Wells, 2023).

Bayesian methods used in modelling tick abundances or related tick-borne diseases have employed the Markov Chain Monte Carlo (MCMC) sampling technique (Zipkin, 2012; Liu et al., 2017; Neupane et al., 2021; Clark and Wells, 2023; Wang et al., 2023). The Integrated Nested Laplace Approximation (INLA) technique can be used to conduct Bayesian inference as an alternative method to MCMC. The INLA technique approximates the posterior distribution based on the Laplace formulation, then it draws from the approximated posterior distribution to estimate the desired quantities and turns out to be more computationally efficient when compared to the MCMC approach (Arab, 2015; Ferkingstad et al., 2017; Khana et al., 2018; Van Niekerk et al., 2023).

A. americanum ticks have long been identified as nuisance biters because of their aggressive questing behaviour (Hair and Howell, 1970; Merten et al., 2000; Otálora-Luna et al., 2022). Despite its prominence as a nuisance biter and vector of human pathogens, efforts to define the species' variations within counties remain limited (Benham et al., 2021). Therefore, one of the strategies to help reduce tick-borne diseases is to have an in-depth analysis of the life stages abundance patterns and factors contributing in each of the randomly sampled locations. The aim of this dissertation was to unmask the heterogeneity within southeastern counties in Virginia. Thus, we modelled tick life stage abundance variations

within the randomly sampled locations. Tick counts were modelled as a function of temporal and environmental covariates.

1.2 The aim and objectives of this dissertation

Several applications of time series of count responses have shown the need to develop and relax model assumptions for accurate statistical modeling. For the $J-$dimensional vector of count responses observed at n locations over T regularly spaced times, the objective is to understand stochastic temporal patterns as a function of observed location-specific and/or time-varying covariates. For example, in ecology, understanding the causes and consequences of variation in the abundance of organisms as a function of environmental and topographical covariates has been a long-standing goal Scheiner and Willig (2011). The modeling described in this dissertation enables adequate incorporation of the association in the response over time as well as association between the components of the response vector.

Specifically, the main objectives are to build on existing models:

- that can jointly model tick life stages, determine the association among life stages, and find the relationship between life stage specific abundance and the environmental and temporal predictor variables.
- and extend them in order to identify tick hotspot areas using sample sites as random effects.
- to investigate spatial and temporal effects on the distribution of tick life stages.

We model multivariate time series of tick life-stage counts as a function of location-specific and time-dependent covariates. In practice, proper specification of the underlying distribution can be challenging due to the fact that count data can

exhibit over- or under-dispersion relative to the Poisson distribution, or an excessive number of zero counts. Another challenge is the complex modelling and implementation of multivariate time series spatial models. The study employs the Integrated Nested Laplace Approximation (INLA) technique by Rue et al. (2009) to model counts using spatial univariate models models. This technique is compared with the Monte Carlo Markov Chain (MCMC) for estimation of posterior parameters in Bayesian inference.

1.3 Significance of the study

The significance of this study is to develop appropriate and accurate dynamic statistical models for areal, multivariate and hierarchical time series of count data. We wish to develop hierarchical dynamic time series models and spatiotemporal models to accommodate the hierarchy in the data, while taking into account temporal and environmental or spatial covariate effects. The use and development of the suggested models is important to understand tick life stage dynamics, their correlations as well as the relationship between environmental or spatial and temporal effects. The models are important for unpacking tick hotspot areas, where these could be linked with their favourable hosts and temporal abundance fluctuations to avoid tick-borne diseases.

1.4 Dissertation layout

Chapter 2 briefly reviews modelling of count data, time series of counts, multivariate Poisson for count data modelling, the MCMC sampling algorithms and the approximate Bayesian inference using INLA.

Chapter 3 present the exploratory data analysis. This chapter describes the data and present summary statistics for tick life stages counts according the habitat

types.

Chapter 4 uses a multilevel dynamic model to investigate the influence of environmental and temporal predictors on the tick life cycle stages count outcome, and accounts for the association among the components of the response vector. This chapter addresses the first research objective through models that were implemented. The model implemented accounts for over-dispersion and allows for both positive and negative associations between the components of the response vector. The model is implemented because of its flexibility which allows joint modelling through a combination of different marginal count data distributions while allowing the building of dynamic models for the vector time series. The multivariate aspect is brought in through the introduction of the variance covariance matrix. The tick life-stage counts data time series is modeled as a function of habitat-type-specific and/or time-dependent covariates. The INLA approach is employed and compared with the MCMC to demonstrate its computational speed through a simulation study. Furthermore, the vector of counts is modelled using different marginal distributions to assess the model capabilities in fitting the data.

In Chapter 5, a hierarchical multivariate model is described for modelling tick abundance and investigating the relationship between location-time specific predictors while taking into account the random effects variations that are accounted for by the tick abundance variations within the sample sites. The aim was specifically to investigate and account for the heterogeneity in the sample sites by means of random effects. As a result, findings in this chapter will address the second research objective. The approximate Bayesian framework for fast estimation is described and the multivariate Poisson and negative binomial distributions were assumed for count data responses. This model helped to assess overall tick abundance across all sample sites that were within the habitat types. The use of the Poisson and negative binomial multivariate distribution also enabled modelling

the association between the components of the response vector as a function of location and time specific covariates that vary over time. Firstly, the study shows the use of multivariate Poisson to model the association between the components of the response vector and application of the methodology to an example from ecology.

In Chapter 6 the Poisson and NB count data models are applied and compared to explore the influence of environmental and temporal predictors on the distribution of tick counts. An assumption that the relationship between the predictors and the response variables in a regression model are constant across the study region and over time is relaxed. This assumption is unrealistic for spatial processes as factors such as sampling variation and different relationships across regions (for example, attitudes, preferences, culture and others) contribute to different responses to the same stimuli as one moves across regions and over time. The approach in this chapter address the last research objective. A frequent approach to spatial modeling dates back to the work by Besag et al. (1991) which was extended by Bernardinelli et al. (1995) to include a linear term for space-time interaction. Many studies have relaxed the assumption, for example, Assunçao et al. (1999) introduced a Bayesian space-varying parameter model to examine micro-region factor productivity and the degree of factor substitution in the Brazilian agriculture. In this study, the Bayesian spatio-temporal process model was used to allow covariates to vary spatially and over time.

Chapter 7.4 presents recommendations for future research and possible extensions of the work covered in this dissertation. Furthermore, the limitations are briefly described.

The R code used for simulating data is provided in the appendices at the end of this dissertation.

Chapter 2

Literature review and preliminary concepts

2.1 Literature review

2.1.1 Count data modeling

Increasing tick abundance and tick-borne pathogens constitute a growing threat to public health (Laaksonen et al., 2017; Paull et al., 2022; Wimms et al., 2023). Nadolny et al. (2014) recorded more than 66 000 questing tick populations in the southeastern Virginia, comprising 7 species from a variety of habitats, with *Amblyomma americanum* (Lone Star tick) constituting over 95% of the ticks collected. Five major tick species were reported to be active and responsible for transmitting a variety of human and animal pathogens namely, *Ixodes scapularis*, *A. americanum*, *Dermacentor variabilis*, *Amblyomma maculatum*, and *Haemaphysalis longicornis* (Nadolny and Gaff, 2018). The *H. longicornis* is of high veterinary importance for transmission of *Theileria orientalis* Ikeda in cattle even though it is not a key human-biting species (Oakes et al., 2019). The most common human-biting tick in Virginia is *A. americanum*, and it transmits several

pathogens of medical and veterinary significance, including *Ehrlichia chaffeensis* (the causative agent of human monocytic ehrlichiosis), Heartland virus (HRTV) and Bourbon virus (BRBV) among others (Goddard and Varela-Stokes, 2009; Savage et al., 2013, 2017)

Due to the effects of climate change and transportation by hosts, tick ranges are ever-changing. In recent years the *A. americanum* species has expanded its range into northeastern and midwestern United States and nothward into Canada (Molaei et al., 2019). A recent study in Kentucky reported detection of *Ehrlichia chaffeensis* from 32 counties out of the 77 sampled counties in 2019 through 2020 (Pasternak and Palli, 2023). The 2011 Centers for Disease Control and Prevention (CDC) report indicated that there were 863 human cases of ehrlichiosis in the United States. An increase was observed over a seven-year period, with human cases steadily increasing to 1 832, with 1 799 as a result of *E. chaffeensis infection* (Adams et al., 2013; CDC, 2021).

A. americanum ticks have long been identified as a nuisance biter because of their aggressive questing behaviour (Hair and Howell, 1970; Merten et al., 2000; Otálora-Luna et al., 2022). Despite its prominence as a nuisance biter and vector of human pathogens, efforts to define the species' variations within counties remain limited (Benham et al., 2021). Therefore, one of the strategies to help reduce tick-borne diseases is to have an in-depth analysis of the life stages abundance patterns and factors contributing in each of the randomly sampled locations.

Regression modeling of problems where the response variables are independent counts is diverse. The Poisson distribution is the most common approach used for modeling of count data. This approach implements generalized linear models as proposed by (McCullagh and Nelder, 1989). The Poisson regression model assumes that the mean and the variance are equal. In practice, counts often exhibit over-dispersion, that is, show evidence that the variance is larger than the

mean (Hinde and Demétrio, 1998; Cameron and Trivedi, 2001). When the data exhibit over-dispersion, the general solution is to introduce a dispersion parameter in the Poisson regression model or revert to the negative binomial model (Hinde and Demétrio, 1998; Cameron and Trivedi, 2001; Zeileis et al., 2008).

Alternatively, the Poisson-lognormal regression model, first proposed by Aitchison and Ho (1989), can be used. This model is a result from a lognormal distribution used as a random effect in the Poisson model (Agresti, 2010). The model is versatile because of its simple form which provides a convenient framework to carry out a series of typical multivariate statistical analyses (Chiquet et al., 2021), and includes multivariate sample comparison via linear discriminant analysis, dimension reduction through the use of principal component analysis (Chiquet et al., 2018), and network inference (Chiquet et al., 2019). Zhou et al. (2012) suggested lognormal and gamma mixed negative binomial regression model for counts since these have one extra degree of freedom to incorporate different kinds of random effects.

Wedderburn (1974) relied on the quasi-Poisson (correct-specification of the mean and variance relationship) in modeling non-standard response distribution as an alternative to the standard Poisson regression model. A Poisson-mixture model, also known as the negative binomial model as proposed by Greenwood and Yule (1920) is the standard parametric model used to account for over-dispersion.

Lambert (1992) introduced zero-inflated count data models to handle excess zero counts in the data. Species zero-count data may arise from many situations; that is, species zero-counts at a location or site due to observations that are attributable to unsuitable habitats or suitable habitats that are unoccupied (Martin et al., 2005). Some examples of zero-inflated Poisson (ZIP) models include studies by Kuhnert et al. (2005) to assess the impact of grazing on bird densities in woodland habitats and Martin et al. (2005) to model prevention in dental epidemiology,

among others.

Univariate analysis of count data is no longer sufficient when the primary focus of the analysis of multivariate count data is to describe the association among the counts (Alfò and Trovato, 2004). Although count data has been the subject of an increasing number of proposals in many scientific areas, models taking into account multivariate counts are still very rare. Bivariate Poisson distribution modelling has been conducted by Johnson and Kotz (1969), Cameron and Johansson (1997), and Karlis (2003). Some of the works that have analyzed multivariate count data include the semi-parametric mixture model by Alfò and Trovato (2004), the negative binomial regression model by Famoye (2010), and the multivariate generalized Poisson regression model proposed by Famoye (2015) based on a multivariate distribution with several parameters to model the over-dispersion, and several parameters to model the correlation between the count variables.

The disadvantage in the bivariate Poisson models by Johnson and Kotz (1969); Cameron and Johansson (1997); Karlis (2003) is that the association between components of the vector-valued count variable is restricted to being positive. This assumption is not realistic for several applications. Karlis and Meligkotsidou (2007) relaxed the assumption by proposing finite multivariate Poisson mixture as an alternative class of models for multivariate count data. The models allow for both negative and positive association and over-dispersion, although computations can take a long time.

Aitchison and Ho (1989) described a multivariate Poisson-lognormal model approach which allows for modeling of the association within the response vector for data that are possibly over-dispersed. A Bayesian approach for the regression model was developed by Chib and Winkelmann (2001). However, the possibility of temporal association was not accounted for when the data was collected

on different sites or locations over time. The literature describing properties for the induced association was still missing, although Aitchison and Ho (1989) noted that the range of the induced associations is not as wide as that of the corresponding lognormal or normal distributions. Serhiyenko (2015) carefully quantified the values of the minima and maxima of the induced associations between the count variables as a function of the marginal Poisson means, correlations and variances of the components from the underlying lognormal distributions.

There is a rising interest in the development and application of spatial and spatio-temporal statistical methods for analysis of geographically correlated count data as well as data that vary over time. This can be attributed to the increasing availability of space-time data in many fields of study including ecology as discussed by (Khaemba and Stein, 2001). Khaemba and Stein (2001) illustrated the use of spatial statistical methods to understand wildlife abundance, distributions and variability of diversity in space and time. Spatio-temporal modeling has seen many applications in modeling of tick-borne disease cases Rosà et al. (2018); Rau et al. (2020); Neupane et al. (2021); Mutizhe et al. (2022). Although these models have been applied extensively, their application regarding tick abundance is sparse. It would be interesting to know how environmental and temporal factors contribute to the tick life-stage abundance distribution.

2.1.2 Modelling time series of counts

The literature on time series modeling of count data includes parameter driven and observation driven models. The GLM (generalized linear models) framework by McCullagh and Nelder (1989) allows combining of traditional time series models such as autoregressive (AR) models, moving average (MA) models, autoregressive moving average (ARMA) models, autoregressive integrated moving average (ARIMA) models, seasonal ARIMA models, autoregressive conditional het-

eroskedasticity (ARCH) models and generalized ARCH (GARCH) models (Davis and Liu, 2012; Douc et al., 2017).

Various statistical models have been developed to model count data, but the Poisson distribution is usually the first choice of the underlying distribution for the time series modeling of counts under the GLM framework. In practice, the classical Poisson regression model is often of limited use because count data sets typically exhibit over- (or under-) dispersion and/or excess number of zeroes. The former issue is usually addressed by extending the standard Poisson model in various directions such as using the sandwich covariance structure or estimating an additional dispersion parameter in a quasi-Poisson model, or using the negative binomial regression model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). The latter issue of excess zeroes is addressed by introduction of zero-inflated Poisson and negative binomial models Mullahy (1986); Lambert (1992).

Various statistical packages can be used for model estimation, diagnostics, and forecasting. The GLM framework is implemented in R (R Core Team, 2022) statistical package through model fitting functions such as `glm()` for Poisson and `glm.nb()` for negative binomial in the **stats** and **MASS** packages, respectively (Bates et al., 1992). The **tscount** R packages provide a flexible framework for estimation of count time series in GLM. Another package in R, called **glarma**, gives functions for fitting generalized linear autoregressive moving average (GLARMA) models for discrete-valued time series (Dunsmuir and Scott, 2015).

Count model estimation can be done using likelihood methods. The method of maximum likelihood inference for Poisson and negative binomial time series models has been developed by Davis et al. (2003); Christou and Fokianos (2014). The method of quasi-likelihood was discussed by Zeger (1988). Chib et al. (1998) discussed Bayesian modeling of panel count data under the Poisson-lognormal model, and Chib and Winkelmann (2001) discussed models with latent effects for

count correlated data.

Another approach to develop models for count time series, proposed by Steutel and van Harn (1979), is based on the thinning operator where the thinning operator is generated by counting series of Bernoulli distributed random variables. The INAR (1) (first-order integer-valued autoregressive) model was developed by McKenzie (1985), and later by Al-Osh and Alzaid (1987). However, integer-valued time series based on thinning operations is more restrictive in its construction than models under the GLM framework (Tjøstheim, 2012). A drawback is that these models have assumed a positive autocorrelation. A review of these models has been done by (Weiß, 2008). Multivariate INAR models for counts were also discussed in Pedeli and Karlis (2011), and recently in Bermúdez and Karlis (2021).

Montesinos-López et al. (2017) developed a MCMC (Bayesian Markov Chain Monte Carlo) model with noninformative priors, which allows obtaining of all required full conditional distributions of the parameters leading to an exact Gibbs sampler for the posterior distribution. These models were developed to help and provide tools for plant breeders performing genomic selection for multiple-traits and multiple-environments for count data. However, one drawback might be the computational time taken by the MCMC. Serhiyenko (2015) proposed LCMs (level correlated models) and implemented them in R, under the INLA (integrated nested laplace approximation) framework, proposed by Rue and Martino (2007); Rue et al. (2009), to take into account the association among the components of the multivariate response vector as well as over-dispersion. The INLA method is an appealing alternative to the MCMC because of its speed and ease of use through the R-INLA package in R statistical software (R Core Team, 2022). In this study, computational times are compared between INLA and the MCMC through multilevel multivariate modelling to tackle both negative and positive correlations in multivariate count outcome data.

2.2 Preliminary concepts

2.2.1 Multivariate Poisson distribution

The definition of a J-variate Poisson distribution is based on a mapping $g : \mathbb{N}^q \to \mathbb{N}^J$, $q \geq J$, such that $\boldsymbol{Y} = g(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{X}$ (Mahamunulu, 1967), where $\boldsymbol{X} = (X_1, \ldots, X_q)'$, is a vector of unobserved independent Poisson random variables; that is, $X_r \sim \text{Poisson}(\lambda_r)$ for $r = 1, \ldots, q$, and $\boldsymbol{A}$ is an arbitrary $J \times q$ matrix which determines the properties of the multivariate Poisson distribution (MP). Here, the J-dimensional vector $\boldsymbol{Y} = (Y_1, \ldots, Y_J)' = \boldsymbol{A}\boldsymbol{X}$ follows a multivariate Poisson distribution (MP_J) with parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)'$ and its probability mass function (PMF) is

$$\begin{aligned} MP_J(\boldsymbol{y}|\boldsymbol{\lambda}) &= P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\lambda}) \\ &= \sum_{x \in g^{-1}(y)} P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\lambda}) \\ &= \sum_{x \in g^{-1}(\boldsymbol{y})} \prod_{r=1}^{q} P(X_r = x_r|\lambda_r), \end{aligned} \tag{2.1}$$

where $\boldsymbol{y}$ denotes a vector of realisations y_j's, $g^{-1}(\boldsymbol{y})$ denotes the inverse image of $\boldsymbol{y} \in \mathbb{N}^J$ and for $r = 1, \ldots, q$, the PMF of the univariate Poisson distribution is

$$P(X_r = x_r|\lambda_r) = \frac{\exp(-\lambda_r)\lambda_r^{x_r}}{x_r!}.$$

A two-way covariance structured multivariate Poisson distribution was proposed by Karlis and Meligkotsidou (2005) in order to permit a more realistic modeling of multivariate counts for several practical applications. The construction of the distribution is achieved by setting $\boldsymbol{A} = [\boldsymbol{A}_1 \boldsymbol{A}_2]$, where $\boldsymbol{A}_1 = \boldsymbol{I}_J$ is a J dimensional identity matrix, and $\boldsymbol{A}_2$ is a $J \times \left(\frac{J(J-1)}{2}\right)$ binary matrix. In a sense of Karlis and Meligkotsidou (2005), $\boldsymbol{A}_1$ is referred to as the "main effects" and $\boldsymbol{A}_2$ is referred

to as the matrix of "two-way covariance effects"; where each column of $\boldsymbol{A}_2$ has 2 ones and $(J-2)$ zeros and no duplicate columns exist; and $q = J + [J(J-1)]/2$. Correspondingly, when the parameter $\boldsymbol{\lambda}$ is split in two parts, then $\boldsymbol{\lambda}^{(1)} = (\lambda_1, \ldots, \lambda_J)'$, which corresponds to the J main effects, and $\boldsymbol{\lambda}^{(2)} = (\lambda_{J+1}, \ldots, \lambda_q)'$ which corresponds to the $J(J-1)/2$ pairwise covariance effects. Suppose $J = 2$, the bivariate Poisson distribution with two-way covariance structure for $\boldsymbol{Y} = (Y_1, Y_2)'$ is expressed through $q = 3$ independent Poisson variables as follows:

$$\begin{aligned} Y_1 &= X_1 + X_3 \\ Y_2 &= X_2 + X_3, \end{aligned} \tag{2.2}$$

where $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, 2, 3$. The joint PMF of Y_1 and Y_2 is given as:

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2|\boldsymbol{\lambda}) = \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\}\frac{\lambda_1^{y_1}\lambda_2^{y_2}}{y_1!y_2!} \\ \times \sum_{i=0}^{s} \binom{y_i}{i}\binom{y_2}{i} i! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i, \end{aligned} \tag{2.3}$$

where $s = \min(y_1, y_2)$. Similarly, for $J = 3$, $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$ can be written as:

$$\begin{aligned} Y_1 &= X_1 + X_4 + X_5 \\ Y_2 &= X_2 + X_4 + X_6 \\ Y_3 &= X_3 + X_5 + X_6, \end{aligned} \tag{2.4}$$

where $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, 6$, see (Karlis and Meligkotsidou, 2005). The joint PMF of Y_1, Y_2 and Y_3 is given as:

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3|\boldsymbol{\lambda}) = \exp\{-\sum_{i=1}^{6}\lambda_i\}\frac{\lambda_3^{y_3-X_5-X_6}\lambda_4^{X_4}\lambda_5^{X_5}\lambda_6^{X_6}}{(y_3-X_5-X_6)!X_4!X_5!X_6!} \times \sum_{(X_4,X_5,X_6)\in D}\frac{\lambda_1^{y_1-X_4-X_5}\lambda_2^{y_2-X_4-X_6}}{(y_1-X_4-X_5)!(y_2-X_4-X_6)!} \tag{2.5}$$

where the summation is over the set D such that $D = [(X_4, X_5, X_6) \in \mathbb{N} : (X_4 + X_5 \leq y_1) \cap (X_4 + X_6 \leq y_2) \cap (X_5 + X_6 \leq y_3)] \neq \emptyset$

Respectively, matrix $\boldsymbol{A}$ has the forms for $J = 2$ and $J = 3$ as follows:

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix} \tag{2.6}$$

The mean vector is given by $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\lambda}) = \boldsymbol{A}\boldsymbol{\lambda}$ and the variance-covariance matrix is given by $\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{\lambda}) = \boldsymbol{A}\sum\boldsymbol{A}'$, where $\sum = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. According to Karlis and Meligkotsidou (2005), this covariance cannot accommodate negative associations among the components of response vector, $\boldsymbol{Y}$. Note that when $J = 1$, the multivariate Poisson, $\mathrm{MP}_J(\boldsymbol{y}|\boldsymbol{\lambda})$ in Equation 2.1 reduces to a univariate Poisson PMF $P(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$.

2.2.2 Markov Chain Monte Carlo sampling algorithms

The Markov Chain Monte Carlo (MCMC) is a large class of methods that enables inference in high dimensional problems with unknown quantities and is able to handle complicated distributions. A detailed and broad overview of the MCMC techniques in Bayesian computation can be found in Gamerman and Lopes (2006); Robert et al. (1999); Chen et al. (2012). These methods are commonly used for

computing posterior quantities such as means, standard deviations, medians, credible intervals and quantiles from the known, simulated and/or approximated posterior distributions. The Gibbs sampler is the most popular and basic technique used (Geman and Geman, 1984). The technique is usually used when the joint distribution is fairly complex, and the conditional distributions are relatively simple. The Metropolis-Hastings (MH) algorithm by Metropolis et al. (1953); Hastings (1970) is another method used when the form of conditional distribution is not available as a known density.

Bayesian computation of the MCMC uses the Gibbs sampler as its sampling algorithm. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ denotes a vector of all parameters associated with a model and suppose $\boldsymbol{y}$ is the observed data. Let $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ be the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. Therefore, the Gibbs sampling algorithm is given as:

- **Step 0**: Choose an arbitrary starting point. $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_p^0)'$, and set $i = 0$;
- **Step 1**: Generate the next value . $\boldsymbol{\theta}^{i+1} = (\theta_1^{i+1}, \ldots, \theta_p^{i+1})'$, and set $i = 0$;
  - Generate $\theta_1^{i+1} \sim \pi(\theta_1^{i+1}|\theta_2^i, \ldots, \theta_p^i, \boldsymbol{y})$;
  - Generate $\theta_2^{i+1} \sim \pi(\theta_2^{i+1}|\theta_1^{i+1}, \theta_3^i, \ldots, \theta_p^i, \boldsymbol{y})$;
  - $\vdots$
  - Generate $\theta_p^{i+1} \sim \pi(\theta_p^{i+1}|\theta_1^{i+1}, \ldots, \theta_{p-1}^{i+1}, \boldsymbol{y})$;
- **Step 2**: Set $i = i + 1$ and repeat **Step 1**.

The algorithm will repeat until the chain converges, thereafter, the value $\boldsymbol{\theta}^i$ will be drawn from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. For a high number of iterations, the algorithm is expected to converge to the equilibrium. The approach requires the conditional distributions $\theta_k \sim \pi(\boldsymbol{\theta}|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_p, \boldsymbol{y})$ to be readily available and the analytical forms to be known. If this is not the case, then the

MH algorithms can be used. The MH algorithm has two parts; that is, a proposal and an acceptance of the proposal. Suppose that $q(\boldsymbol{\theta}|\boldsymbol{y})$ is a proposal density and $\mathrm{U}(0,1)$ is the uniform distribution defined on the interval $(0,1)$. The general form of MH sampling algorithm can be written as follows:

- **Step 0**: Choose an arbitrary starting point $\boldsymbol{\theta}^0$, and set $i = 0$;
- **Step 1**: Generate a candidate starting point $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^i, .)$ and u from $\mathrm{U}(0,1)$;
- **Step 2: Set** $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^*$ if $u \leq a(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i$ otherwise, the acceptance probability is given by:

$$a(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = \min\left(\frac{\pi(\boldsymbol{\theta}^*|D)q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i|D)q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)}\right) \tag{2.7}$$

- **Step 3**: Set $i = i + 1$ and go to **Step 1**.

Tierney (1994) stated that the general MH algorithm can be reduced to an independent chain Metropolis algorithm if $q(\boldsymbol{\theta}, \boldsymbol{\alpha}) = q(\boldsymbol{\alpha})$. When $q(\boldsymbol{\theta}, \boldsymbol{\alpha}) = q(\boldsymbol{\alpha} - \boldsymbol{\theta})$, where $q(\cdot)$ is the multivariate density and the candidate $\boldsymbol{\theta}^*$ in Step 2 is drawn such that $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{\omega}$ (where $\boldsymbol{\omega}$ is the increment random variable following the distribution $q(\cdot)$), then the procedure is referred to as a random walk chain, see Chib and Greenberg (1995).

2.2.3 Approximate Bayesian inference using INLA

The MCMC framework is useful in case where no analytical form of the posterior distribution is available. The generalized dynamic models are a complex class of models in terms of the dependence between different effects and states. It is worth noting that the sources for dependence arise from the time evolution of the state equation and the possible relationships within components of the response vector. For this reason, the MCMC method tends to be slow in convergence of the sam-

pling scheme for such complicated models. As a result, performance improvement of this method was done by Gamerman and Lopes (1997); Knorr-Held and Rue (2002); Fruehwirth-Schnatter and Frühwirth (2007); Rue and Held (2005). Nevertheless, the MCMC method algorithms remain cumbersome and time consuming in terms of speed and accuracy. Rue et al. (2009) proposed the Integrated Nested Laplace Approximations (INLA) to provide fast approximate Bayesian inference. This method provides accurate approximations to the posterior distributions of the parameters, and, since it does not rely on multiple sampling schemes, these approximations reduce computational time compared to the MCMC inference. Further information can be found in (Rue et al., 2009).

Usually, the approach is discussed relative to the latent Gaussian models or SAR (structured additive regression) models (Fahrmeir and Lang, 2001). This setup assumes that the response variable y_t belongs to the exponential family and that it is observed over time. The mean μ_t is linked to the structural additive predictor η_t through a link function, $\log(\mu_t) = \eta_t$. The dynamic model predictor can be written as follows:

$$\eta_t = \alpha + \gamma_t + \boldsymbol{z}_t'\boldsymbol{\beta}, \tag{2.8}$$

where α is the intercept, γ_t introduces a temporal dependence in the model, and $\boldsymbol{\beta}$ is the linear coefficient vector of covariates $\boldsymbol{z}$. Suppose $\boldsymbol{x}$ is the vector of all latent Gaussian variables, and $\boldsymbol{\theta}$ vector of hyperparameters is associated with a model. According to Rue and Martino (2007) the latent field $\boldsymbol{x}$ is assumed to have conditional independence properties and the number of hyperparameters is relatively small (≤ 6).

The marginal posterior distribution can be written as

$$\pi(x_i|\boldsymbol{y}) = \int_{\theta} \pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{2.9}$$

where x_i denotes each component of the latent Gaussian $\boldsymbol{x}$, $\boldsymbol{y}$ is a vector of observed data, and $\boldsymbol{\theta}$ is the vector of hyperparameters. In the form of the hierarchical structure of the joint distribution, $\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{y})$. Therefore, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can be approximated by the Laplace approximation of a marginal posterior distribution.

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} \right|_{x=x^*(\theta)} \tag{2.10}$$

where $\boldsymbol{x}^*$ is the mode of the full conditional $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. The denominator $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ denotes the Gaussian approximation to $\pi_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ (Rue and Held, 2005). In order to integrate out $\boldsymbol{\theta}$, a good set of evaluation points θ_k for numerical integration in Equation 2.9 must be found through exploring the properties of Equation 2.10. That is, an iterative algorithm with appropriate choice of weights Δ_k, which are assigned to each θ_k must be initiated (Rue et al., 2009). According to Rue and Martino (2007) and Rue et al. (2009), there are three alternatives that can be used to approximate $\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})$. A simplified Laplace approximation, a Gaussian approximation, and a Laplace approximation. Rue and Held (2005) stated that the non-normal distribution under this alternative is approximated with a Gaussian density matching the mode and the curvature at the mode. This method produces results that are reasonable, though the approximation may be improved by applying the simplified Laplace or the Laplace approximation to $\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})$. Briefly, the numerical integration of

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{y}, \theta_k)\tilde{\pi}(\theta_k|\boldsymbol{y})\Delta_k. \tag{2.11}$$

can be used to obtain an approximation of the posterior marginal density in Equation 2.9. The integration points θ_k can be chosen using the central composite design or the grid strategy. A detailed description can be found in Rue and Martino (2007). As a result, the approximate posterior quantities can be obtained and used as posterior summaries for the parameters of interest.

Chapter 3

Exploratory data analysis

3.1 Data description

The data were collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, in Virginia, United States (US). Standard flagging techniques (CDC, 2020) were used along established transects and identified species and life-stage according to (Sonenshine, 1979). Ticks life stages count data were recorded for each different location, habitat and habitat type through the four seasons. Data collection was conducted from independent cities, namely; City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County. For brevity, all will be referred to as counties. Data collection was done at least once a month on varying days of the week and at 12 different sites in southeast Virginia from May 2009 through December 2018 (Lephoto et al., 2021). The data were collected at random from multiple areas referred to as habitats, and each area was designated by a unique number ranging from 1 to 5 for different habitats. The habitat type was used to designate the kind of area (woods, edges or grass) where the data were collected. The number of the week (from 1 to 53)

was also recorded during data collection, with week 1 being the first week of the year. While fewer ticks were observed in winter, the adult stage of *I. scapularis* is active in winter. To help align the information from year to year, data collected during the last week of December were recorded as week 53, which was also the first week of January of the following year. Track records of the month and year were also kept during data collection.

3.2 Exploratory results

Table 3.1 shows environmental and time predictor variables used in this study.

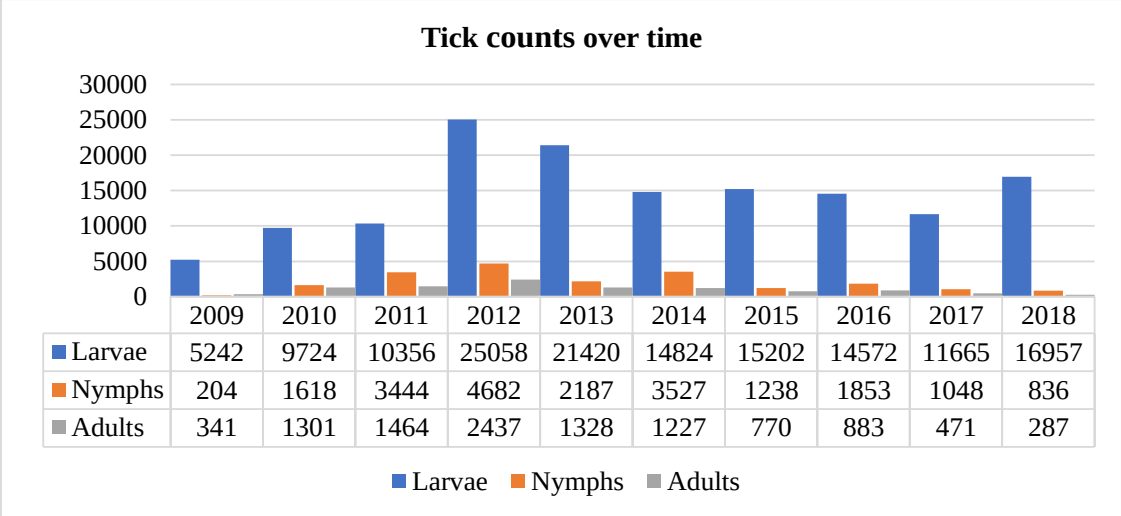
Table 3.1: Predictor variables and their type.

Predictor	Type
Habitat	Environmental
Habitat type	Environmental
Location(sample site)	Environmental
Year	Time
Month	Time
Season	Time

Table 3.2 shows the number of positive and zero counts in the data. It is obvious that the data have more zero counts than positive counts for all tick life stages. Our records show that the larvae were more abundant in the environment than the nymph and adult stage ticks at all times, see Figure 3.1. The blue bars for larvae are on average, approximately 7 and 14 times taller when compared with the orange and grey bars for the nymph and adults, respectively, during the study period from 2009 through 2018.

Table 3.2: Tick counts in the data, 2009-2018.

Variable	Positive counts	Nil counts
Larvae	726	3 767
Nymphs	1 770	2 723
Adults	1 365	3 128



	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Larvae	5242	9724	10356	25058	21420	14824	15202	14572	11665	16957
Nymphs	204	1618	3444	4682	2187	3527	1238	1853	1048	836
Adults	341	1301	1464	2437	1328	1227	770	883	471	287

Figure 3.1: Tick counts of larvae, nymph and adult species in Virginia, 2009-2018.

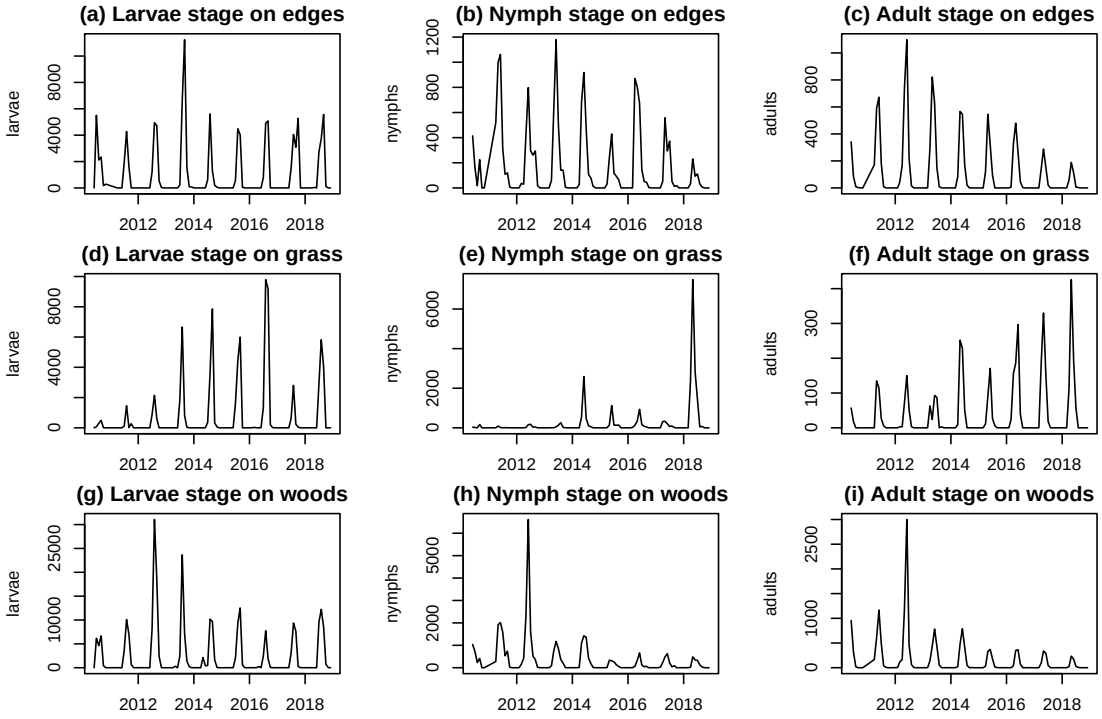


Figure 3.2: Plots showing time series of tick life stage counts in each habitat type, 2009 through 2018.

Figure 3.2 depicts the monthly tick life stage counts by habitat type from June 2010 through December 2018. The data shows strong seasonality throughout the years, indicated by peaks during warmer seasons. The plots show that larvae counts were higher in all habitat types relative to nymph and adult counts. The woods habitat type shows that higher tick counts were recorded in 2012, with the larvae counts reaching approximately 30 000. However, we observe a decline in the abundances for all life stages over the study period. The grass habitat type had lower nymphal tick numbers from 2010 through 2014, after which there were considerable peaks during summer seasons. We also observe a steady increase over the study period in all stages. In the edges habitat type we observe a decline over time for nymph and adult abundances, while larvae abundances were constant throughout the study period.

Table 3.3 present summary statistics of tick life stage counts by habitat types. It is clear from the table that the mean is less than the standard deviation for all stages in each habitat type. The inequality in this case suggest presence of overdispersion in the data and, as a result, the assumption of equal mean and variance in the Poisson distribution is violated. The assumption can be relaxed by considering the negative binomial distribution among other count distribution models.

Table 3.3: Summary statistics of the data by habitat type, 2009-2018.

Habitat type	Tick stage	N	Mean	St. Dev.	Min	Max
Edge	Larvae	108	328.94	661.33	0	3,746
	Nymphs	108	53.33	89.33	0	393
	Adults	108	35.80	69.27	0	366
Grass	Larvae	111	235.86	630.19	0	3,265
	Nymphs	111	28.06	96.77	0	863
	Adults	111	8.36	18.42	0	99
Woods	Larvae	106	785.98	1,682.64	0	10,355
	Nymphs	106	110.96	248.61	0	2,070
	Adults	106	53.92	122.93	0	967

Chapter 4

Multilevel dynamic multivariate model applied to tick life-stage count data

4.1 Introduction

Increasing tick abundance and tick-borne pathogens constitute a growing threat to public health (Laaksonen et al., 2017; Paull et al., 2022; Wimms et al., 2023). Nadolny et al. (2014) recorded more than 66 000 questing tick populations in the southeastern Virginia, comprising 7 species from a variety of habitats, with *Amblyomma americanum* (Lone Star tick) constituting over 95% of the ticks collected. Our research is focused on modelling *A. americanum* tick life-stage abundance collected over time from different habitat types. The objective of this study is to develop a multilevel model for tick count data collected over time for tick life stages, namely: larvae, nymph and adult stages.

Models in literature for tick modelling involve GLMs, maximum entropy (MaxEnt), classification and regression tree (CART), species distribution modelling

(SDM), ecological niche factor analysis (ENFA), and Bayesian hierarchical models among many others. Generalized linear models have been employed to investigate tick aggregation at different spatial ranges, provide robust estimates of tick densities between landscapes, and to determine and disaggregate the drivers of tick density and presence probabilities in an area (De Clercq et al., 2015; Kaizer et al., 2015; Estrada-Peña et al., 2016; Requena-García et al., 2017; Sagurova et al., 2019; Fan et al., 2023). Maximum entropy methods have been used to explore the limits of potential distribution by extrapolating the environmental requirements of ticks. These methods were also utilised to analyse the possible spatial range of tick species, and to explore how climate change can shape their distribution (de Oliveira et al., 2017; Raghavan et al., 2019; Ma et al., 2021; Lippi et al., 2021; Rochlin et al., 2023). Studies using SDM techniques lead to precise boundaries of the range of tick presence based on computational map modelling and demonstrate the way in which local populations of ticks differ in abundance towards the boundaries of the range (Qviller et al., 2016; Estrada-Peña et al., 2016; Lado et al., 2020; Kopsco et al., 2022; Rochlin et al., 2022, 2023). The CART technique has been used to review data on tick distribution and prevalence of tick-borne disease (TBD) for a national TBD management approach using the ecological and epidemiological information of ticks and the related diseases they transmit (Lynen et al., 2008; De Clercq et al., 2015; Kjær et al., 2019). The ENFA technique has been used to measure the extent to which the preferences of a given species deviate from average conditions and the extent to which the species is selective over the range of environmental conditions available in a country (Estrada-Pena and Venzal, 2007; Lynen et al., 2008). Bayesian hierarchical models have been used in modelling prevalence of tick-borne pathogens and spatio-temporal distributions (Wimberly et al., 2008; Simpson et al., 2019; Lephoto et al., 2021; Clark and Wells, 2023).

Bayesian methods used in modelling tick abundances or related tick-borne dis-

eases have employed the Markov Chain Monte Carlo (MCMC) sampling technique (Zipkin, 2012; Liu et al., 2017; Neupane et al., 2021; Clark and Wells, 2023; Wang et al., 2023). The Integrated Nested Laplace Approximation (INLA) technique can be used to conduct Bayesian inference as an alternative method to MCMC. The INLA technique approximates the posterior distribution based on the Laplace formulation, then it draws from the approximated posterior distribution to estimate the desired quantities and turns out to be more computationally efficient when compared to the MCMC approach (Arab, 2015; Ferkingstad et al., 2017; Khana et al., 2018; Van Niekerk et al., 2023).

We extend the application of a multilevel multivariate model proposed by Serhiyenko et al. (2018), also called the level correlated model (LCM), which they used to study temporal patterns in prescription drug counts for competing drugs in therapeutics. The model allows the flexibility of modelling multivariate hierarchical data at different cluster levels and employs the INLA algorithm for fast and accurate approximate Bayesian posterior distributions. The model is appealing in our case because it helps relax the assumption of positive associations between the components of the response vector (which is an assumption made when dealing with hierarchical life stages of the life cycle of a biological population) to include negative associations through the variance-covariance matrix, where components of the response vector are modeled jointly using univariate distributions and the components are correlated at the location level. This study focuses on models for multivariate time series of tick life stages count data collected at different habitat types to investigate covariate effects and determine dependence between the life stages. Univariate regression models cannot account for the dependence between tick life stages (Aitchison and Ho, 1989; Chib and Winkelmann, 2001; Ma et al., 2008). Since the dependence between tick life-stages may be due to biological processes, where eggs hatch to emerging larvae, larvae to nymphs and nymphs to adults, it is clear that numbers in a later stage will depend on previous

stages in a life cycle, and correlations can occur due to factors such as habitat type, sample sites and other factors which may simultaneously affect life stages. Multilevel models are used instead of the univariate regression models to account for the dependence between tick life stages. Our search did not find similar methods applied to ecological data, particularly tick life-stage count data. Our aim is to understand the causes and consequences of variation in the abundance as a function of environmental and temporal covariates.

Section 4.2 gives a detailed description of the structure of the multilevel model and its framework in modelling multivariate time series data. A brief description of model diagnostics is given in Section 4.3. Section 4.4 details the multilevel model simulation process and results. Section 4.5 describes the application of the model, where the first subsection gives a description of the tick life-stage data, then a detailed description of multilevel Poisson and zero-inflated Poisson model fitting, model application to data, model diagnostics, and lastly presents and interprets model results. The conclusion is presented in the last section.

4.2 Model framework

Let $\boldsymbol{Y}_{it} = (Y_{1,it}, \ldots, Y_{J,it})'$ be a $J-$dimensional vector of count responses observed over T regularly spaced time points and n locations, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The assumption of the Poisson or zero inflated Poisson (ZIP) marginal distributions for the multilevel time series model corresponds to

$$Y_{j,it}|\delta_j, \lambda_{j,it} \sim \begin{cases} \text{Poi}(\lambda_{j,it}); \text{ when } \delta_j = 1; \text{ or} \\ \text{ZIP}(\lambda_{j,it}, \delta_j); \text{ otherwise,} \end{cases} \tag{4.1}$$

and to include the effect for covariates we use the log-link regression model given by

$$\log(\lambda_{j,it}) = \beta_{j,i0} + \gamma_{j,t} + \boldsymbol{z}'_{j,it}\boldsymbol{\beta}_j + \xi_{j,it}, \tag{4.2}$$

where $\lambda_{j,it}$ and δ_j are the mean count and dispersion parameters for the jth component, respectively. $\beta_{j,i0}$ is a location-specific intercept for the jth component of the response vector, $\boldsymbol{z}'_{j,it}$ denotes a p_j-dimensional vector of predictors, $\boldsymbol{\beta}_j$ is a p_j-dimensional vector of coefficients corresponding to the predictors, and $\xi_{j,it}$ is the random effects term explaining the dependence between the components of the response vector through the variance covariance matrix $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4.3}$$

where $\mathbf{0}$ is a $J-$dimensional zero vector, $\boldsymbol{\Sigma}$ is a variance-covariance matrix with σ_{rs} elements, such that $1 \leq r \leq s \leq J$, and under Bayesian estimation will be part of hyperparameters. The dependence between components of the response vector can be postulated either though $\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ or $\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. Postulating correlations through the matrix $\boldsymbol{\Sigma}$ means that the correlations between the J tick-life cycle developmental stages (which are the same as the response vector components) are location independent whereas $\boldsymbol{\Sigma}_i$ is a location-specific variance covariance matrix. Consider $1 \leq r \leq s \leq J$, and suppose $r \neq s$, then the off-diagonal elements of $\boldsymbol{\Sigma}$ are given as follows:

$$\sigma_{rs} = \rho_{rs}\sqrt{\sigma_{rr}\sigma_{ss}},$$

where σ_{rr} and σ_{ss} are the marginal variances and ρ_{rs} is the correlation coefficient in the matrix. To introduce location specificity we need to introduce an index i in the equation above. The parameter, $\gamma_{j,t}$, is the jth component specific dynamic

time effect modeled as a random walk such that:

$$\gamma_{j,t} = \gamma_{j,(t-1)} + \omega_{j,t}, \tag{4.4}$$

where $\omega_{j,t} \sim \mathrm{N}(0, 1/W_j)$ denotes the error term for component j of the response vector assumed to follow the normal distribution. W is the precision parameter based on the inverse-gamma distribution, and the LogGamma prior is specified for $\log(W_j) \sim \mathrm{LogGamma}(a, b)$ in the application, where $a = 1$ and $b = 0.00005$. The time effect in Equation 4.4 can be modified such that:

$$\gamma_t = \gamma_{(t-1)} + \omega_t, \tag{4.5}$$

which means that the model in Equation 4.2 is a model where the time effect is not stage (or response component) specific, and $\omega_t \sim \mathrm{N}(0, 1/W)$. This implies a model assuming shared time effects across all locations and all components of the response vector. The second case is

$$\gamma_{it} = \gamma_{i(t-1)} + \omega_{it}, \tag{4.6}$$

and this implies that Equation 4.2 is a model where the time effect is location specific and life-cycle stage (or response component) independent, and $\omega_{it} \sim \mathrm{N}(0, 1/W_i)$ (an assumption where the errors share the same distribution over all components of the response vector at a given location i). Thus, the model in Equation 4.2 modifies into a model assuming shared time effects across all components of the response vector. Lastly, Equation 4.4 can be modified to

$$\gamma_{j,it} = \gamma_{j,i(t-1)} + \omega_{j,it}, \tag{4.7}$$

which is a model where the time effect is assumed to have separate time evolution on each of the components of the response vector and each location, where $\omega_{j,it} \sim \mathrm{N}(0, 1/W_i)$ or $\omega_{j,it} \sim \mathrm{N}(0, 1/W_j)$ (an assumption that state errors share the same distribution for all components of the response vector or that state errors follow response component specific normal distributions). Thus, the model in Equation 4.2 modifies into a model assuming different time effects across all locations and all components of the response vector. The dispersion parameter $\delta_j = 1$ in Equation 4.1 implies the assumption of the Poisson distribution for the jth component, and $\delta_j \neq 1$ implies the ZIP distribution for the jth component. The variance-covariance matrix, $\boldsymbol{\Sigma}$, plays a crucial role in understanding the dependence between the components of the vector of counts where each component of the response vector follows a univariate distribution. Consequently, the model in Equations 4.1 to 4.3 accounts for overdispersion through the use of the ZIP model and in this case the diagonal terms of the matrix $\boldsymbol{\Sigma}$; for $r = s$ in $\boldsymbol{\Sigma}_{1 \leq r \leq s \leq J}$, $\sigma_{jj} > 0$, so that $\mathrm{Var}[Y_{j,it}] > \mathrm{E}[Y_{j,it}]$. Note that

$$\mathrm{E}[Y_{j,it}] = \exp(z'_{j,it}\beta_j \exp(\sigma_{jj}/2) = m_{j,it},$$

$$\mathrm{Var}[Y_{j,it}] = m_{j,it} + m^2_{j,it}(\exp(\sigma_{jj}) - 1),$$

$$\mathrm{Cov}[Y_{r,it}, Y_{s,it}] = m_{r,it}m_{s,it}(\exp(\sigma_{rs}) - 1).$$

Moreover, the model can allow for the positive or negative dependence between the components of the response vector (Chib and Winkelmann, 2001). This follows from the sign of the covariance in Equation 4.3 which depends directly on the value of σ_{rs}, and a negative value of σ_{rs} yields a negative value of the association,

while a positive value results in a positive association between the components Y_{ri} and Y_{si}.

4.3 Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \bar{D}(\theta) + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter et al. (2002) state that p_D values less than zero indicate substantial conflict between the prior distribution and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of deviance suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter et al. (2002).

4.4 Simulation study

Suppose $\boldsymbol{Y}_i = (Y_{1i}, Y_{2i})$ represent a bivariate ($J = 2$) count data vector that is observed at $i = 1, \ldots, n$ locations when there is no over-dispersion in the data. The data is simulated according to the simplified model described in Equations 4.8 and 4.9, and does not assume any temporal dependence. That is, the marginal Poisson multilevel model is given by:

$$Y_{ji}|\lambda_{ji} \sim \text{Poi}(\lambda_{ji}), \tag{4.8}$$

$$\log(\lambda_{ji}) = \beta_{j,0} + \xi_{ji}, \tag{4.9}$$

where $\beta_{j,0}$ and $\xi_{j,i}$ are the intercept and random effects parameters for the jth component, respectively. The bivariate multilevel model data simulation comprises $\boldsymbol{\beta}_0 = (\beta_{1,0}, \beta_{2,0})'$ and $\boldsymbol{\xi}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with σ_{11}, σ_{22} and ρ_{12}, the elements of $\boldsymbol{\Sigma}$. We assume that the true parameter values were $\sigma_{11} = \sigma_{22} = 0.4$, $\rho_{12} = -0.95$, and $\beta_{1,0} = \beta_{2,0} = 10$ and run the simulation for $n = 100$, $n = 500$, $n = 5\ 000$ and $n = 10\ 000$ using the R-INLA package in R software proposed by Rue et al. (2009) and fitted the four different bivariate multilevel model. The bivariate multilevel model is formulated through joint modelling of the four models below.

Model 1:

$$Y_{1i}|\lambda_{1i} \sim \mathrm{Poi}(\lambda_{1i}); Y_{2i}|\lambda_{2i} \sim \mathrm{Poi}(\lambda_{2i}). \tag{4.10}$$

Model 2:

$$Y_{1i}|\lambda_{1i}, \delta_1 \sim \mathrm{ZIP}(\lambda_{1i}, \delta_1); Y_{2i}|\lambda_{2i}, \delta_2 \sim \mathrm{ZIP}(\lambda_{2i}, \delta_2). \tag{4.11}$$

Model 3:

$$Y_{1i}|\lambda_{1i}, \delta_1 \sim \mathrm{ZIP}(\lambda_{1i}, \delta_1); Y_{2i}|\lambda_{2i} \sim \mathrm{Poi}(\lambda_{2i}). \tag{4.12}$$

Model 4:

$$Y_{1i}|\lambda_{1i} \sim \mathrm{Poi}(\lambda_{1i}); Y_{2i}|\lambda_{2i}, \delta_2 \sim \mathrm{ZIP}(\lambda_{2i}, \delta_2). \tag{4.13}$$

λ_{ji} is modeled as a function of ξ_{ji} and $\beta_{j,0}$ as in Equation 4.9. We assume a Normal prior for $\beta_{j,0}$ and the matrix $\boldsymbol{\Sigma}$ has a Wishart prior with $r = 2 \times J + 1$ degrees of freedom and identity matrix as a prior precision matrix. We also use default hyperparameter specifications in the `inla()` function from the R-INLA package. In particular we assume $\beta_{j,0} \sim \mathrm{N}(0, 10^3)$ and the precision matrix $\boldsymbol{\Sigma}^{-1} \sim W(r, \boldsymbol{\Sigma}^{-1})$. For computational time comparison, a fully Bayesian inference model using Markov Chain Monte Carlo (MCMC) was fitted using the `MCMCglmm()`function in the MCMCglmm R package (Hadfield, 2010). Here, we assume the default Normal prior for $\beta_{j,0}$, with a zero mean vector and a diagonal

variance matrix with large variances $N(0, 1 \times 10^{10})$ (Hadfield, 2010). We assumed a default inverse Wishart prior for each component in the covariance matrix of the random effects.

Estimated parameters, CPU (Central Processing Unit) times between INLA and fully Bayesian inference using MCMC, and estimated averages of posterior means are given in the subsections that follow.

4.4.1 Estimation using INLA

The CPU time comparisons between a fully Bayesian inference using MCMC and an approximate Bayesian inference using INLA are presented in Table 4.1. The MCMC sampling was run with 105 000 iterations, where 5 000 iterations were used for burn-in, and 100 iterations were for thinning. As a result, only 1 000 posterior samples were retained from which posterior summaries were calculated. It can be seen from Table 4.1 that the computational times quickly increase with the increase in sample size for the MCMC runs, and that INLA runs faster than the MCMC, as expected in all cases. The INLA technique is approximately 20 times faster for Poisson distributions, 6 times faster for ZIP, and between 10 and 17 times faster for Poisson and ZIP mixed distributions of counts when n is large ($\geq$ 5 000) compared to the MCMC, and its posterior estimates are closer to those obtained using the MCMC (Table 4.2). Note that the speed under M2 - M4 is approximately double the speed under M1. This indicates that the distributions involving the inclusion of ZIP models are expected to run twice as slow as models including Poisson distributions.

For brevity, 500 replications were run for n = 5 000, and M1 model parameters were estimated using MCMC and INLA. The model averages of posterior means and MSEs (mean square errors) of 500 replicates are provided in Table 4.3. It is clear from the table that INLA average posterior estimates are reasonably close

Table 4.1: CPU time of model estimation using INLA and MCMC (seconds)

n	INLA				MCMC
	(M1)	(M2)	(M3)	(M4)	
100	1	2	2	2	11
500	2	4	3	3	48
1 000	5	10	7	7	95
5 000	21	80	27	49	467

Table 4.2: Estimated parameters in simulated models.

n	Model	M1			M2			M3			M4			True
		Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI		
	parameter		2.5%	97.5%		2.5%	97.5%		2.5%	97.5%		2.5%	97.5%	value
100	σ_{11}	0.48	0.34	0.69	0.47	0.34	0.69	0.48	0.34	0.70	0.47	0.34	0.69	0.40
	σ_{22}	0.43	0.31	0.62	0.43	0.31	0.62	0.43	0.31	0.62	0.43	0.31	0.62	0.40
	ρ_{12}	-0.86	-0.92	-0.77	-0.86	-0.92	-0.77	-0.86	-0.92	-0.77	-0.86	-0.92	-0.78	-0.95
	$\beta_{1,0}$	9.93	8.48	11.56	9.98	8.53	11.61	9.97	8.51	11.61	9.93	8.49	11.57	10.00
	$\beta_{2,0}$	9.18	7.89	10.63	9.18	7.88	10.63	9.18	7.89	10.64	9.18	7.88	10.63	10.00
500	σ_{11}	0.39	0.34	0.46	0.38	0.34	0.43	0.40	0.35	0.47	0.40	0.35	0.47	0.40
	σ_{22}	0.42	0.36	0.50	0.42	0.38	0.48	0.43	0.38	0.51	0.43	0.37	0.50	0.40
	ρ_{12}	-0.88	-0.92	-0.84	-0.88	-0.90	-0.85	-0.89	-0.92	-0.84	-0.89	-0.92	-0.86	-0.95
	$\beta_{1,0}$	9.72	9.12	10.35	9.73	9.13	10.36	9.72	9.12	10.35	9.72	9.12	10.35	10.00
	$\beta_{2,0}$	9.75	9.13	10.40	9.74	9.12	10.39	9.76	9.14	10.41	9.75	9.12	10.40	10.00
5 000	σ_{11}	0.42	0.37	0.47	0.40	0.37	0.43	0.41	0.37	0.46	0.42	0.39	0.47	0.40
	σ_{22}	0.42	0.38	0.48	0.42	0.38	0.45	0.43	0.38	0.48	0.42	0.38	0.47	0.40
	ρ_{12}	-0.90	-0.93	-0.87	-0.91	-0.93	-0.90	-0.90	-0.93	-0.87	-0.90	-0.93	-0.88	-0.95
	$\beta_{1,0}$	9.70	9.26	10.15	9.71	9.28	10.16	9.72	9.29	10.17	9.70	9.26	10.15	10.00
	$\beta_{2,0}$	10.09	9.63	10.56	10.11	9.66	10.57	10.08	9.63	10.55	10.11	9.66	10.58	10.00
10 000	σ_{11}	0.40	0.38	0.42	0.41	0.39	0.42	0.40	0.38	0.42	0.40	0.38	0.42	0.40
	σ_{22}	0.40	0.38	0.42	0.40	0.38	0.42	0.40	0.38	0.43	0.40	0.38	0.42	0.40
	ρ_{12}	-0.94	-0.95	-0.93	-0.95	-0.95	-0.94	-0.95	-0.96	-0.93	-0.94	-0.95	-0.93	-0.95
	$\beta_{1,0}$	9.86	9.67	10.06	9.87	9.67	10.07	9.87	9.67	10.06	9.85	9.65	10.05	10.00
	$\beta_{2,0}$	10.05	9.85	10.25	10.05	9.85	10.25	10.05	9.85	10.26	10.06	9.86	10.26	10.00

to MCMC average posterior estimates in all cases. As a result, this study will use the INLA method because of its fast computation time when compared to the MCMC as well as the accuracy of its estimates.

4.5 Tick life-stage modelling using multilevel model

In order to understand the dynamics and habitat type variation in the abundance of tick life-stages, we describe the statistical analysis pertaining to tick life-stage count data using multilevel dynamic models. In addition, the study focuses on determining factors contributing to tick life-stage abundances across different habitat types. Most of the existing research focuses on location level

Table 4.3: Averages of posterior means and MSEs of 500 replicates for $n = 5\,000$

Parameters	MCMC		INLA		True value
	Posterior mean	MSE	Posterior mean	MSE	
σ_{11}	0.401	0.0001037147	0.401	0.0001042946	0.40
σ_{22}	0.401	0.0001043847	0.401	0.0001062165	0.40
ρ_{12}	-0.947	0.0000516038	-0.941	0.0001146017	-0.95
$\beta_{1,0}$	10.002	0.0094999830	9.997	0.0094975680	10.0
$\beta_{2,0}$	9.992	0.0092616110	9.987	0.0093305050	10.0

counts for the different life stages of the tick cycle by investigating the tick aggregation at different spatial ranges, determining and disaggregating the drivers of tick density and probability of presence, and providing robust estimates of tick densities (Zannou et al., 2021). The association between tick life-stages and their locations is of interest. Furthermore, there is interest in knowing temporal effects on the distribution of counts.

4.5.1 Data description

We used data collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, Virginia, United States. Ticks were collected using standard flagging techniques (CDC, 2020) along established transects and identified species and life-stage according to (Sonenshine, 1979). Larvae-, nymph- and adult-stage count data were recorded for each location, habitat and habitat type through the four seasons of the year. Eight counties and independent cities were sampled in this state, namely; City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County. For brevity, all will be referred to as counties. Data was collected at least once a month on varying days of the week and at 12 different sites in southeast Virginia from May 2009 through December 2018. The data were collected from multiple areas referred to as habitats, and

each area was designated by a unique number ranging from 1 to 5 for different habitats. The habitat type was used to designate the kind of area (woods, edges or grass) where the data were collected. The number of the week (from 1 to 53) was also recorded during data collection. Week 1 is the first week of the year, and while fewer ticks were observed in winter, the adult stage of *I. scapularis* is active in winter. To help align the information from year to year, data collected during the last week of December were recorded as week 53, which was also the first week of January of the following year. Track records for the month and year were also kept. This study makes use of habitat types (edges, woods and grass), time in months, and three tick life stage counts for modelling. After removing unused variables from the raw dataset, we were left with 103 months out of the 116 total monthly data points. Note that the removal of certain data points was due to life stages being collected at different months during the first 13 months. The retained dataset consisted of three habitat types, within which there were three tick life stages each of length 103 (months).

4.5.2 Model framework for tick life stage data

Let $\boldsymbol{Y}_{it} = (Y_{L,it}, Y_{N,it}, Y_{A,it})'$ be a vector of count responses for tick life stage $j = L, N, A$ in habitat type $i = E, G, W$ (E = Edges, G = Grass, W = Woods) at equally spaced monthly times $t = 1, \ldots, 103$. The multivariate vector of tick life-stage responses $\boldsymbol{Y}_{it}$ is modelled using marginal Poisson and/or zero-inflated Poisson distributions as follows:

$$Y_{j,it}|\delta_j, \lambda_{j,it} \sim \begin{cases} \text{Poi}(\lambda_{j,it}); \text{ when } \delta_j = 1; \text{ or} \\ \text{ZIP}(\lambda_{j,it}, \delta_j); \text{ otherwise,} \end{cases} \tag{4.14}$$

where $\lambda_{j,it}$ is the jth tick life-stage mean count and δ_j is a dispersion parameter at stage j. The log-link is given by

$$\log(\lambda_{j,it}) = \beta_{j,it,0} + \gamma_t + \alpha_i + z'_{j,it}\beta_j + \xi_{j,it}, \tag{4.15}$$

where $\beta_{j,it,0}$ is the dynamic intercept, γ_{it} is the month time effect in the ith habitat type, $z'_{j,it}$ denotes a p_j-dimensional vector corresponding to the fixed effects of seasonal changes and sinusoidal predictors $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$ included to handle the seasonality with period 12 (in months), $\xi_{j,it}$ is the random effects term explaining the dependence between tick stages. The random vector parameter, $\boldsymbol{\xi}_{it} = (\xi_{L,it}, \xi_{N,it}, \xi_{A,it})'$, is the vector of the unobserved random variables introduced to capture the tick-stage dependence through the multivariate normal distribution as follows:

$$\boldsymbol{\xi}_{it} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4.16}$$

where $\mathbf{0}$ is a $J-$dimensional zero vector, $\boldsymbol{\Sigma}$ is the variance-covariance matrix for the random effect terms such that for $r \neq s$ where $\{r \text{ and } s \in j = 1, ..., J\}$, $\boldsymbol{\rho}_{rs} = (\rho_{LN}, \rho_{LA}, \rho_{NA})'$ denotes the tick life stage correlation vector. The month time effects vector $\boldsymbol{\gamma}_t = (\gamma_{L,t}, \gamma_{N,t}, \gamma_{A,t})$ and the dynamic intercept $\beta_{j,it,0}$ are assumed to be independent and evolve according to a random walk process in the state equation of the above multilevel model. The terms also capture the autoregressive time series nature of the data and hence can be specified as

$$\gamma_t = \gamma_{(t-1)} + \omega_t,$$

while

$$\beta_{j,it,0} = \beta_{j,i(t-1)+\varpi_{j,it}},$$

where $\omega_t \sim \mathrm{N}(0, 1/W)$ and $\varpi_{j,it} \sim \mathrm{N}(0, 1/\eta_j)$; that is, the errors are assumed to follow the normal distribution, where W and η_j are the precision parameters based on the inverse-gamma distribution. The reference level for the seasonal changes categorical predictor variable was taken to be winter since it is the season expected to have lower tick activity, and hence lower abundance. It therefore makes sense to contrast other seasons to the winter season. The type of habitat was also included in the model to assess the influence of habitat type on tick abundance and the baseline was taken to be the edge habitat type. Different multilevel mixed effects models taking the form in Equation 4.15 were fitted and compared using the DIC values. Model 1 denotes the multivariate model which assumes Poisson marginal distributions for all life stages in the response vector Y_{it}. Models 2 to 7 assume combinations of ZIP and Poisson mixed effects for tick life stages in the response vector. Model 8 denotes the multivariate ZIP model for tick life stage components in the response vector. The multivariate multilevel model is formulated by introducing the variance covariance matrix and jointly modelling the response vector assuming marginal count distributions as shown below:

Model 1:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \mathrm{Poi}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it} &\sim \mathrm{Poi}(\lambda_{N,it});\\
Y_{A,it}|\lambda_{A,it} &\sim \mathrm{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 2:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poisson}(\lambda_{N,it});\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

Model 3:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 4:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it},\delta_N);\\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poi}(\lambda_{N,it});\\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poisson}(\lambda_{A,it})
\end{aligned}
$$

Model 5:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

Model 6:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poi}(\lambda_{N,it});\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

Model 7:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it},\delta_N);\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 8:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it},\delta_N);\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

where $j = L, N$ or A, and δ_j the dispersion parameter for the jth life stage. All models were implemented through an approximate sampling based framework which provide a mechanism for Bayesian inference based on accurate approximations to the posterior distributions of the parameters. The approximate sample Bayesian framework requires the usual prior specification on the parameters. We assume a Wishart prior for $\boldsymbol{\Sigma}$, and a Log-Gamma prior for $\log(W)$ and $\log(\eta_j)$. The default hyperparameter specifications assume the default prior from the R-INLA package in R. The matrix $\boldsymbol{\Sigma}$ has a Wishart prior with $2 \times J + 1$ degrees of freedom and identity matrix as a prior precision matrix. The DIC values for the

fitted model given by Equation 4.15 are presented in Table 4.4.

Table 4.4: DIC values for the fitted models 1 to 8.

Model	$\text{MDC}_j(\theta_{j,it})$	DIC	p_{DIC}
1	Poi Poi Poi	5007.80	798.47
4	ZIP Poi Poi	7513432.10	-332850.08
7	ZIP ZIP Poi	7539545.18	-308560.64

Even though we do not compare the INLA model with the MCMC model, we report that the INLA approach took 32.3 seconds (approx half a minute) to run using R-INLA on Intel(R) Core(TM) i7-8550U CPU 3.80 GHz with 8GB of RAM.

Models 2, 3, 5, 6 and 8 crashed when fitted to the data and therefore we show the DIC values for model 1, 4 and 7 in Table 4.4. It is clear from the table that model 1 fits the data well compared to models 4 and 7 since its DIC value (5007.80) is smaller compared to the DIC values for the other models.

4.5.3 Results

Fixed effects and correlations

Table 4.5 gives the posterior means of parameters of the multivariate multilevel model. We discover that there is a significant, strong and positive association between larvae and nymphs, larvae and adults, and nymphs and adults with posterior correlations $\rho_{LN} = 0.903$, $\rho_{LA} = 0.833$ and $\rho_{NA} = 0.858$, and 95% credible intervals (CI) $(0.806, 0.950)$, $(0.7163, 0.891)$, and $(0.668, 0.860)$, respectively.

Variation in habitat type affects tick life cycle stages abundance in a similar manner. Comparing the grass habitat type with the edges and woods habitat types, the model estimated higher and significant [1.042 and 0.755; CI: $(0.949, 1.137)$ and $(0.926, 0.970)$] tick life cycle stage abundances in the edges and woods habitat types, respectively. Variation in seasons affects variation in tick abundances in a life stage specific manner during fall, but it affect the abundance in a similar

Table 4.5: Estimated coefficients for covariates.

Covariates	Mean	SD	2.5%	97.5%
Habitat type (**ref:** Edge)				
Grass	1.042	1.033	0.949	1.137
Woods	0.755	0.926	0.551	0.970
Season (**ref:** Winter)				
Spring.L	0.927	2.000	-2.997	4.865
Spring.N	4.631	1.745	1.208	8.078
Spring.A	4.613	1.745	1.190	8.059
Summer.L	4.049	2.057	0.007	8.097
Summer.N	4.441	1.927	0.648	8.232
Summer.A	3.113	1.927	-0.680	6.904
Fall.L	4.631	1.964	0.768	8.488
Fall.N	4.205	1.888	0.489	7.911
Fall.A	-0.683	1.917	-4.459	3.076
Cosine term	-0.111	0.174	-0.453	0.231
Sine term	0.206	0.165	-0.118	0.530
Correlations				
ρ_{LN}	0.903	0.037	0.806	0.950
ρ_{LA}	0.833	0.042	0.716	0.891
ρ_{NA}	0.858	0.063	0.668	0.933

manner during summer and spring seasons. In comparison to winter season as the reference category, the model yields significant and higher [4.631 and 0.613; CI: $(1.208, 8.078)$ and $(1.190, 8.059)$] log mean abundances for nymphs and adults and non-significant [0.927; CI: $(-2.997, 4.865)$] log mean abundances for larvae in the spring season. Higher and significant [4.049 and 4.441; CI: $(0.007, 8.097)$ and $(0.648, 8.232)$] and non-significant [3.113; CI: $(-0.680, 6.904)$] in the summer season relative to the winter season. Higher and significant [4.631 and 4.205; CI: (0.768,8.488) and (0.489,7.911)] log mean abundances for larvae and nymphs and non-significant [-0.683; CI: $(-4.459, 3.076)$] log mean abundances for adults during the fall season as compared to the winter season. However, seasonal variations in the data were not significantly [-0.111 and 0.206; CI: $(-0.453, 0.231)$ and $(-0.118, 0.530)$] strengthened by the cosine and sine terms in the model.

Habitat type specific non-linear month effects

Figure 4.1 shows the model estimated non-linear month effects, modelled as the random walk of order one (RW1). The model estimated similar months effects across the three habitat types. The graphs show the posterior mean (solid line) and 95% credible intervals (dotted lines). The curves show an expected gradual increase in tick life cycle stages abundances from January to June, and a steady decrease between June and August, slight increase between August and September, sharp decrease from September to November, and a gradual increase from November to December.

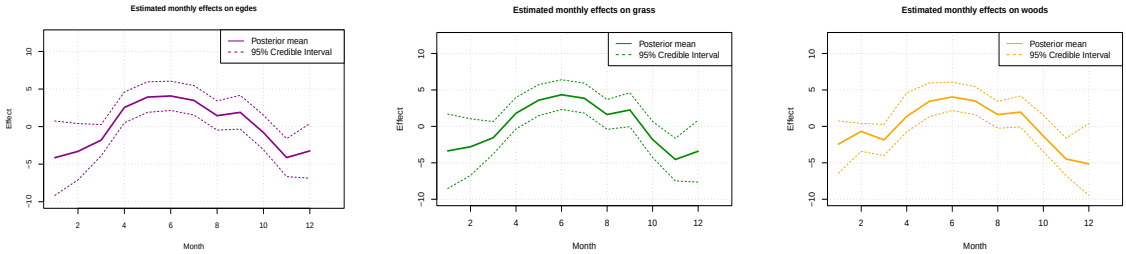


Figure 4.1: Model estimated non-linear month effects in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types.

Habitat type specific trends

Figure 4.2 shows the model estimated dynamic intercept ($\beta_{j,it,0}$), showing trends in the edges (purple curves), grass (green curves) and woods (brown curves) habitat type for larvae, nymph and adult tick life cycle stages. The fitted model poorly tracked the stochastic pattern across the three habitat types for larvae. This suggest that the larvae abundance is affected by environmental characteristics beyond those measured in this study. The top panel (row) are the estimated larvae trends showing constant fluctuations throughout the study period of 2009 through 2018 across the habitat types. The middle and bottom panels shows a declining nymph and adult tick trends since 2012 in the edges and woods habitat types, and an increasing trend in the grass habitat type.

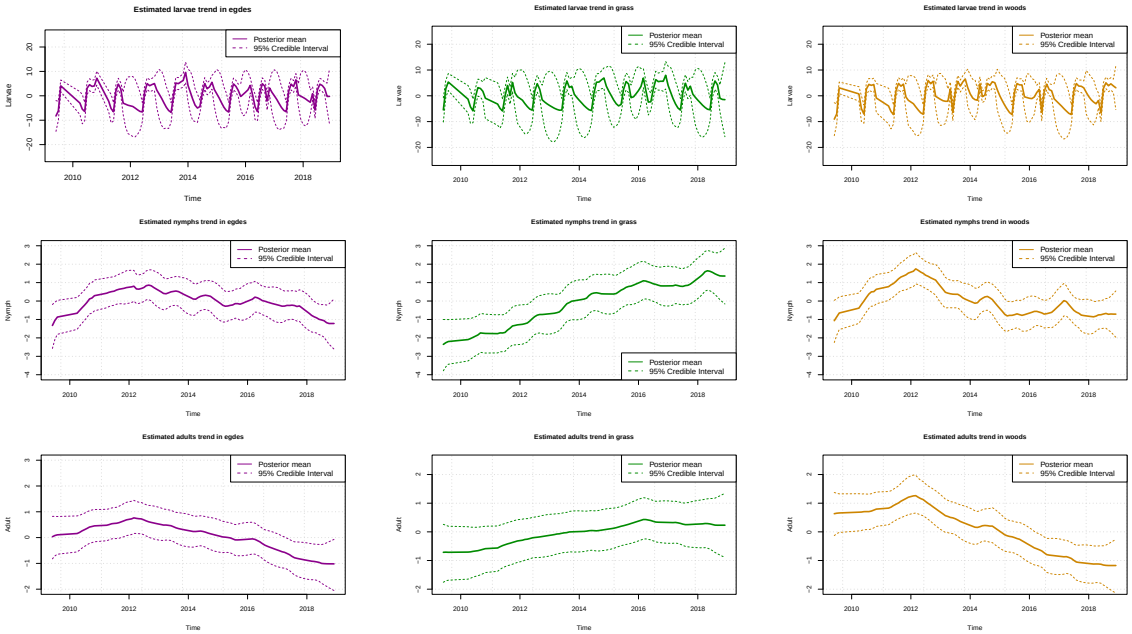


Figure 4.2: Estimated model trends in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types.

4.6 Discussion

This study described the multilevel model for the vector of time series of counts using the R-INLA technique, which is effective and computationally feasible for large datasets. The aim of this study was to investigate the influence of environmental and temporal predictors of tick life cycle stages abundances, determine trends and demonstrate the computational speed of the INLA technique proposed by Rue et al. (2009) relative to the MCMC technique. The multilevel model used in this study has seen applications in prescription counts of physicians in a large pharmaceutical company in Serhiyenko (2015); Serhiyenko et al. (2018) and ridesourcing data in NYC by Ravishanker et al. (2022) among other applications. To our knowledge, these models have not been applied to ecological data, specifically tick life stage count data. The MCMC is a large class of methods that enables inference in highly dimensional problems with unknown quantities and is able to handle complicated distributions. The literature on the methods is well documented in Gamerman and Lopes (2006); Robert et al. (1999); Chen et al.

(2012), and these methods are commonly used for computing posterior quantities through the popular Gibbs sampling methods including the Metropolis-Hastings algorithm (Geman and Geman, 1984; Metropolis et al., 1953; Hastings, 1970). However, the MCMC technique is computationally intensive when complex models are involved, or when the data is large. As an alternative to the computationally intensive MCMC Bayesian technique, the INLA technique for approximating posterior parameters was suggested and its feasibility demonstrated by simulating the data and estimating parameters, while taking into account the time taken to complete in INLA and in MCMC before models were applied to tick data. This study assumed tick life stage specific effects of the covariates on the distribution of tick stages. That is, we assumed that each life stage was affected differently from another tick life stage. However, the results in this study showed that monthly effects are similar across the three habitat types. Seasonal changes were different for each life stage and abundances were estimated to be higher in the grass and woods habitat types compared to the edges habitat type. The model estimated trends revealed an increasing trend for nymphs and adults in grassy habitat types. The results in this chapter are similar to those in Willis et al. (2012) and Lephoto et al. (2021).

4.7 Conclusion

The multilevel model successfully fitted tick life stage data and revealed that assuming common covariates effects on tick life stages may be unrealistic. However, this seems to be the case with monthly effects as similar trends were estimated. The assumption was unrealistic in the sense that factors such as sampling variation and different relationships across regions contribute to different responses to the same stimuli as one moves across regions and over a period of time. Spring and fall seasons affected life stages differently, though the effects of summer were

similar for the three tick life-stages. The model is also beneficial in allowing for a mixture of count distributions in the vector of responses.

Chapter 5

Multivariate hierarchical modelling including sample sites as random effects

5.1 Introduction

Increasing tick abundance and tick-borne pathogens constitute a growing threat to public health (Laaksonen et al., 2017; Paull et al., 2022; Wimms et al., 2023). For example in the USA, Nadolny et al. (2014) recorded more than 66 000 questing tick populations in the southeastern Virginia, comprising 7 species from a variety of habitats, with *Amblyomma americanum* (Lone Star tick) constituting over 95% of the ticks collected. Our research focused on modeling *A. americanum* tick life-stage abundance collected over time from three different habitat types namely: edges, woods and grass, along 12 random sample sites. The objective of this study was to identify abundance levels of *A. americanum* variations within sample sites from eight counties in the state of Virginia, United States. We also wanted to investigate the influence of environmental and temporal predictors and estimate

time trends.

Five major tick species were reported to be active and responsible for transmitting a variety of pathogens of both human and animals. These were *Ixodes scapularis*, *A. americanum*, *Dermacentor variabilis*, *Amblyomma maculatum*, and *Haemaphysalis longicornis* (Nadolny and Gaff, 2018). The *H. longicornis* is of high veterinary importance for transmission of *Theileria orientalis* Ikeda in cattles even though it is not a key human-biting species (Oakes et al., 2019). The most common human-biting tick in Virginia is the *A. americanum*, and it transmits several pathogens of medical and veterinary significance, including *Ehrlichia chaffeensis* (the causative agent of human monocytic ehrlichiosis), Heartland virus (HRTV) and Bourbon virus (BRBV) among others (Goddard and Varela-Stokes, 2009; Savage et al., 2013, 2017)

Due to the effects of climate change and transportation by hosts, tick ranges remain ever-changing. In recent years the *A. americanum* species has expanded its range into northeastern and midwestern United States and nothward into Canada (Molaei et al., 2019). A recent study in Kentucky reported detection of *Ehrlichia chaffeensis* in 32 counties out of the 77 sampled counties in 2019 through 2020 (Pasternak and Palli, 2023). A 2011 Centers for Disease Control and Prevention (CDC) report indicated that there were 863 human cases of ehrlichiosis in the United States. An increase has been observed over the seven-year period, with human cases steadily increasing to 1 832, reporting 1 799 as a result of *E. chaffeensis* infection (Adams et al., 2013; CDC, 2021).

A. americanum ticks have long been identified as nuisance biters because of their aggressive questing behaviour (Hair and Howell, 1970; Merten et al., 2000; Otálora-Luna et al., 2022). Despite their prominence as a nuisance biters and vector of human pathogens, efforts to define the species' variations within counties remain limited (Benham et al., 2021). Therefore, one of the strategies to help re-

duce tick-borne diseases is to have an in-depth analysis of the life stages abundance patterns and factors contributing in each of the randomly sampled locations. The aim of this chapter was to unmask the heterogeneity within southeastern counties in Virginia. Thus, we modelled tick life stage abundance variations within the randomly sampled locations. The multivariate time series of tick counts was modelled as a function of season, habitat type, non-linear effects of month and year, and sampling site random effects.

In the next Section 5.2 we describe the data for this study, give a modelling framework in Section 5.3, and simulate the data in Section 5.5 . We describe model diagnostics in Section 5.6, give model results in Section 5.7, and conclude in Section 5.9.

5.2 Data description

This chapter used data collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, Virginia, United States. Data collection was done at least once a month on varying days of the week in southeastern Virginia from May 2009 through December 2018. Ticks were collected using standard flagging techniques (CDC, 2020) along established transects and species and life-stages were identified according to (Sonenshine, 1979). Larvae-, nymph- and adult-stage count data were recorded for each edge, wooded and grassy habitat type in York, Chesapeake, Norfolk, Hampton, Isle of Wight, Virginia Beach, Northampton and Portsmouth counties through the four seasons of the year. The datafile included sample sites/locations which designate the area where the data were collected. Sample sites were abbreviated JC and ST (Jacobson Track and Stephens Tract) in Chesapeake, LA (Langley) in Hampton, BW (Blackwater Ecological Preserve) in Isle of Wight, WS (Weyanoke) in Norfolk, KP (Kiptopeke State Park) in Northampton, HC and PC (Hoffler Creek Wildlife

Preserve and Paradise Creek Nature Park) in Portsmouth, BB and OD (Back Bay NWR and Oceana/Dam Neck) in Virginia Beach, and CA and NN (Naval Supply Center, Cheatham Annex and Newport News Park) in York.

5.3 Model framework

For practical situations where responses arise as a vector of counts that vary across different observational sample sites/locations, univariate Poisson regression models for each of the components in the response vector cannot account for the association or dependence among the components of the response vector. The dependence may be due to omitted variables which simultaneously affect the response vector Aitchison and Ho (1989); Chib and Winkelmann (2001); Ma et al. (2008), and therefore multivariate modelling approach is needed. However, in the case of tick stages which define the components of the response vector, dependence between the stages is expected, because individuals in stage $k+1$ depend on the survival of individuals from stage k. Suppose the response is a $J-$variate vector of counts from n habitat types at $t = 1, \ldots, T$ time points. A hierarchical multivariate Poisson model is defined as follows for $j = 1, \ldots, J$; $i = 1, \ldots, n$; $t = 1, \ldots, T$:

$$\boldsymbol{Y}_{it}|\boldsymbol{\lambda}_{it} \sim \text{MVP}_J(\boldsymbol{\lambda}_{it}), \tag{5.1}$$

$$\log(\lambda_{j,it}) = \mathbf{D}'_{j,it}\boldsymbol{\gamma}_{j,it} + \boldsymbol{S}'_{j,it}\boldsymbol{\beta}_j + \upsilon_{j,it}, \tag{5.2}$$

where $\boldsymbol{D}_{j,it} = (D_{j,it,1}, \ldots, D_{j,it,a})'$ is an $a-$dimensional vector of exogenous (dynamic) location-time varying predictors with the corresponding dynamic coefficients $\boldsymbol{\gamma}_{j,it} = (\gamma_{j,it,1}, \ldots, \gamma_{j,it,a})$, $\boldsymbol{S}_{j,it} = (S_{j,it,1}, \ldots, S_{j,it,c})$ is an $c-$dimensional

vector of exogenous predictors with static coefficients $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,c})$, and the error term $\upsilon_{j,it}$ incorporates the dependence between the components of the response vector by assuming a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{V}$ for each habitat type i at time t. For identifiability, we assume that if the model includes the habitat type and time varying intercept $\gamma_{j,it,1} = 1$, then there is no intercept $\beta_{j,1} = 1$ in the static part.

For all i and t, we assume that the errors $\upsilon_{j,it} \sim \mathrm{N}(0, V_{jj})$, and that the correlation $Corr(\upsilon_{j,it}, \upsilon_{\ell,it}) = 1$ if $j = \ell$ and $\rho_{j\ell}$ if $j \neq \ell$, $j, \ell = 1, \ldots, J$. Suppose $\mathrm{Cov}(\boldsymbol{\upsilon}_{it}) = \boldsymbol{V} = \{V_{j\ell}\}$ and $\mathrm{Corr}(\upsilon_{it}) = \boldsymbol{R} = \{\rho_{j\ell}\}$. Let $\boldsymbol{M}_\upsilon = \mathrm{diag}(V_{11}, \ldots, V_{JJ})$. Note that $\boldsymbol{V}, \boldsymbol{R}$ and $\boldsymbol{M}_\upsilon$ are $J \times J$ symmetric positive definite matrices, and $\boldsymbol{R} = \boldsymbol{M}_\upsilon^{-1/2} \boldsymbol{V} \boldsymbol{D}_\upsilon^{-1/2}$. For example, if $j = 2$, then the covariance matrix is:

$$\boldsymbol{V} = \begin{pmatrix} \upsilon_{11} & \upsilon_{12} \\ \upsilon_{12} & \upsilon_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\tau_{11}} & \frac{R_{12}}{\sqrt{\tau_{11}\tau_{22}}} \\ \frac{R_{12}}{\sqrt{\tau_{11}\tau_{22}}} & \frac{1}{\tau_{22}} \end{pmatrix}$$

where the marginal precisions of $\upsilon_{1,t}$ and $\upsilon_{2,t}$ are $\tau_{11} = 1/V_{11}$ and $\tau_{22} = 1/V_{22}$, the correlation between $\upsilon_{1,t}$ and $\upsilon_{2,t}$ is $\rho_{12} = V_{12}/\sqrt{V_{11}V_{22}}$. We assume that the precision matrix $\boldsymbol{V}^{-1}$ follows an inverse Wishart distribution with r degrees of freedom and scale matrix $\boldsymbol{\Sigma}^{-1}$; that is $\boldsymbol{V}^{-1} \sim \mathrm{IW}(r, \boldsymbol{\Sigma}^{-1})$.

A detailed exposition of the models described in 5.2 can be found in Fokianos (2015) and Ravishanker et al. (2022) among many others. We extend the model in 5.2 to include the random effect, b_k, which quantify the variation in abundance that is accounted for by the m sample sites' variance. The resulting multivariate dynamic Poisson model is defined as follows for $j = 1, \ldots, J$; $i = 1, \ldots, n$; $k = 1, \ldots, m$; $t = 1, \ldots, T$:

$$\boldsymbol{Y}_{it} | \boldsymbol{\lambda}_{it} \sim \mathrm{MVP}_J(\boldsymbol{\lambda}_{it}), \tag{5.3}$$

$$\log(\lambda_{j,ikt}) = \mathbf{D}'_{j,ikt}\boldsymbol{\gamma}_{j,ikt} + \boldsymbol{S}'_{j,ikt}\boldsymbol{\beta}_j + \upsilon_{j,ikt} + b_k, \tag{5.4}$$

where $\boldsymbol{D}, \boldsymbol{\gamma}, \boldsymbol{S}$, $\boldsymbol{\beta}$ and $\upsilon_{j,it}$ are defined as in 5.2, and b_k is the random effects term which follows independent and identical normal distribution (iid) such that, $b_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}_{b_k})$. For $j = 2$, the diagonal covariance matrix $\boldsymbol{W}_{b_k}$ is given as:

$$\boldsymbol{W}_{b_k} = \begin{pmatrix} \omega_{11} & 0 \\ 0 & \omega_{22} \end{pmatrix}$$

where the diagonal precisions of ω_{11} and ω_{22} are $\varrho_{11} = 1/\omega_{11}$ and $\varrho_{22} = 1/\omega_{22}$.

5.4 Simulations

This section presents a simulation study to mimic the observed data generation process. We assume a two-component outcome vector for simplicity. Let $\boldsymbol{Y}_{ik} = (Y_{1ik}, Y_{2ik})$ represent a bivariate ($J = 2$) vector of counts data observed at $k = 1, \ldots, m$ random sample sites that are within $i = 1, \ldots, n$ locations. We simulate the data according to the simplified model described as follows:

$$\boldsymbol{Y}_{ik}|\boldsymbol{\lambda}_{ik} \sim \mathrm{MVP}_J(\boldsymbol{\lambda}_{ik}), \tag{5.5}$$

$$\log(\lambda_{j,ik}) = \boldsymbol{S}_{j,ik}\boldsymbol{\beta}_{j,ik} + \upsilon_{j,i} + b_k, \tag{5.6}$$

where $\boldsymbol{S}$ and $\boldsymbol{\beta}$ represent vectors of static predictors and coefficients respectively, each simulated from a multivariate normal distribution, such that $\boldsymbol{S} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{V}_s)$ and $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{V}_\beta)$. The prior distribution for $\boldsymbol{S}_i \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(10^3))$ and $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(10^3))$. The error term $\upsilon_{j,i}$ is simulated from $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \mathrm{diag}(V_{11}, V_{22})$ and $\rho_{12} = \rho_{21}$ for $i = 1, \ldots, n$. We let the precisions terms be

$V_{11} = V_{22} = 0.4$, and $\rho_{12} = 0.90$.

The errors in the random effects term b_k are simulated from a multivariate normal i.i.d, that is $b_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}_{b_k})$ with $\boldsymbol{W}_{b_k} = \mathrm{diag}(\omega_{11}, \omega_{22})$, where the precision parameters are based on the inverse-gamma distribution; that is $\varrho_{11} \sim \mathrm{IG}(a_1, b_1)$ and $\varrho_{22} \sim \mathrm{IG}(a_2, b_2)$ (Grilli et al., 2015). The hyperpriors a and b are the shape and rate parameters in the gamma distribution respectively. Since the choice of priors can have a significant impact on the posterior distributions of the model parameters and model performance can be sensitive to the choice of sample site-specific random effects variance priors Gelman (2006), we considered three hyperpriors for ω_{11} and ω_{22} in $\boldsymbol{W}_{b_k}$ through the precision parameters ϱ_{11} and ϱ_{22} in the precision matrix $\boldsymbol{\Omega}_{b_k}^{-1}$. The hyperpriors are as follows:

- $\Gamma(1, 0.0005)$, the default choice of the `inla()` function in the R-INLA package (Rue et al., 2009);
- $\Gamma(0.001, 0.001)$, the default choice of the BUGS software (Lunn et al., 2012) and
- $\Gamma(0.5, 0.0164)$ proposed by Fong et al. (2010), corresponding to random effects b_k with marginal Cauchy distribution such that $e^{b_k} \in [0.1, 10]$ with probability 0.95.

Bayesian estimation was carried out using the `inla()` function from the R-INLA package by (Rue et al., 2009). The aim was to investigate the influence of n and m on the posterior estimates. Since the tick data under study has $n = 3$ habitat types and $m = 8$ or 9 random sample sites, we consider two scenarios in estimating the posterior parameters: In the first scenario we allow the location type of size n to change, while the random sample site of size is fixed at $m = 10$ while $n = 3, 10, 20, 50, 100$ and 1 000. In the second scenario, $n = 5\ 000$ and $m = 10, 50, 100, 500, 1\ 000$ and 2 500. The second scenario investigates the effects

in the posterior parameters using the hyperprior in the `inla()` function by Rue et al. (2009), BUGS software by (Lunn et al., 2012) and that by Fong et al. (2010). We set a large value of $n = 5\ 000$ and observe the change in the posterior parameter as $m = 10, 50, 100, 500, 1\ 000$ and $2\ 500$. The results of the simulation are presented in Table 5.1.

5.4.1 Simulation results

Table 5.1 presents the results of the simulations for different values of n when m is fixed in scenario 1, and simulations for different values of m when n is fixed in scenario 2. In scenario 1 we note that posterior estimates are close to the true values when $n \geq 50$ and $m = 10$, and that credible intervals become narrower as n increases. In scenario 2, posterior estimates of the sample site-specific random effects are close to the true values for $n \geq 1\ 000$ for INLA and BUGS hyperpriors as compared to the BUGS hyperpriors. Credible intervals are narrower throughout the values of m, and the results show that posterior estimates are approximately equal to the true values when there is $\frac{m}{n} = \frac{1\ 000}{5\ 000} \approx 20\%$ of m random sample sites that are within n habitat types locations. The results of the simulation reveal that posterior estimates are not accurate for $n \leq 20$ and $m = 10$, and these sizes include the sizes for our tick data.

5.5 Model application on tick life stage data

In order to understand the dynamic causes and consequences of variation in the abundance of ticks, multivariate hierarchical dynamic and static count data model, which include sample site random effects terms, were used to model the tick life-stage data. In the application, we let $\boldsymbol{Y}_{ikt} = (larvae, nymphs, adults)'$ be a $3-$dimensional time series vector of tick counts collected from location $i = 1, \ldots, 3$, comprising $k = 1, \ldots, 12$ random sampling sites over $t = 1, \ldots, 10$ years

Table 5.1: Estimated posterior parameters in the simulated model.

Parameter	SCENARIO 1 ($m = 10$)				SCENARIO 2 ($n = 5000$)									True value
	INLA					INLA			BUGS			FONG			
	n	Posterior mean	95% CI 2.5%	95% CI 97.5%	m	Posterior mean	95% CI 2.5%	95% CI 97.5%	Posterior mean	95% CI 2.5%	95% CI 97.5%	Posterior mean	95% CI 2.5%	95% CI 97.5%	
β_1		4.773	1.995	11.007		5.028	4.632	5.459	5.024	4.743	5.323	5.026	4.686	5.394	5.0
β_2		3.585	1.458	8.577		4.967	4.609	5.351	4.964	4.628	5.326	4.968	4.590	5.375	5.0
V_{11}		0.232	0.087	1.089		0.410	0.386	0.435	0.410	0.385	0.435	0.411	0.387	0.437	0.4
V_{22}	3	0.220	0.083	1.002	10	0.424	0.400	0.449	0.425	0.399	0.451	0.424	0.400	0.451	0.4
ρ_{12}		-0.039	-0.661	0.597		0.876	0.858	0.895	0.877	0.857	0.894	0.877	0.858	0.896	0.9
ω_{11}		0.000	0.000	0.004		0.019	0.014	0.027	0.009	0.006	0.015	0.013	0.005	0.036	0.01
ω_{22}		0.000	0.000	0.004		0.015	0.011	0.022	0.014	0.009	0.020	0.018	0.011	0.031	0.01
β_1		3.145	1.936	4.980		5.031	4.917	5.147	5.032	4.915	5.150	5.032	4.915	5.150	5.0
β_2		2.785	1.249	5.424		4.969	4.855	5.087	4.974	4.857	5.093	4.974	4.857	5.093	5.0
V_{11}		0.201	0.084	0.695		0.409	0.386	0.432	0.408	0.387	0.432	0.408	0.387	0.432	0.4
V_{22}	10	0.613	0.235	2.366	50	0.425	0.400	0.449	0.423	0.401	0.448	0.423	0.401	0.448	0.4
ρ_{12}		0.304	-0.394	0.824		0.877	0.859	0.895	0.878	0.860	0.894	0.878	0.860	0.894	0.9
ω_{11}		0.000	0.000	0.004		0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.01
ω_{22}		0.000	0.000	0.004		0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.01
β_1		3.667	2.454	5.292		5.033	4.920	5.148	5.032	4.914	5.152	5.030	4.903	5.159	5.0
β_2		3.337	2.220	4.831		4.972	4.859	5.087	4.972	4.855	5.091	4.974	4.850	5.100	5.0
V_{11}		0.383	0.185	0.974		0.407	0.384	0.433	0.405	0.382	0.429	0.407	0.385	0.431	0.4
V_{22}	20	0.370	0.178	0.984	100	0.423	0.400	0.449	0.426	0.401	0.451	0.419	0.398	0.444	0.4
ρ_{12}		0.596	0.161	0.872		0.878	0.859	0.895	0.876	0.858	0.895	0.880	0.863	0.895	0.9
ω_{11}		0.001	0.000	0.004		0.001	0.000	0.003	0.002	0.001	0.004	0.005	0.003	0.007	0.01
ω_{22}		0.000	0.000	0.004		0.001	0.000	0.002	0.001	0.001	0.003	0.004	0.002	0.007	0.01
β_1		4.843	3.907	5.925		5.030	4.917	5.145	5.029	4.916	5.144	5.033	4.917	5.150	5.0
β_2		4.189	3.240	5.317		4.975	4.863	5.087	4.972	4.861	5.084	4.966	4.851	5.082	5.0
V_{11}		0.293	0.170	0.551		0.405	0.382	0.429	0.407	0.384	0.432	0.400	0.378	0.427	0.4
V_{22}	50	0.463	0.277	0.841	500	0.421	0.398	0.446	0.422	0.400	0.447	0.421	0.397	0.447	0.4
ρ_{12}		0.724	0.484	0.879		0.884	0.864	0.901	0.880	0.862	0.897	0.886	0.868	0.904	0.9
ω_{11}		0.000	0.000	0.006		0.005	0.003	0.009	0.001	0.001	0.002	0.008	0.005	0.011	0.01
ω_{22}		0.000	0.000	0.006		0.003	0.002	0.005	0.001	0.000	0.001	0.005	0.004	0.008	0.01
β_1		4.383	3.640	5.225		5.017	4.906	5.131	5.029	4.916	5.144	5.038	4.921	5.155	5.0
β_2		4.610	3.963	5.325		4.963	4.853	5.076	4.973	4.860	5.087	4.962	4.849	5.077	5.0
V_{11}		0.517	0.359	0.775		0.402	0.378	0.426	0.405	0.382	0.431	0.397	0.374	0.430	0.4
V_{22}	100	0.297	0.199	0.472	1000	0.416	0.393	0.442	0.421	0.399	0.446	0.418	0.395	0.443	0.4
ρ_{12}		0.738	0.555	0.867		0.895	0.871	0.926	0.882	0.858	0.904	0.890	0.872	0.907	0.9
ω_{11}		0.000	0.000	0.004		0.010	0.007	0.013	0.003	0.002	0.008	0.010	0.007	0.014	0.01
ω_{22}		0.000	0.000	0.004		0.009	0.006	0.012	0.002	0.001	0.005	0.010	0.007	0.013	0.01
β_1		5.028	4.632	5.459		5.029	4.920	5.140	5.041	4.930	5.153	5.024	4.913	5.138	5.0
β_2		4.967	4.609	5.351		4.963	4.853	5.075	4.971	4.861	5.083	4.970	4.859	5.083	5.0
V_{11}		0.410	0.386	0.435		0.403	0.374	0.428	0.400	0.378	0.429	0.404	0.376	0.429	0.4
V_{22}	1000	0.424	0.400	0.449	2500	0.418	0.391	0.444	0.418	0.395	0.447	0.415	0.391	0.440	0.4
ρ_{12}		0.876	0.858	0.895		0.889	0.870	0.908	0.887	0.868	0.904	0.894	0.873	0.911	0.9
ω_{11}		0.019	0.014	0.027		0.008	0.005	0.011	0.007	0.004	0.012	0.010	0.007	0.016	0.01
ω_{22}		0.015	0.011	0.022		0.007	0.005	0.011	0.004	0.002	0.010	0.009	0.005	0.015	0.01

time points. Suppose that counts $\boldsymbol{Y}_{ikt}$ follow a multivariate Poisson (MVP) or the negative binomial distribution (MVNB). The model can be written as:

$$\boldsymbol{Y}_{ikt}|\boldsymbol{\delta}, \boldsymbol{\lambda}_{ikt} \sim \begin{cases} \text{MVP}_J(\boldsymbol{y}_{ikt}|\boldsymbol{\lambda}_{ikt}); \text{ when } \boldsymbol{\delta} = \mathbf{1}; \text{ or} \\ \text{MVNB}_J(\boldsymbol{\lambda}_{ikt}, \boldsymbol{\delta}); \text{ otherwise,} \end{cases} \tag{5.7}$$

where $\boldsymbol{\lambda}_{ikt}$ and $\boldsymbol{\delta}$ are the vectors of mean and dispersion parameters, respectively. The mean count vector of parameters $\boldsymbol{\lambda}_{ikt}$ is linked to the dynamic and static

covariates through the log link function as follows:

$$\begin{aligned}\log(\lambda_{j,ikt}) &= \boldsymbol{\beta}_0 + \boldsymbol{Month}_{it} + \boldsymbol{Year}_{it} \\ &+ \boldsymbol{\beta}_1 I(Spring = 1)_{it} + \boldsymbol{\beta}_2 I(Summer = 2)_{it} + \boldsymbol{\beta}_3 I(Fall = 3)_{it} \\ &+ \boldsymbol{\beta}_4 I(Grass = 2)_i + \boldsymbol{\beta}_5 I(Woods = 3)_i \\ &+ \boldsymbol{\beta}_6 Cosine_t + \boldsymbol{\beta}_7 Sine_t + \upsilon_t + b_k \\ &= \boldsymbol{D}'_{j,ikt}\boldsymbol{\gamma}_{j,ikt} + \boldsymbol{S}'_{j,ikt}\boldsymbol{\beta}_j + \upsilon_{j,t} + b_k, \end{aligned} \tag{5.8}$$

where the $\boldsymbol{D}$, $\boldsymbol{\gamma}$, $\boldsymbol{S}$, $\boldsymbol{\beta}$, $\boldsymbol{\upsilon}$ and b are defined as in 5.2 and 5.4. Suppose that the number of coefficients in the dynamic and static vectors are given by $p_d = \sum_{j=1}^{J} a_j$ and $p_s = \sum_{j=1}^{J} b_j$ respectively, and let $\boldsymbol{\pi}_{it}$ be a p_d-dimensional vector constructed by stacking the a_j coefficients, $\boldsymbol{\gamma}_{it}$. That is,

$$\boldsymbol{\pi}_{it} = \boldsymbol{\theta}_t + \boldsymbol{\upsilon}_t \tag{5.9}$$

where $\boldsymbol{\theta}_t$ is the state parameter and $\boldsymbol{\upsilon}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{V})$. The state equation is given as

$$\boldsymbol{\theta}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\alpha}_t \tag{5.10}$$

where $\boldsymbol{G}$ is a $p_d \times p_d$ state transition matrix, and the state errors $\boldsymbol{\alpha}_t$ are assumed to be i.i.d $\mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda})$. Thus, the hierarchical equation in 5.8 can be written in a more compact form as follows:

$$\log(\lambda_{j,ikt}) = \boldsymbol{I}\boldsymbol{\pi}_{j,ikt} + \boldsymbol{\beta}\boldsymbol{S}_{j,ikt} + b_k. \tag{5.11}$$

We assume multivariate normal priors for the static predictor $\boldsymbol{\beta}_j$ as $\mathrm{N}(\mathbf{0}, \mathrm{diag}(10^3))$, and inverse-Wishart priors in $\boldsymbol{V}_i$, $\boldsymbol{\Lambda}_j$, and $\boldsymbol{W}_{b_k}$ for $\boldsymbol{\upsilon}_t$, $\boldsymbol{\alpha}_t$ and b_k respectively.

The reference category for the habitat type covariate is taken to be 1 corresponding

to the edge habitat type, while the season variable is taken to be the winter season corresponding to low tick activity. Month and year covariates entered the model as continuous variables and we assume that they evolve according to a random walk process, and the sample sites' covariates were taken as random effects.

We employed Bayesian estimation using the R-INLA package by Rue et al. (2009) given its fast computation time compared to the MCMC counterpart. The Bayesian approach treats parameters as unknown random variables having prior probability distributions which are updated with data leading to posterior distributions in contrast to how likelihood methods treat parameters as fixed constants. Here, the likelihood of the model describes the data generating process given the parameters, while the prior usually reflects any previous information about the model parameters. Consequently, when the prior knowledge is scarce then we assume vague or non-informative priors so that the posterior distribution is driven by the observed data (Gelman, 2006).

5.6 Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \bar{D}(\theta) + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter et al. (2002) state that p_D values less than zero indicate substantial conflict between the prior and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of deviance suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter et al. (2002).

5.7 Model application results

To identify factors associated with tick life stage abundances, several models were fitted assuming the Poisson and the negative binomial multivariate distributions. Since the choice of priors can have an important impact on the posterior distributions of the model parameters and the performance of the model can be sensitive to the choice of the location-specific random effects variance priors Gelman (2006), we consided three inverse-gamma distribution priors for sample site random effects term $b_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}_{b_k})$, postulated through the precision matrix $\boldsymbol{\Omega}_{b_k}^{-1}$, that is $\boldsymbol{W}_{b_k} \sim IW(r_{b_k}, \boldsymbol{\Omega}_{b_k}^{-1})$. The fitted model included the intercept, sinusoidal terms, season and habitat type fixed effects, month and year non-linear effects, and the abundance sample site random effects. The model was fitted using three priors of the hyperparmeter for the variance of sample site random effects. Figure 5.1 displays tick abundance variations in the posterior densities of the hyperpriors. The plots suggest that the Rue et al. (2009) prior (black curve) was better in modelling the tick life stages since its log-likelihood curve was higher than the other curves. However, the DIC value of the Lunn et al. (2012) prior was smaller for the NB model compared to other models (Table 5.2), suggesting the results of this prior to be preferred for our multivariate data. We note that the posterior estimates of the models are similar and are presented in Table 5.3.

Table 5.2: Summary of Poisson and negative binomial model DIC values for different priors.

	DIC	
Prior	Poisson	Neg.Bin
$\Gamma(1, 0.0005)$	-206894.20	41717.78
$\Gamma(0.001, 0.001)$	-206887.00	41706.62
$\Gamma(0.5, 0.0164)$	-206891.10	41716.74

Table 5.3 displays the posterior means and the 95% credible interval estimates from the NB model assuming three sample sites random effects priors. Since the results from the fitted models are similar, we only interpret the the results from

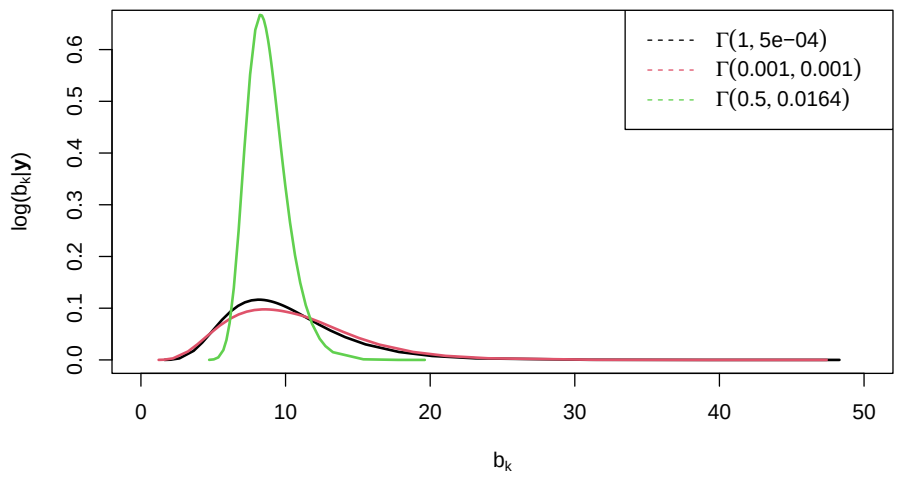


Figure 5.1: Sensitivity analysis on the priors of precision parameter of fitted negative binomial model. The black, red and green curves are for $\Gamma(1, 0.0005)$, $\Gamma(0.001, 0.001)$ and $\Gamma(0.5, 0.0164)$, respectively.

the first model M1.

The estimated effects of the predictors of the tick abundance reveal no significant effects due to the type of habitat. That is, the model estimated non-significant and higher [0.33; CI:(-0.04,0.69) and 0.12; CI:(-1.03, 1.28)] log mean overall tick abundances in the grass and woods habitat types compared to the edges. However, the model estimated significant and higher [4.24; CI:(0.92,7.56)] log mean overall tick abundances in summer season, while non-significant [3.34; CI:(-0.14,7.96) and 2.74, (-1.14,6.50)] abundances were estimated during spring and fall seasons. The model estimated non-significant [-0.77; CI:(-2.99,1.36) and 0.02; CI:(-2.11)] cosine and sine sinusoidal terms. This indicates the model's inability to capture the seasonal patterns in the data. The model estimated a significant [5.09; CI:(2.61,12.65)] overall sample site random effect, which indicates differing overall tick abundances within sample sites with similar environmental and temporal conditions. The model estimated non-significant [-0.02; CI:(-0.41,0.33) and -0.04; CI:(-0.46,0.38))] correlations between larvae and nymph and nymph and adult ticks, while non-significant [0.03; CI:(-0.44,0.54)] correlation was estimated be-

tween larvae and adults.

Table 5.3: Posterior estimates of the negative binomial models assuming **M1**: $\Gamma(0.1, 0.0005)$, **M2**: $\Gamma(0.001, 0.001)$ and **M3**: $\Gamma(0.5, 0.0164)$ priors in the precision matrix $\boldsymbol{\Omega}$.

Covariates	**M1**		**M2**		**M3**	
Fixed Effects	Mean	95% CI	Mean	95% CI	Mean	95% CI
(Intercept)	-4.87	(-7.67, -2.07)	-4.88	(-7.91, -1.86)	-4.90	(-8.14, -1.65)
Season (**ref**: Winter)						
Spring	3.91	(-0.40, 8.13)	3.89	(-0.68, 8.37)	3.84	(-1.17, 8.68)
Summer	4.24	(0.73, 7.75)	4.25	(0.51, 7.99)	4.25	(0.16, 8.34)
Fall	2.70	(-1.42, 6.66)	2.66	(-1.71, 6.90)	2.60	(-2.23, 7.19)
Habitat type (**ref**: Edge)						
Grass	0.33	(-0.03, 0.69)	0.33	(-0.03, 0.70)	0.33	(-0.03, 0.70)
Woods	0.12	(-1.03, 1.28)	0.12	(-1.03, 1.29)	0.14	(-1.03, 1.33)
Cosine term	-0.77	(-2.99, 1.36)	-0.84	(-3.35, 1.57)	-0.89	(-3.70, 1.71)
Sine term	0.02	(-2.12, 2.11)	-0.02	(-2.46, 2.36)	-0.05	(-2.75, 2.55)
Random Effects						
b_k	5.09	(2.61, 12.65)	4.82	(2.75, 8.77)	5.34	(2.04, 12.37)
Correlations						
ρ_{LN}	-0.02	(-0.41, 0.33)	0.00	(-0.24, 0.25)	0.08	(-0.35, 0.54)
ρ_{LA}	0.03	(-0.44, 0.54)	-0.01	(-0.26, 0.23)	0.04	(-0.39, 0.50)
ρ_{NA}	-0.04	(-0.46, 0.38)	0.03	(-0.38, 0.46)	0.02	(-0.41, 0.46)

5.7.1 Sample sites random effects

Figure 5.2 shows the caterpillar plot for the estimated sampling sites random effects posterior means (the dots) and 95% credible intervals. A negative posterior mean value for the sampling site random effects implies that lower variations are estimated, while a positive value implies higher variations are estimated within the sampling site. The model estimated significant higher tick variation in Stephens tract (ST), Jacobson tract (JC), Newport News Park (NN), Kiptopek (KP) and Hoffer Creek Wildlife Preserve (HC) and significant lower abundances in Paradise Creek Nature Park (PC) and Weyanoke (WS). Non-significant lower variations were estimated in Blackwater Ecological Preserve (BW), Oceana Dam/Neck (OD), Back Bay (BB) and Langley (LA), while non-significant higher abundances were estimated in Cheatham Annex (CA). The results from the model suggest that ST and JC sampled sites in Cheasapeake county, CA in York county, and KP Portsmouth county and KP in Northampton county had higher tick variations,

while lower variations were estimated in PC sampling site in Portsmouth county and in WS sampling site in Norfolk county. On the other hand, non-significant lower variations were estimated in Virginia Beach, Isle of Wight, Hampton and York counties. The *A. americanum* is found throughout the southeastern United States, and its populations extend west to central Texas and north to Iowa (Childs and Paddock, 2003).

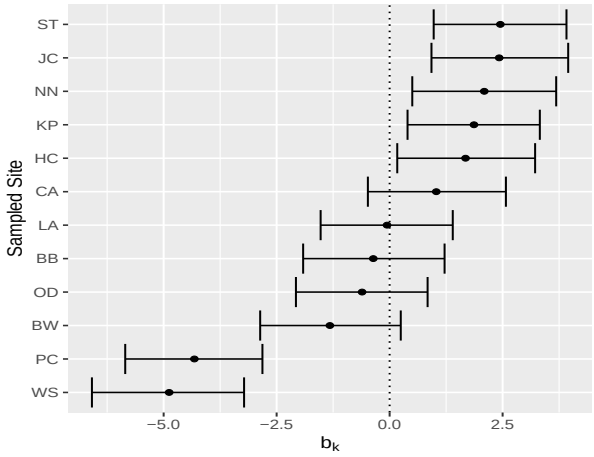


Figure 5.2: Caterpillar plots for predictors of the random effects of sampling locations with posterior means (dots) and 95% credible intervals in Virgnia, USA.

5.7.2 Non-linear effects

Figure 5.3 shows the shared trend effects of months and years over the study period, 2009 - 2018. The month trend plot shows that the log of mean tick abundances increase during the first six months (January to June), then decrease between June and July, increase between July and September, decrease between September and November, and increase from November to December. The year trend plot shows a steady decline in the log of mean tick abundances between 2009 through 2012, followed by a sharp decrease from 2012 through 2015, and then a steady decline from 2015 to 2017, after which a steady increase from 2017 through 2018 can be observed.

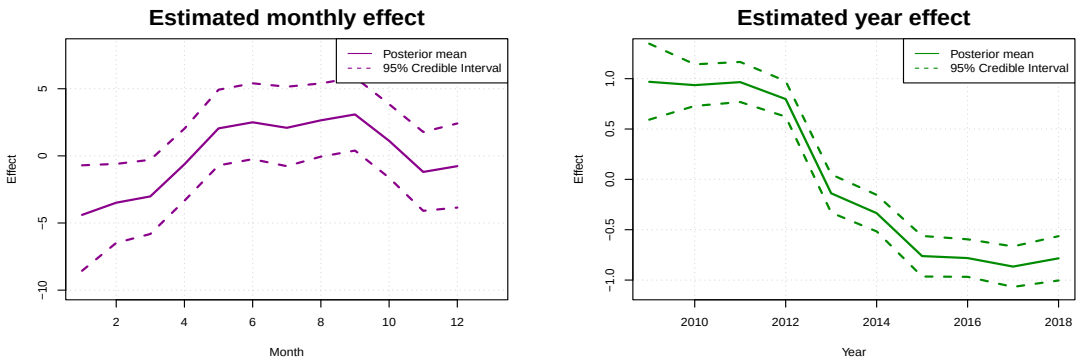


Figure 5.3: Estimated shared month and year time trends for larvae, nymph and adult tick life stages.

5.8 Discussion

This study used hierarchical dynamic Poisson and negative binomial count data models to investigate the influence of temporal and location specific effects on the distribution of *A. americanum* abundance in Virginia, United States. The model included the sampling sites random effects term to assess tick abundance variations within the sampling sites. The proposed model was investigated through simulations at different locations and sampling site sizes. It is worth noting that for small samples sizes the random effects model proposed may be susceptible to bias (Grilli et al., 2015). Furthermore, the model was fitted to the tick data and results showed that there were no significant effects of habitat types. We also noted that credible intervals were wider as a result of small $n = 3$ and $m = 12$ values. However, summer season effects were significant and positive indicating an expected increase in tick abundance during summer seasons. This aligns with the results in Childs and Paddock (2003); Kennedy et al. (2021); Lephoto et al. (2021); Alkishe and Peterson (2022), where the larvae life stage abundance peaks in August, the nymph stage in July, and the adult stage in June. The evidence of sample site abundance variations was detected in Chesapeake, York, Portsmouth, Northampton, and Norfolk counties suggesting potential disease hotspots. These results confirm the results in (Gaines et al., 2014; Cohen et al., 2010). The esti-

mate of non-linear effects of the model showed an upward monthly tick abundance trend from January to June, a dip between June and July, steady peak between July and September, a sharp decrease between September and November, and a steady peak from November to December. However, the year tick abundance trend showed a steady decrease between 2009 and 2011, a steep decrease between 2011 and 2012, and a sharp decrease from 2012 through 2018. These trends are well documented in Davidson et al. (1994); Kollars Jr et al. (2000); Mangan et al. (2018).

5.9 Conclusion

In this study, we investigated the effects of dynamic, static and random sample site effects on the distribution of Lone Star tick abundance using the hierarchical dynamic Poisson and negative binomial count data models. The simulation study showed that the proposed model accurately captured the data for larger sample sizes. We conclude that application of the proposed model to data did not fit well, and wide credible intervals were obtained. Although the model showed insignificant relationships between tick abundance and the habitat type, the model was able to reveal that significant higher tick abundances were expected in summer compared to winter seasons. The model revealed that there were significant variations due to the random sample sites.

Chapter 6

Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients

6.1 Introduction

There is a vast amount of geo-referenced data in many fields of study including ecological studies. Geo-referencing is usually by point referencing; that is, latitudes and longitudes or by areal referencing, which includes districts, counties, states, provinces and other administrative units. The availability of large geo-referenced datasets for modelling has necessitated the development and application of spatial statistical methods. However, spatial varying coefficients models exploring the abundance of tick counts remain limited. Efficient assessment of spatial and temporal patterns in species distributions, for example, to draw species distribution maps and to estimate population trends, has been a challenging but essential endeavour for ecologists and wildlife managers for a long time (Barker and Sauer, 1992; Inger et al., 2015). Species distribution or abundance models that specify

population trends help understand basic ecological questions (Sattler et al., 2007). Reliable distribution and trend estimates are essential for a smart resource allocation in conservation (Rodrigues et al., 2006) and species management (Sutherland et al., 2004).

6.2 Data description

The data were collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, in Virginia, United States (US). Standard flagging techniques (CDC, 2020) were used along established transects and identified species and life-stage according to (Sonenshine, 1979). Ticks life stages count data were recorded for each different location, habitat and habitat type through the four seasons. Data collection was conducted from independent cities, namely; City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County. For brevity, all will be referred to as counties. Data collection was done at least once a month on varying days of the week and at 12 different sites in southeast Virginia from May 2009 through December 2018 (Lephoto et al., 2021). This chapter uses data arranged by habitat type, collected throughout the four seasons of the year as per the objectives.

The data were collected at random from multiple areas referred to as habitats, and each area was designated by a unique number ranging from 1 to 5 for different habitats. The habitat type was used to designate the kind of area (woods, edges or grass) where the data were collected. The number of the week (from 1 to 53) was also recorded during data collection, with week 1 being the first week of the year. While fewer ticks were observed in winter, the adult stage of *I. scapularis* is active in winter. To help align the information from year to year, we recorded data collected during the last week of December as week 53, which was also the

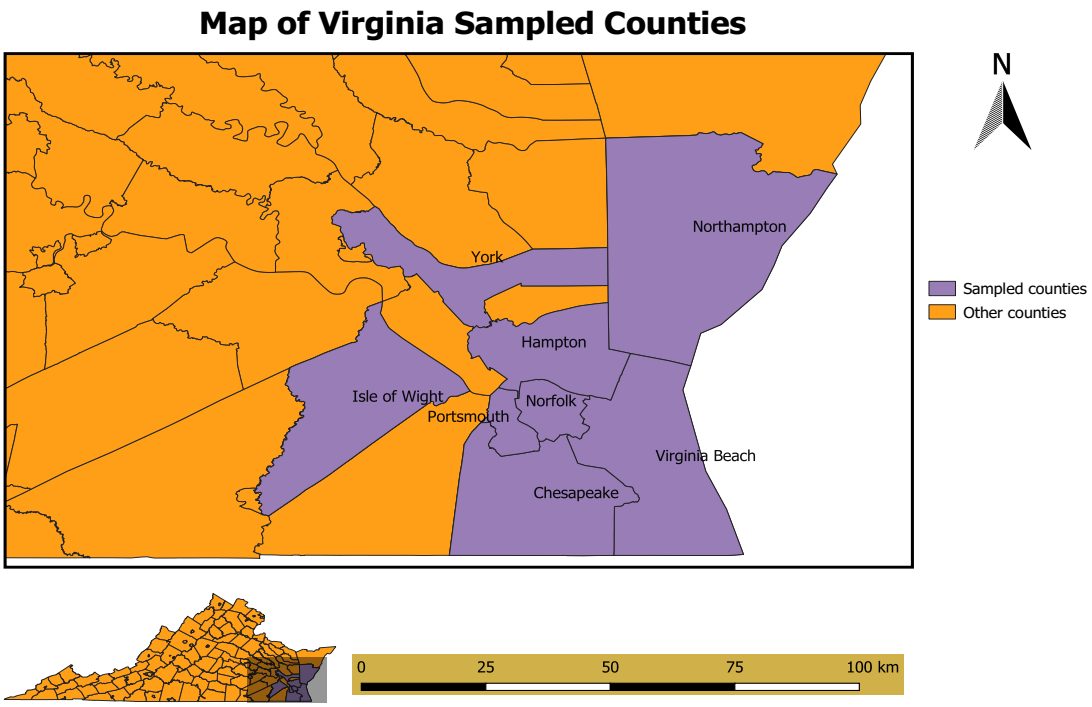


Figure 6.1: Map of Virginia with names of counties where ticks were sampled.

first week of January of the following year. We also kept track of the month and year for the data collection. The ten-year study period was grouped into two-year successive periods so that 2009 and 2010 were grouped together, 2011 and 2012, 2013 and 2014, 2015 and 2016 and 2017 and 2018. This predictor variable was used to capture residual spatial effects on the abundance of tick counts. The paired time segmentation was to help find out if there were times when counts were significantly high or low compared with the rest of the data. The above-mentioned variables were used as predictors of tick life-stage count outcome data to find relationships using count regression models.

6.3 Model description

Various statistical models have been developed to model count data. In this study, we applied the Poisson and the negative binomial distribution models. The classical Poisson model, which assumes that mean and variance of the count

responses are equal, is often of limited use when the empirical data sets exhibit over-dispersion or have more zeros than expected. This can be addressed by introducing a dispersion parameter in this model or by extension to models that can account for excessive zeros in the data (Zeileis et al., 2008). As mean and variance are identical in the standard Poisson model, this means that the dispersion parameter is fixed to 1. In the presence of over-dispersion, the Poisson model underestimates the variance and renders all model-based tests more conservative. Violation of the equal mean and variance assumption indicates correlation in the data, which may affect both standard errors of the parameter estimates and the model. In this study, the tick count data showed a greater proportion of zeros than positive counts, which is an indication of zero-inflation and over-dispersion.

The NB distribution is commonly used to model over-dispersed count data. A dispersion parameter in the NB model caters for over-dispersion allowing the variance to be greater than the mean while also accommodating the unobserved heterogeneity in the data. In addition to over-dispersion, it is common that many empirical count data sets exhibit more zero observations than would be expected by the classical Poisson model. A model capable of capturing both properties is the zero-inflated Poisson (ZIP), which assumes that zero counts occur with some probability, while a Possion (λ) random variable is observed with probability 1 – p. The ZIP distribution model approaches the classical Poisson when $p \rightarrow 0$. It is worth noting that zero observations arise from both the zero-component distribution and the classical Poisson distribution. The zero-component distribution accounts for the inflated zeros that are observed in addition to zeros that are expected to be observed under the classical Poisson distribution. A more detailed account of the development of zeroinflated models can be found in Lambert (1992).

Let y_{ijkm} be a tick count for life stage m ($m = 1:$ Larvae, $m = 2:$ Nymphs, and $m =$

3 : Adults), in habitat-type k, habitat j, and location i, modeled by multi-hierarchical Poisson model. Under the Poisson model, we assume that the dependent variable y_{ijkm} is Poisson distributed, i.e

$$y_{ijkm}|\mu_{ijkm} \sim \text{Poisson}(\mu_{ijkm})$$

where μ_{ijkm} is the mean tick count in the respective life-stage, habitat type, habitat and location.

Let X_{ijkm}=$(\boldsymbol{\alpha}_m; x_{ijkm1}, x_{ijkm2}, \ldots, x_{ijkmp})'$ be a vector of p continuous predictors with the first component accounting for the constant and $W_{ijkm} = (w_{ijkm1}, w_{ijkm2}, \ldots, w_{ijkmr})'$ be a vector of r categorical predictors. The link function $h(\cdot)$ relates the mean μ_{ijkm} to the predictors as follows:

$$h(\mu_{ijkm}) = X^T\beta_m + W^T\gamma_m \tag{6.1}$$

where $W = W_{ijkm}$ and $X = X_{ijkm}$; while $h(\cdot)$ is the log link function, β_p an p-dimensional vector of regression coefficients for continuous predictors; and γ_m a vector of the categorical predictors.

In order to cater for non-linear effects of the continuous covariates as well as the spatial autocorrelation and temporal effects in the data, we incorporated the random walk model of order 2 (RW2); the convolution models; the linear trend over years; and space-time interactions into the model such that Equation 6.1 becomes:

$$\begin{aligned} h(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_i(S_i, x_{ijkm}) + f_{month}(m) \\ + f_{spat}(S_{im}) + f_{year}(t) + f_{it}(S_i, t), \end{aligned} \tag{6.2}$$

where the function $f_i(S_i, x_{ijkm})$ represents the space-covariate interaction; f_{month} a non-linear twice differentiable smooth function for the continuous month covariate effect; and the functions $f_{spat}(S_{im})$, $f_{year}(T)$ and $f_{it}(S_i, t)$ are space, time (years) and space-time interactions, respectively. The convolution model assumes that the spatial effect can be decomposed into two; namely, the spatially structured and spatially unstructured components. This means that $f_{spat}(S_{im}) = f_{str}(S_{im}) + f_{unstr}(S_{im})$ where $m = 1, 2$ or 3 (Manda and Leyland, 2007; Ngesa et al., 2014; Okango et al., 2015).

The spatially structured random effects account for the unobserved covariates, which vary spatially across counties, while the spatially unstructured random effects account for the unobserved covariates that are inherent within counties or correlations within counties, for example, the climate and common cultural practices among other things. The spatially structured random effects capture the spatial autocorrelation and they are technically defined as the correlation computed among the values of a single geo-referenced variable that is attributable to the geographic proximity of the objects, to which the values are attached (Cliff and Ord, 1981). Moreover, $f_{year}(t)$ represents random time effects which can be modeled as a first-order random walk (RW1) or a first-order autoregressive process (AR1), while $f_{it}(S_i, t)$ is a space-time interaction.

In the presence of over-dispersion, the Poisson model is replaced by the negative binomial model, where the variance depends on the mean as $\mu(1+\mu\Psi)$, and where Ψ is an over-dispersion parameter, which captures the extent at which variance deviates from the mean. We assume that a tick count variable y_{ijkm} follows a negative binomial distribution, such that

$$y_{ijkm}|\mu_{ijkm} \sim \text{NB}(\boldsymbol{\mu}, \boldsymbol{\psi}).$$

The mean function $E(Y) = \mu$ relates to the predictors in the same way as that of the Poisson distribution model in Equation 6.1. In this chapter, a full Bayesian estimation approach was used, where parameters were assigned prior distributions.

6.4 Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \bar{D}(\theta) + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter et al. (2002) state that p_D values less than zero indicate substantial conflict between the prior and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of deviance suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter et al. (2002).

6.5 Fitted models

Four spatio-temporal models were fitted in the R statistical software, version 3.5.1 using the **INLA** package to predict the effects of the ecological covariates on the distribution of larvae, nymph and adult tick counts in the eight previously mentioned counties of Virginia, U.S. Preparation, organization and merging of the data with the map was done using QGIS software, version 3.6.3-Noosa (QGIS, 2009). The first approach modelled the space-covariate effects $f(S_i, x_{ijkm})$, non-linear effects of the month covariate $f_{month}(m)$, the temporal time effects $f_{year}(t)$ and the space-time effects $f_{it}(S_i, t)$. This model (M1) does not consider the spatially structured and spatially unstructured random effects, and the three life-stage counts were modelled independently as follows:

$$\textbf{M1: } \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_(S_i, x_{ijkm}) + f_{month}(m) + f_{year}(t) + f_{it}(S_i, t),$$

where m equals 1 for larvae; 2 for nymphs; and 3 for adults with respect to counts. The second model (M2) is the same as M1 but with spatially unstructured effects to cater for the unobserved covariates that are inherent within the counties. The spatially unstructured effects were specified by the identically and independently distributed (iid) with the normal distribution.

$$\textbf{M2: } \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_(S_i, x_{ijkm}) + f_{month}(m) + f_{unstr}(S_i) + f_{year}(t) + f_{it}(S_i, t), \tag{6.3}$$

The third model (M3) is the same as M1 but with spatially structured effects which cater for any unobserved covariates which vary spatially across counties, specified by the conditional autoregressive model (CAR).

$$\textbf{M3: } \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_(S_i, x_{ijkm}) + f_{month}(m) + f_{str}(S_i) + f_{year}(t) + f_{it}(S_i, t), \tag{6.4}$$

The fourth model (M4) is the same as M1 with a convolution of spatially unstructured and spatially structured effects, which are specified by the iid normal distribution and CAR model respectively.

$$\textbf{M4: } \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_(S_i, x_{ijkm}) + f_{month}(m)$$
$$+ f_{unstr}(S_i) + f_{str}(S_i) + f_{year}(t) + f_{it}(S_i, t), \qquad (6.5)$$

We used the Poisson and the negative binomial count data distributions and compared them using the DIC. We then interpreted the results from the best performing models based on the DIC.

6.6 Model comparisons

Table 6.1 shows the DICs for four spatio-temporal Poisson and NB models, where the model with the smallest DIC is the one with the best fit. As seen, all the spatio-temporal NB models have lower DICs compared to the corresponding spatio-temporal Poisson models. M1 showed the best fit for the tick larvae count data compared to all the other models, which also suggests that unobserved covariates vary significantly over time. Similarly, M4 showed the best fit for the nymph and adult tick count data. The differences in the DIC values suggest that the spatio-temporal NB model would be the best model compared to the spatio-temporal Poisson model. This outcome also suggests that unobserved covariates vary spatially across counties and over time.

6.7 Results

6.7.1 Space-covariate effects

Spatio-temporal NB models out-performed the spatio-temporal Poisson models in our case. We show the choropleth maps (Figures 2-4) of all the models with the smallest DICs for the fitted larvae, nymph and adult tick count data, respectively.

Table 6.1: DIC values for spatio-temporal Poisson and NB models.

Response	Total counts	Model	Spatio-temporal models			
			Poisson		NB	
			DIC	p_D	DIC	p_D
Larvae	145 020		281147.18	71.79	12563.79	13.31
Nymphs	20 637	1	29982.53	79.52	-	-
Adults	10 509		13990.35	62.55	-	-
Larvae	145 020		277483.97	73.81	12591.82	9.64
Nymphs	20 637	2	30176.78	67.67	14203.41	49.58
Adults	10 509		14308.39	62.55	10101.87	47.52
Larvae	145 020		295800.95	71.16	12600.72	10.82
Nymphs	20 637	3	34093.74	65.24	14518.67	45.52
Adults	10 509		14317.57	62.60	10101.87	47.52
Larvae	145 020		277468.27	73.88	12591.73	9.64
Nymphs	20 637	4	30176.53	67.64	14202.29	49.36
Adults	10 509		14308.23	62.55	10100.97	47.13

The choropleth maps show the space-covariate interaction effects for the selected counties in Virginia. A yellow shade was used if the effects were greater and black or dark shade if the effects were lower.

Larvae

The effect of location on the log mean tick counts was highest in Hampton County followed by York and Chesapeake counties (Figure 6.2). The effect was lower in Norfolk, Portsmouth and Isle of Wight counties. Virginia Beach County showed the lowest effect of the location variable. The effect of habitat was almost the same across all the counties except for Isle of Wight County. The effect of change in weeks was evident in York and Norfolk counties. The effect of habitat type on the log of mean larvae tick counts was high in Chesapeake County. The effect of change in season was very low in Isle of Wight County compared to other counties.

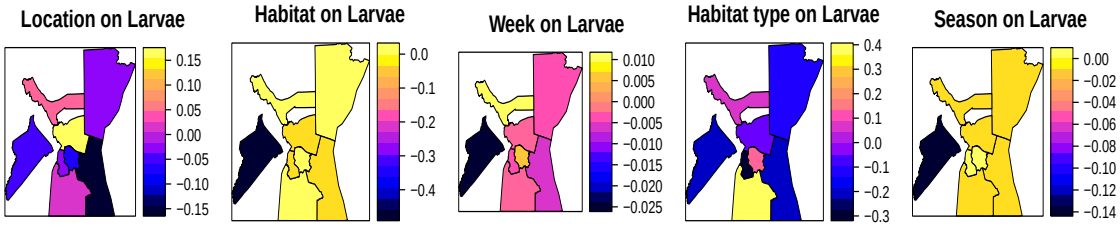


Figure 6.2: Choropleth maps showing the space-covariate effects on larvae.[1]

Nymphs

The effect of location on the log mean nymph counts was high in Hampton County and very low in Norfolk County compared to the others (Figure 6.3). The effect of habitat was very low in Isle of Wight County, and low in Northampton County followed by Chesapeake, Hampton and Portsmouth counties. However, there was a high effect of habitat in Norfolk County. The effect of change in week was very low in York County and Northampton County but high in Portsmouth County. The effect of habitat type on the log nymphal counts was evident in Chesapeake County but low in Isle of Wight County. The effect of change in seasons was very low in Isle of Wight County, while a higher effect can be observed in York, Hampton and Norfolk counties (Figure 6.3).

Adults

The effect of location on the log mean adult counts was high in York County, while it was generally low in other counties. The effect of habitat was high in Norfolk and Hampton counties, but it was a very low in Isle of Wight County. The effect of change in week was also evident in the latter. A very low effect in this respect can be observed in York County (Figure 6.4). The effect of habitat type was very high in Chesapeake County and very low in York County compared with other

[1]Brighter colors = higher spatial effects and dark colors = lower spatial effects.

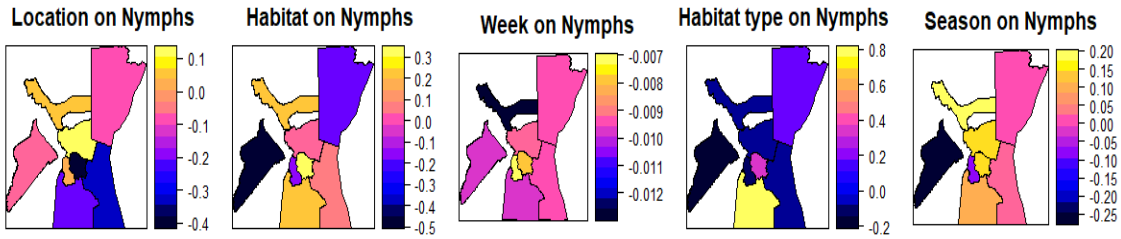


Figure 6.3: Choropleth maps showing the space-covariate effects on nymphs.[1]

counties. The effect of change in season was very high in Norfolk County and very low in Isle of Wight County. Other counties had low effects compared to the effect in the latter.

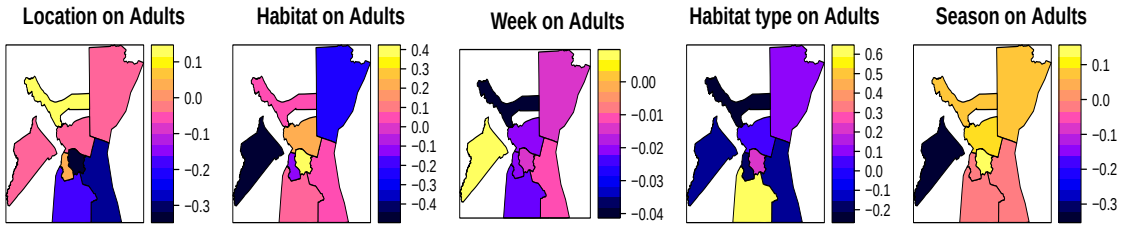


Figure 6.4: Choropleth maps showing the space-covariate effects on adults.[1]

The non-linear effects of the month

Figure 6.5 shows non-linear associations between time (month) and larvae, nymph and adult tick counts. These figures give the posterior mean (black line in the figure) of the smooth function and their corresponding 95% credible intervals (red and blue lines). From the figures it is evident that there is a non-linear relationship between month and tick counts. We confirmed general changes in line with the seasons in the study area. Thus we report the probability of larvae increasing be-

[1]Brighter colors = higher spatial effects and dark colors = lower spatial effects.

tween January and April, fluctuation between April and June, increasing further between June and August, then decreasing until December. Regarding nymphs, the probability is that they increase slightly between January and February, increase abruptly between February and June, decreases between June and August, steadily decrease between August and September, and decrease sharply between September and December. Regarding adults, there is a probability that they increase steadily between January and February, and abrupt increase from February to May, increase steadily between May and June, and decrease sharply between June to September, then start to increase steadily from September to December.

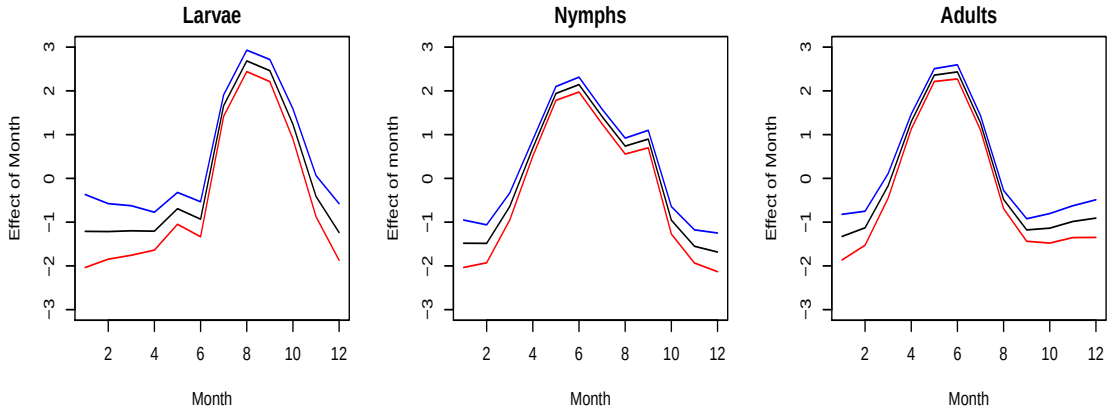


Figure 6.5: Non-linear effects of month variable on larvae, nymphs and adults, 2009-2018.[1]

6.7.2 Space-time effects

Larvae

Figure 6.6 shows the mapped estimated residual spatial effects on the abundance of larvae counts between 2009 and 2018. The residual spatial effects that represent unobserved spatial factors either not measured in the surveys or abducting the effects of cultural patterns, are evident. High effects were observed in Portsmouth, Chesapeake and Hampton counties in 2009/2010 while the other counties showed low residual spatial effects. Two counties showed decreased residual spatial ef-

[1]Note: Red and blue lines = 95% credible intervals; black line = posterior mean.

fects in the period 2011/2012, namely, Portsmouth and Hampton counties. In the period 2013/2014 the effects were evident in Hampton, Portsmouth, Chesapeake and York counties. The effects were high in Chesapeake, Norfolk and York counties in 2015/2016, while they were very low in Portsmouth County. During 2017/2018 all counties, except for Isle of Wight and Portsmouth counties, showed high effects regarding the abundance of larvae.

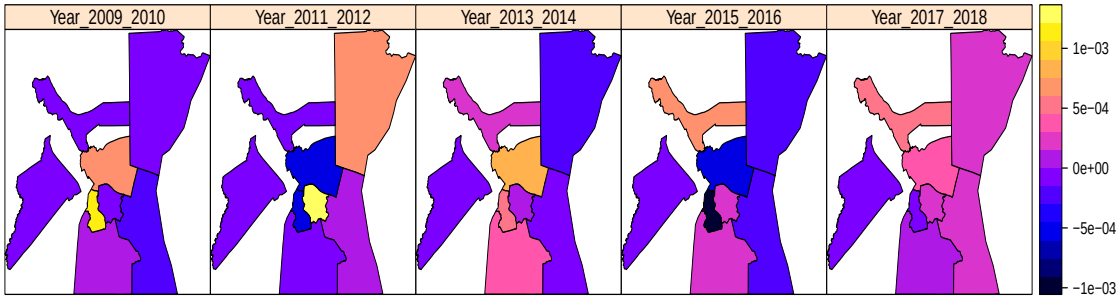


Figure 6.6: Choropleth maps showing residual spatial effect of larvae tick abundance, 2009-2018.[1]

Nymphs

The effects on the nymph abundance decreased over the whole study period, particularly in York County (Figure 6.7).

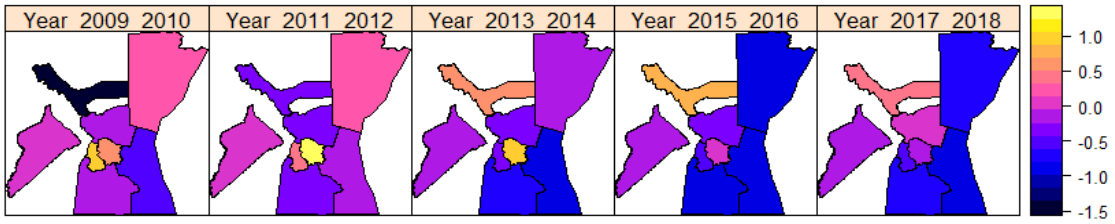


Figure 6.7: Choropleth maps showing residual spatial effect of nymphs tick abundance, 2009-2018.[1]

[1]Brighter colors = higher spatial effects and dark colors = lower spatial effects.

Adults

Residual spatial effects on the distribution of abundance of adult ticks increased between 2009 and 2016 in York County, while all other counties showed a decrease of spatial residual effects throughout the study period 2009 to 2018. In York, a decrease in residual spatial effects was observed in 2017/2018 relatively to the 2015/2016 study period. Higher residual spatial effects were evident in Norfolk County during the periods 2009/2010, 2011/2012 and 2013/2014 (Figure 6.8).

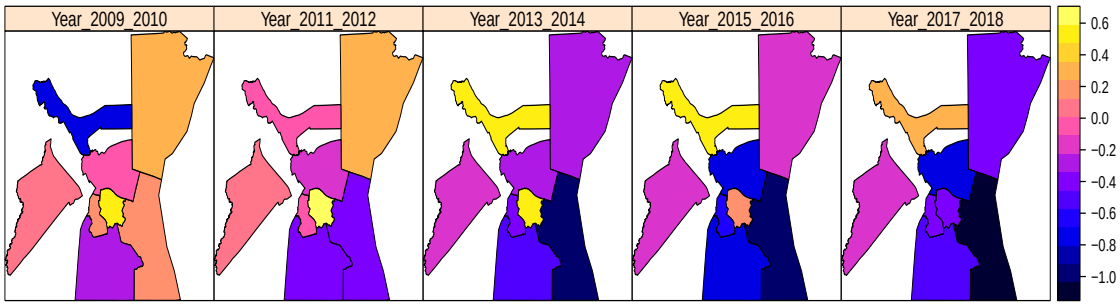


Figure 6.8: Choropleth maps showing residual spatial effect of adult tick abundance, 2009-2018.[1]

6.7.3 Temporal effects

For larvae, there was a steady decrease in the mean count between periods 1 and 4, then a steady increase from period 4. The graph indicates that larvae abundance declined between 2009 and 2016, after which there was a steady increase in the abundance of larvae ticks. The graph indicates that nymph abundance increased between periods 2009 and 2012 then declined between 2012 and 2016, after which there was a steady increase in the abundance of nymph ticks. For adults, the abundance was constant from 2009 through 2012, after which a gradual decline was observed from 2012 until 2018 (Figure 6.9).

[1]Brighter colors = higher spatial effects and dark colors = lower spatial effects.

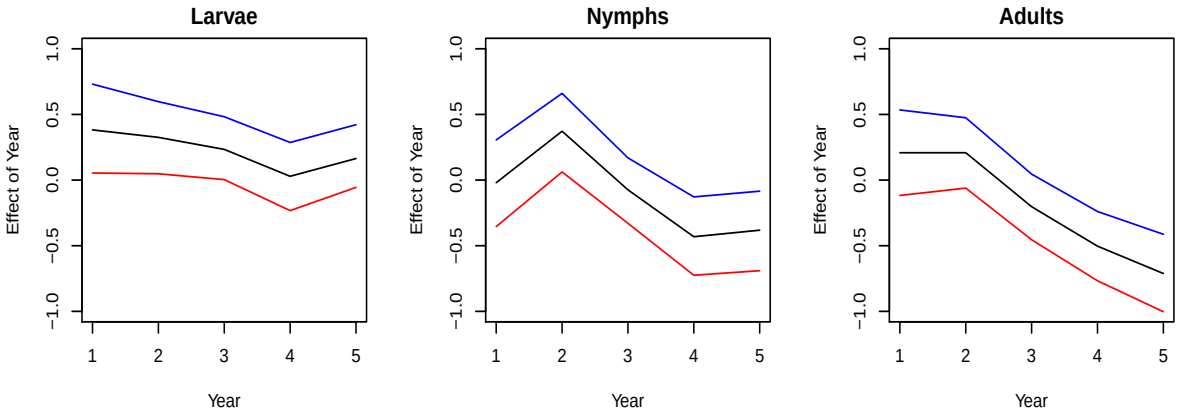


Figure 6.9: The temporal year random effect for the cumulative best fitting model.[1] [2]

6.8 Discussion

This study applied the Poisson and NB count data models to tick count data with the aim of exploring the influence of environmental and temporal predictors on the distribution of tick counts in Virginia, U.S. We relaxed the assumption that the relationship between the predictors and the response variables in a regression model are constant across the study region and over time. This assumption is unrealistic for spatial processes as factors such as sampling variation and different relationships across regions (for example, attitudes, preferences, culture and others) contribute to different responses to the same stimuli across regions and over time. A frequent approach to spatial modelling dates back to the work by Besag et al. (1991) which was extended by Bernardinelli et al. (1995) to include a linear term for space-time interaction. Many studies have relaxed this assumption, for example, Assunçao et al. (1999) introduced a Bayesian space-varying parameter model to examine micro-region factor productivity and the degree of factor substitution in the Brazilian agriculture; Gamerman et al. (2003) developed a flexible modelling approach for space-varying regression models; and Okango et al. (2015) modelled the HIV and HSV-2 viruses using spatially varying coefficient models.

[1]Note: Red and blue lines = 95% credible intervals; black line = posterior mean.

[2]Note: Year 1 = 2009 and 2010; Year 2 = 2011 and 2012; Year 3 = 2013 and 2014; Year 4 = 2015 and 2016; Year 5 = 2017 and 2018.

The Bayesian spatio-temporal process model was used to allow covariates to vary spatially and over time. We specified the CAR prior for the structured random effects; autoregressive of order one (AR1) prior for the temporal random effects; and normal iid priors for the unstructured random effects. Non-linear effects of the month variable on tick counts were also evident. As a result, an assumption of linear relationship would have led to misleading results and consequently to wrong interpretations. The exploratory data analysis showed that more larvae counts were observed compared with nymphs and adult tick counts. We found that the effects of covariates on tick counts varied spatially across counties and over time. Spatio-temporal models were powerful, in the sense that they were capable of revealing county specific effects of each covariate, county space-time effects, and the effects of change in time on the distribution of life-stages of Lone Star ticks. We were able to show that tick abundance has been increasing over the study time in Virginia, which confirms the previous results by Lantos et al. (2015) that observed the significant expansion of Lyme disease distribution in Virginia between 2000 and 2014, particularly southward into the Virginia mountain ranges. It is clear that change in temperature affects tick numbers, such that they decrease in winter (Linske et al., 2020). We confirmed this unsurprising fact and noted also that larvae count remained low up to May. Non-linear effects of month showed that nymphs and adults were observed in spring and summer, meaning that the distribution of larvae would be expected to increase later that summer and thus adults the following spring. This happens in locations in early summer, thus determining the distribution of larvae and adults in late summer and spring the following year, respectively (Stein et al., 2008). Our study also revealed that the effects of habitat type were high in Chesapeake County. This could be because of the abundant white-tailed deer population found in forests, farms, parks and backyards throughout the Chesapeake Bay watershed. The Lone Star tick is said to be very aggressive and specific when looking for hosts (Goddard and Varela-

Stokes, 2009), but they are unspecific during each life-stage, as this species is found on humans, domesticated animals, ground-dwelling birds as well as on small and large wild animals (Kollars Jr, 1993). The white-tailed deer feeds on fruits and vegetation that are available to them each season, which makes it simple for ticks to attach and feed from these animals. During warm seasons (summer and spring), these animals feed on green plants, during fall they feed on nuts, acorns and crops and in winter they feed on woody vegetation, such as bark, twigs and buds of hardwood and pine trees, where indeed ticks are found (Willis et al., 2012).

6.9 Conclusions

Tick counts are influenced by environmental factors and seasonal changes. There is no dominant weekly influence or observable change in the number of ticks due to the changing weeks. We conclude that tick numbers depend on the type of habitat where they are closer to their hosts and the time when their hosts are more likely to be targeted. Grassy and wooded places are the most liked by ticks as hosts feed and live in such places. Larvae counts are to be expected during summer between June and August, while nymphs are found in abundance between February and May, while adult counts appear mainly between February and May.

Chapter 7

Summary, conclusion, limitations and future research

7.1 Summary of the main findings

This dissertation looked at various methods to model the vector of tick life-stage time series of counts collected from random sample sites within three different habitat types namely: edges, grass and woods. The data in this study was collected by the researchers from the Old Dominion University in Virginia, United States, and used in this study as part of the joint NIH-NSF-USDA Ecology and Evolution of Infectious Diseases program.

The main objective for this study was to determine the relationship between *A. americanum* tick abundance and static, temporal and spatial predictors in the southeastern Virginia, United States (US). The data used spanned 2009 through 2018 and was collected at least once a month from eight counties and independent cities in Virginia, US. The counties included the City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County.

Various models were applied to the data in Chapters4, 5 and 6 to investigate the environmental, temporal and spatial effects including sample site random effects. The models were successful in revealing the effects of covariates on the distribution of tick abundance across the eight counties.

In Chapter 4, we introduced multilevel models using the INLA technique and demonstrated its flexibility in allowing for a mixture of marginal count data distributions in the response vector. The study described the model for the vector of time series of counts using the R-INLA technique for fast approximate Bayesian computing compared to the MCMC method. The aim of this study was to unpack estimate trends for the tick life stage abundances and investigate predictors that contribute to higher tick abundance. We also demonstrated the speed differences between the MCMC and the INLA techniques. Throughout our research study period, we did not find similar methods applied to ecological data, specifically tick life stage count data. Results revealed that INLA and MCMC techniques produced similar parameter estimates. However, INLA CPU times were shorter than the MCMC times, and thus the INLA technique was preferred over the MCMC.

The MCMC is a large class of methods that enables inference in highly dimensional problems with unknown quantities and is able to handle complicated distributions. The literature on the methods is well documented in Gamerman and Lopes (2006); Robert et al. (1999); Chen et al. (2012), and these methods are commonly used for computing posterior quantities through the popular Gibbs sampling methods including the Metropolis-Hastings algorithm (Geman and Geman, 1984; Metropolis et al., 1953; Hastings, 1970). However, the MCMC technique is computationally intensive when complex models are involved, or when the data is large. As an alternative to the computationally intensive MCMC Bayesian technique, the INLA technique for approximating posterior parameters was suggested and its feasibil-

ity demonstrated by simulating the data and estimating parameters, while taking into account the time taken to complete in INLA and in MCMC before models were applied to tick data. This study assumed tick life stage specific effects of the covariates on the distribution of tick stages. That is, we assumed that each life stage was affected differently from another tick life stage. The results in this study indicated that the assumption of shared effects was somehow unrealistic.

In Chapter 5 we employed multivariate hierarchical dynamic Poisson and negative binomial count data models to investigate the influence of temporal and location specific effects on the distribution of *A. americanum* abundance in Virginia, United States. The model included the sampling site random effects term to assess the within tick abundance variations. Data simulation was conducted and the model was fitted to ascertain accurate model parameter estimates before it was fitted to tick data. We noted that for small sample sizes the random effects model proposed was susceptible to bias (Grilli et al., 2015). As a result, the model failed to reveal significant fixed effects of the habitat type as credible intervals were wider. However, higher abundance was estimated in summer and this was a significant factor. Consequently, this result confirmed those in the literature Childs and Paddock (2003); Kennedy et al. (2021); Lephoto et al. (2021); Alkishe and Peterson (2022), where life stage abundance of larvae stage ticks peak in August, nymph stage in July and adult stage in June. The evidence of sample site tick abundance variations detected in Chesapeake York, Portsmouth, Northampton, and Norfolk counties suggest the potential disease hotspots. These results confirm the results in (Cohen et al., 2010; Gaines et al., 2014).

The estimate of non-linear effects of the model showed an upward monthly tick abundance trend from January to June, a dip between June and July, steady increase between July and September, a sharp decrease between September and November, and a steady peak from November to December. However, the year

tick abundance overall trend showed a steady decrease between 2009 and 2011, a steep decrease between 2011 and 2012, and a sharp decrease from 2012 through 2018. These trends are well documented in Davidson et al. (1994); Kollars Jr et al. (2000); Mangan et al. (2018) among others.

In Chapter 6 we applied univariate Poisson and NB count data models to tick count data with the aim to explore the influence of environmental and temporal predictors on the distribution of tick counts. We relaxed the assumption that the relationship between the predictors and the response variables in a regression model are constant across the study region and over time. This assumption is unrealistic for spatial processes as factors such as sampling variation and different relationships across regions, for example, attitudes, preferences, culture and others, contribute to different responses to the same stimuli as one moves across regions and over time. Many studies have relaxed this assumption, for example: Assunçao et al. (1999) introduced a Bayesian space-varying parameter model to examine micro-region factor productivity and the degree of factor substitution in Brazilian agriculture; Gamerman et al. (2003) developed a flexible modelling approach for space-varying regression models; and Okango et al. (2015) modelled the HIV and HSV-2 viruses using spatially varying coefficient models. A frequently used approach to spatial modelling dates back to the work by Besag et al. (1991) which was extended by Bernardinelli et al. (1995) to include a linear term for space-time interaction. The Bayesian spatio-temporal process model was used to allow covariates to vary spatially and over time. We specified the CAR prior for the structured random effects; autoregressive of order one (AR1) prior for the temporal random effects; and normal iid priors for the unstructured random effects. Non-linear effects of the month variable on tick counts were also evident. As a result, an assumption of linear relationship would have led to misleading results and consequently to wrong interpretations. The exploratory data analysis showed that more larvae counts were observed compared with nymphs and adult

tick counts. We found that the effects of covariates on tick counts varied spatially across counties and over time. Spatio-temporal models were powerful in the sense that they were capable of revealing county specific effects of each covariate, county space-time effects and the effects of change in time on the distribution of life-stages of Lone Star ticks. We were able to show that tick abundance has been increasing over the study time in Virginia, which confirms the previous results by Lantos et al. (2015) that observed the significant expansion of Lyme disease distribution in Virginia between 2000 and 2014, particularly southward into the Virginia mountain ranges. It is clear that a change in temperature affects tick numbers, such that they decrease in winter (Linske et al., 2020). We confirmed this unsurprising fact and noted also that larvae count remained low up to May. Non-linear effects of month showed that nymphs and adults were observed in spring and summer, meaning that the distribution of larvae would be expected to increase later that summer and thus adults the following spring. This happens in locations in early summer thus determining the distribution of larvae and adults in late summer and spring the following year, respectively (Stein et al., 2008). Our study also revealed that the effects of habitat type were high in Chesapeake County. This could be because of the abundant white-tailed deer population found in forests, farms, parks and backyards throughout the Chesapeake Bay watershed. The Lone Star tick is said to be very aggressive and specific when looking for hosts (Goddard and Varela-Stokes, 2009), but they are unspecific during each life-stage, as this species is found on humans, domesticated animals, ground-dwelling birds as well as on small and large wild animals (Kollars Jr, 1993). The white-tailed deer feeds on fruit and vegetation that are available to them each season, which makes it simple for ticks to attach to and feed from these animals. During warm seasons (summer and spring), these animals feed on green plants, during fall they feed on nuts, acorns and crops and in winter they feed on woody vegetation, such as bark, twigs and buds of hardwood and pine trees, where indeed ticks are found

(Willis et al., 2012).

7.2 Conclusion

The models applied in this dissertation were successful in modeling the time series of tick life stage counts. The models were able to reveal variations within the sampling sites. All models indicated that higher abundances were expected in summer seasons relative to winter seasons. We conclude that variations in seasons affect variations in tick abundances in a life stage specific manner. We also noted considerably higher abundances in the woods compared to the edges. Thus tick abundances depend on being closer to where hosts feed and live, such as grassy and wooded places where deer feed and live. Larvae abundance was higher during summer between June and August, while nymphs were higher in abundance between February and May, and adult abundance was higher between February and May.

7.3 Limitations

This study did not account for excess zeros in the discrete distribution. In future, zero-inflated multivariate models should be developed and applied to data. The incorporation of covariate-time-space interaction effects could also reveal important information on the distribution tick life stage abundance. However, due to limited data, the study only looked at the three habitat types and monthly and seasonal time predictor variables.

7.4 Future research

The work done in this study can be extended to include statistical learning techniques such as supervised and unsupervised learning. These techniques play a

crucial role in many areas of science, finance, and industry among others, and they encompass a diverse range of techniques for understanding data (Vapnik, 1999). A supervised learning technique aims to build a model that predicts the output based on given inputs, whereas the unsupervised learning has inputs with no given outputs (Cunningham et al., 2008; Liu et al., 2011).

In this case, the observed data can be split into a *training* and *testing* portion to predict future trends in each habitat type. Suppose that for $t = 1, \ldots, T$ time points and $i = 1, \ldots, n$ habitat types, $\boldsymbol{Y}_{it} = (Y_{1,it}, \ldots, Y_{n,it})$ denote a $J-$dimensional vector of the observed counts. The data can be split into a *training* portion $\boldsymbol{Y}_{1t}, \ldots, \boldsymbol{Y}_{nt}$, and a *testing* portion $\boldsymbol{Y}_{(n+1)t}, \ldots, \boldsymbol{Y}_{(n+h)t}$, where the *training* and *testing* portions are of length n and $n+h$ respectively. Since R-INLA does not have a separate forecasting function like the `predict()` function in the **stats** package which is commonly used with functions like `lm()`, then we can append the end of the *training* set of length n with NA's to represent the times to holdout the last h data points. In Bayesian dynamic modelling framework, given the *training* data $\boldsymbol{Y}_{1t}, \ldots, \boldsymbol{Y}_{(n-h)t}$, estimates of the states $\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{(n-h)t}$ and estimates of the model hyperparameters $\boldsymbol{\theta}$ we can forecast $\boldsymbol{x}_{(n+h)t}$, $h = 1, \ldots, n$ by $f(\boldsymbol{x}_{(n+h)t}|\boldsymbol{Y}^n)$.

A detailed exposition of hierarchical dynamic non-Gaussian modelling is described in Prado and West (2010), Ravishanker et al. (2022) and Davis et al. (2016) among others. In the Bayesian framework, these models can be compared using the conditional predictive ordinate, Bayes factors, the deviance criterion (DIC), and Watanabe Akaike information criterion (WAIC). However, these may be useful for comparing the relative performance of the fitted models based on the *training* and *testing* data. Models may be compared based on their predictive performance using the frequentist metrics such as the mean absolute percentage error or the mean absolute error, see (Ravishanker et al., 2022).

Another area of extension of the research in this thesis is by considering multivariate spatial models for count data. Suppose $Y_{j,it}$ is the number of ticks for the jth tick life stage in the ith county ($i = 1, \ldots, n$) observed at t time points. The number of ticks in each county, time and life stage stratum, given the mean number $\lambda_{j,it}$ follows a Poisson distribution such that

$$Y_{j,it}|\lambda_{j,it} \sim \text{Poi}(\lambda_{j,it})$$

and the link function is given as

$$\log(\lambda_{j,it}) = \alpha_j + \theta_{ji} + \gamma_{jt} + \delta_{jit}$$

where α_j is an intercept for the jth stage, θ_{ji} and γ_{jt} are the jth spatial and temporal main effects, and δ_{jit} is the spatio-temporal interaction within the jth stage. Suppose $\Theta = \{\theta : i = 1, \ldots, n; j = 1, \ldots, J\}$ and $\Gamma = \{\gamma_{jt} : t = 1, \ldots, T; j = 1, \ldots, J\}$ are two matrices whose columns are the spatial and temporal random effect respectively, and $\Delta = \{\delta_{j,it} : i = 1, \ldots, n; t = 1, \ldots, T\}$ is a matrix capturing the spatio-temporal interaction within each tick life stage. The advantage of multivariate modelling is that dependency between the spatial and temporal patterns of the different tick life stages can be included in the model so that the latent association between the life stages can help to improve the estimates and to discover factors related to the phenomena being studied.

To understand how dependence between the spatial counties and between the global temporal trends of the different life stages are included in the model, let Θ and Γ be expressed as

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}_\theta \boldsymbol{M}_{\boldsymbol{\theta}},$$

and

$$\boldsymbol{\Gamma} = \boldsymbol{\Phi}_{\gamma}\boldsymbol{M}_{\boldsymbol{\gamma}},$$

where $\boldsymbol{\Phi}_{\theta}$ and $\boldsymbol{\Phi}_{\gamma}$ are the random effects matrices of order $I \times K_{\theta}$ and $T \times K_{\gamma}$ whose columns are distributed independently following a spatially correlated distribution and a temporally correlated distribution respectively. The matrices $\boldsymbol{M}_{\theta}$ and $\boldsymbol{M}_{\gamma}$, of orders $K_{\theta} \times J$ and $K_{\gamma} \times J$, are responsible for inducing dependence between the different columns of $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ respectively. More precisely, dependence between the columns of $\boldsymbol{\Theta}$ means correlation between spatial patterns of the life stages, whereas the dependence between their rows indicates spatial correlation within life stages. Similarly, dependence between columns of $\boldsymbol{\Gamma}$ means correlation between the temporal patterns of the life stages, and dependence between rows leads to temporal correlation within life stages, see Vicente et al. (2020).

References

Adams, D. A., Gallagher, K. M., Jajosky, R. A., Kriseman, J., Sharp, P., Anderson, W. J., Aranas, A. E., Mayes, M., Wodajo, M. S., Onweh, D. H., et al. (2013). Summary of notifiable diseases–united states, 2011. *Morbodity and Mortality Weekly Report*, 60(53):1–17.

Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Alfò, M. and Trovato, G. (2004). Semiparametric mixture models for multivariate count data, with application. *The Econometrics Journal*, 7(2):426–454.

Alkishe, A. and Peterson, A. T. (2022). Climate change influences on the geographic distributional potential of the spotted fever vectors amblyomma maculatum and dermacentor andersoni. *PeerJ*, 10:e13279.

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health*, 12(9):10536–10548.

Assunçao, J., Gamerman, D., and Assunçao, R. (1999). Regional differences in factor productivities of Brazilian agriculture: a space varying parameter approach. In *Proceedings of the XV Latin American Meeting of the Econometric Society*.

Barker, R. J. and Sauer, J. R. (1992). Modelling population change from time series data. In *Wildlife 2001: populations*, pages 182–194. Springer.

Bates, D., Chambers, J., and Hastie, T. (1992). Statistical models in s. In *Computer science and statistics: proceedings of the 19th Symposium on the Interface*. Wadsworth & Brooks.

Benham, S. A., Gaff, H. D., Bement, Z. J., Blaise, C., Cummins, H. K., Ferrara, R., Moreno, J., Parker, E., Phan, A., Rose, T., et al. (2021). Comparative population genetics of Amblyomma maculatum and Amblyomma americanum in the mid-Atlantic United States. *Ticks and Tick-borne Diseases*, 12(1):101600.

Bermúdez, L. and Karlis, D. (2021). Multivariate INAR (1) regression models based on the Sarmanov distribution. *Mathematics*, 9(5):505.

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14(21-22):2411–2431.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Cameron, A. C. and Johansson, P. (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics*, 12(3):203–223.

Cameron, A. C. and Trivedi, P. K. (2001). Essentials of count data regression. *A Companion to Theoretical Econometrics*, pages 331–348.

CDC (2020). Guide to the surveillance of metastriate ticks (Acari: Ixodidae) and their pathogens in the United States.

CDC (2021). National notifiable diseases surveillance system, 2019 annual tables of infectious disease data.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Chib, S., Greenberg, E., and Winkelmann, R. (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, 86(1):33–54.

Chib, S. and Winkelmann, R. (2001). Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435.

Childs, J. E. and Paddock, C. D. (2003). The ascendancy of Amblyomma americanum as a vector of pathogens affecting humans in the United States. *Annual Review of Entomology*, 48(1):307–337.

Chiquet, J., Mariadassou, M., and Robin, S. (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698.

Chiquet, J., Mariadassou, M., and Robin, S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:588292.

Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR.

Christou, V. and Fokianos, K. (2014). Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, 35(1):55–78.

Clark, N. J. and Wells, K. (2023). Dynamic generalised additive models (DGAMs) for forecasting discrete ecological time series. *Methods in Ecology and Evolution*, 14(3):771–784.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*. Taylor & Francis.

Cohen, S. B., Yabsley, M. J., Freye, J. D., Dunlap, B. G., Rowland, M. E., Huang, J., Dunn, J. R., Jones, T. F., and Moncayo, A. C. (2010). Prevalence of Ehrlichia chaffeensis and Ehrlichia ewingii in ticks from Tennessee. *Vector-Borne and Zoonotic Diseases*, 10(5):435–440.

Cunningham, P., Cord, M., and Delany, S. J. (2008). *Supervised learning, machine learning techniques for multimedia*. Springer-Verlag.

Davidson, W. R., Siefken, D. A., and Creekmore, L. H. (1994). Seasonal and annual abundance of Amblyomma americanum (Acari: Ixodidae) in central Georgia. *Journal of Medical Entomology*, 31(1):67–71.

Davis, R. A., Dunsmuir, W. T., and Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika*, 90(4):777–790.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N. (2016). *Handbook of discrete-valued time series*. CRC Press.

Davis, R. A. and Liu, H. (2012). Theory and inference for a class of observation-driven models with application to time series of counts. *arXiv preprint arXiv:1204.3915*.

De Clercq, E., Leta, S., Estrada-Peña, A., Madder, M., Adehan, S., and Vanwambeke, S. O. (2015). Species distribution modelling for Rhipicephalus mi-

croplus (Acari: Ixodidae) in Benin, West Africa: comparing datasets and modelling algorithms. *Preventive Veterinary Medicine*, 118(1):8–21.

de Oliveira, S. V., Romero-Alvarez, D., Martins, T. F., Dos Santos, J. P., Labruna, M. B., Gazeta, G. S., Escobar, L. E., and Gurgel-Gonçalves, R. (2017). Amblyomma ticks and future climate: range contraction due to climate warming. *Acta Tropica*, 176:340–348.

Douc, R., Fokianos, K., and Moulines, E. (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electronic Journal of Statistics*, 11(2):2707–2740.

Dunsmuir, W. T. and Scott, D. J. (2015). The glarma package for observation-driven time series regression of counts. *Journal of Statistical Software*, 67:1–36.

Estrada-Peña, A., de la Fuente, J., and Cabezas-Cruz, A. (2016). A comparison of the performance of regression models of Amblyomma americanum (L.)(Ixodidae) using life cycle or landscape data from administrative divisions. *Ticks and Tick-borne Diseases*, 7(4):624–630.

Estrada-Pena, A. and Venzal, J. M. (2007). Climate niches of tick species in the Mediterranean region: modeling of occurrence data, distributional constraints, and impact of climate change. *Journal of Medical Entomology*, 44(6):1130–1138.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220.

Famoye, F. (2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6):969–981.

Famoye, F. (2015). A multivariate generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 44(3):497–511.

Fan, X., Ma, R., Yue, C., Liu, J., Yue, B., Yang, W., Li, Y., Gu, J., Ayala, J. E., Bunker, D. E., et al. (2023). A snapshot of climate drivers and temporal variation of Ixodes ovatus abundance from a giant panda living in the wild. *International Journal for Parasitology: Parasites and Wildlife*, 20:162–169.

Ferkingstad, E., Held, L., and Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat*, 6(1):331–344.

Fokianos, K. (2015). *Handbook of discrete-valued time series, handbooks of modern statistical methods*. Chapman & Hall.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Fruehwirth-Schnatter, S. and Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, 51(7):3509–3528.

Gaines, D. N., Operario, D. J., Stroup, S., Stromdahl, E., Wright, C., Gaff, H., Broyhill, J., Smith, J., Norris, D. E., Henning, T., et al. (2014). Ehrlichia and spotted fever group Rickettsiae surveillance in Amblyomma americanum in Virginia through use of a novel six-plex real-time PCR assay. *Vector-Borne and Zoonotic Diseases*, 14(5):307–316.

Gamerman, D. and Lopes, H. (1997). *Markov chain Monte Carlo (texts in statistical science)*. Chapman & Hall.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gamerman, D., Moreira, A. R., and Rue, H. (2003). Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis*, 42(3):513–533.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Goddard, J. and Varela-Stokes, A. S. (2009). Role of the Lone Star tick, Amblyomma americanum (L.), in human and animal diseases. *Veterinary Parasitology*, 160(1-2):1–12.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83(2):255–279.

Grilli, L., Metelli, S., and Rampichini, C. (2015). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 85(13):2718–2726.

Hadfield, J. (2010). MCMCglmm: Markov chain Monte Carlo methods for generalised linear mixed models. *Tutorial for MCMCglmm package in R*, 125.

Hair, J. A. and Howell, D. E. (1970). *Lone Star ticks: their biology and control in Ozark recreation areas*. Oklahoma Agricultural Experiment Station, Volume 679 of Bulletin B.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Inger, R., Gregory, R., Duffy, J. P., Stott, I., Voříšek, P., and Gaston, K. J. (2015). Common european birds are declining rapidly while less abundant species' numbers are rising. *Ecology Letters*, 18(1):28–36.

Johnson, N. and Kotz, S. (1969). *Distributions in statistics: discrete distributions*. Hoghton Mifflin Harcout.

Kaizer, A., Foré, S., Kim, H.-J., and York, E. (2015). Modeling the biotic and abiotic factors that describe the number of active off-host Amblyomma americanum larvae. *Journal of Vector Ecology*, 40(1):1–10.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77.

Karlis, D. and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing*, 15(4):255–265.

Karlis, D. and Meligkotsidou, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference*, 137(6):1942–1960.

Kennedy, A. C., Marshall, E., et al. (2021). Lone Star ticks (amblyomma americanum): an emerging threat in Delaware. *Delaware Journal of Public Health*, 7(1):66.

Khaemba, W. M. and Stein, A. (2001). Spatial statistics for modeling of abundance and distribution of wildlife species in the Masai Mara ecosystem, Kenya. *Environmental and Ecological Statistics*, 8(4):345–360.

Khana, D., Rossen, L. M., Hedegaard, H., and Warner, M. (2018). A Bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-INLA. *Journal of Data Science*, 16(1):147.

Kjær, L. J., Soleng, A., Edgar, K. S., Lindstedt, H. E. H., Paulsen, K. M., Andreassen, Å. K., Korslund, L., Kjelland, V., Slettan, A., Stuen, S., et al. (2019). Predicting and mapping human risk of exposure to Ixodes ricinus nymphs using climatic and environmental data, Denmark, Norway and Sweden, 2016. *Eurosurveillance*, 24(9):1800101.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Kollars Jr, T. M. (1993). Ticks (Acari: Ixodidae) infesting medium-sized wild mammals in southwestern Tennessee. *Journal of Medical Entomology*, 30(5):896–900.

Kollars Jr, T. M., Oliver Jr, J. H., Durden, L. A., and Kollars, P. G. (2000). Host associations and seasonal activity of Amblyomma americanum (Acari: Ixodidae) in Missouri. *Journal of Parasitology*, 86(5):1156–1159.

Kopsco, H. L., Smith, R. L., and Halsey, S. J. (2022). A scoping review of species distribution modeling methods for tick vectors. *Frontiers in Ecology and Evolution*, 10:893016.

Kuhnert, P. M., Martin, T. G., Mengersen, K., and Possingham, H. P. (2005). Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environmetrics*, 16(7):717–747.

Laaksonen, M., Sajanti, E., Sormunen, J. J., Penttinen, R., Hänninen, J., Ruohomäki, K., Sääksjärvi, I., Vesterinen, E. J., Vuorinen, I., Hytönen, J., et al. (2017). Crowdsourcing-based nationwide tick collection reveals the distribution of Ixodes ricinus and I. persulcatus and associated pathogens in Finland. *Emerging Microbes & Infections*, 6(1):1–7.

Lado, P., Smith, M. L., Carstens, B. C., and Klompen, H. (2020). Population genetic structure and demographic history of the Lone Star tick, Amblyomma

americanum (Ixodida: Ixodidae): new evidence supporting old records. *Molecular Ecology*, 29(15):2810–2823.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lantos, P. M., Nigrovic, L. E., Auwaerter, P. G., Fowler Jr, V. G., Ruffin, F., Brinkerhoff, R. J., Reber, J., Williams, C., Broyhill, J., Pan, W. K., et al. (2015). Geographic expansion of lyme disease in the southeastern United States, 2000–2014. In *Open forum infectious diseases*, volume 2, page 143. Oxford University Press.

Lephoto, T., Mwambi, H., Bodhlyera, O., and Gaff, H. (2021). Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients. *Geospatial Health*, 16(2).

Linske, M. A., Williams, S. C., Stafford, K. C., Lubelczyk, C. B., Henderson, E. F., Welch, M., and Teel, P. D. (2020). Determining effects of winter weather conditions on adult Amblyomma americanum (Acari: Ixodidae) survival in Connecticut and Maine, USA. *Insects*, 11(1):13.

Lippi, C. A., Gaff, H. D., White, A. L., St. John, H. K., Richards, A. L., and Ryan, S. J. (2021). Exploring the niche of Rickettsia montanensis (Rickettsiales: Rickettsiaceae) infection of the american dog tick (Acari: Ixodidae), using multiple species distribution model approaches. *Journal of Medical Entomology*, 58(3):1083–1092.

Liu, B. et al. (2011). *Web data mining: exploring hyperlinks, contents, and usage data*, volume 1. Springer.

Liu, Y., Lund, R. B., Nordone, S. K., Yabsley, M. J., and McMahan, C. S. (2017). A Bayesian spatio-temporal model for forecasting the prevalence of

antibodies to Ehrlichia species in domestic dogs within the contiguous United States. *Parasites & Vectors*, 10(1):1–14.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press.

Lynen, G., Zeman, P., Bakuname, C., Di Giulio, G., Mtui, P., Sanka, P., and Jongejan, F. (2008). Shifts in the distributional ranges of boophilus ticks in tanzania: evidence that a parapatric boundary between boophilus microplus and b. decoloratus follows climate gradients. *Experimental and Applied Acarology*, 44:147–164.

Ma, D., Lun, X., Li, C., Zhou, R., Zhao, Z., Wang, J., Zhang, Q., and Liu, Q. (2021). Predicting the potential global distribution of Amblyomma americanum (Acari: Ixodidae) under near current and future climatic conditions, using the maximum entropy model. *Biology*, 10(10):1057.

Ma, J., Kockelman, K. M., and Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, 40(3):964–975.

Mahamunulu, D. (1967). A note on regression in the multivariate Poisson distribution. *Journal of the American Statistical Association*, 62(317):251–258.

Manda, S. O. and Leyland, A. (2007). An empirical comparison of maximum likelihood and Bayesian estimation methods for multivariate disease mapping: theory and methods. *South African Statistical Journal*, 41(1):1–21.

Mangan, M. J., Foré, S. A., and Kim, H.-J. (2018). Ecological modeling over seven years to describe the number of host-seeking Amblyomma americanum in each life stage in northeast Missouri. *Journal of Vector Ecology*, 43(2):271–284.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology:

improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8(11):1235–1246.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall.

McKenzie, E. (1985). Some simple models for discrete variate time series 1. *Journal of the American Water Resources Association*, 21(4):645–650.

Merten, H. A., Durden, L. A., et al. (2000). A state-by-state survey of ticks recorded from humans in the United States. *Journal of Vector Ecology*, 25(1):102–113.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Molaei, G., Little, E. A., Williams, S. C., and Stafford, K. C. (2019). Bracing for the worst-range expansion of the Lone Star tick in the northeastern United States. *New England Journal of Medicine*, 381(23):2189–2192.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Montesinos-López, J. C., Singh, P., Juliana, P., and Salinas-Ruiz, J. (2017). A Bayesian poisson-lognormal model for count data for multiple-trait multiple-environment genomic-enabled prediction. *G3: Genes, Genomes, Genetics*, 7(5):1595–1606.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Mutizhe, S. W., Mhlanga, L., Sithole, R., Maya, B. T., Sibanda, A., and Mpofu, P. (2022). Spatio-temporal variation in tick community composition and abundance in a wildlife–livestock interface within Nyanga National Park, Zimbabwe. *African Journal of Ecology*, 60(3):607–620.

Nadolny, R. M. and Gaff, H. D. (2018). Natural history of Amblyomma maculatum in Virginia. *Ticks and Tick-borne Diseases*, 9(2):188–195.

Nadolny, R. M., Wright, C. L., Sonenshine, D. E., Hynes, W. L., and Gaff, H. D. (2014). Ticks and spotted fever group rickettsiae of southeastern Virginia. *Ticks and Tick-borne Diseases*, 5(1):53–57.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Neupane, N., Goldbloom-Helzner, A., and Arab, A. (2021). Spatio-temporal modeling for confirmed cases of lyme disease in virginia. *Ticks and Tick-borne Diseases*, 12(6):101822.

Ngesa, O., Mwambi, H., and Achia, T. (2014). Bayesian spatial semi-parametric modeling of HIV variation in Kenya. *PloS One*, 9(7):e103299.

Oakes, V. J., Yabsley, M. J., Schwartz, D., LeRoith, T., Bissett, C., Broaddus, C., Schlater, J. L., Todd, S. M., Boes, K. M., Brookhart, M., et al. (2019). Theileria orientalis Ikeda genotype in cattle, Virginia, USA. *Emerging Infectious Diseases*, 25(9):1653.

Okango, E., Mwambi, H., Ngesa, O., and Achia, T. (2015). Semi-parametric spatial joint modeling of HIV and HSV-2 among women in Kenya. *PloS One*, 10(8):e0135212.

Otálora-Luna, F., Dickens, J. C., Brinkerhoff, J., and Li, A. Y. (2022). Geotropic, hydrokinetic and random walking differ between sympatric tick species: the deer tick Ixodes scapularis and the Lone Star tick Ambylomma americanum. *Journal of Ethology*, 40(2):133–143.

Pasternak, A. R. and Palli, S. R. (2023). County-level surveillance for the lone star tick, Amblyomma americanum, and its associated pathogen, Ehrlichia chaffeensis, in Kentucky. *Ticks and Tick-borne Diseases*, 14(1):102072.

Paull, S. H., Thibault, K. M., and Benson, A. L. (2022). Tick abundance, diversity and pathogen data collected by the national ecological observatory network. *Gigabyte*, 2022:1–11.

Pedeli, X. and Karlis, D. (2011). A bivariate INAR (1) process with application. *Statistical Modelling*, 11(4):325–349.

Prado, R. and West, M. (2010). *Time series: modeling, computation, and inference*. Chapman and Hall/CRC.

QGIS, D. T. (2009). QGIS geographic information system, open source geospatial foundation.

Qviller, L., Viljugrein, H., Loe, L. E., Meisingset, E. L., and Mysterud, A. (2016). The influence of red deer space use on the distribution of Ixodes ricinus ticks in the landscape. *Parasites & Vectors*, 9(1):1–9.

R Core Team (2022). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghavan, R., Barker, S., Cobos, M. E., Barker, D., Teo, E., Foley, D., Nakao, R., Lawrence, K., Heath, A., and Peterson, A. T. (2019). Potential spatial distribution of the newly introduced long-horned tick, Haemaphysalis longicornis in North America. *Scientific Reports*, 9(1):498.

Rau, A., Munoz-Zanzi, C., Schotthoefer, A. M., Oliver, J. D., and Berman, J. D. (2020). Spatio-temporal dynamics of tick-borne diseases in north-central Wisconsin from 2000–2016. *International Journal of Environmental Research and Public Health*, 17(14):5105.

Ravishanker, N., Raman, B., and Soyer, R. (2022). *Dynamic time series models using R-INLA: an applied perspective*. CRC Press.

Requena-García, F., Cabrero-Sañudo, F., Olmeda-García, S., González, J., and

Valcárcel, F. (2017). Influence of environmental temperature and humidity on questing ticks in central Spain. *Experimental and Applied Acarology*, 71:277–290.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Rochlin, I., Egizi, A., and Ginsberg, H. S. (2023). Modeling of historical and current distributions of Lone Star tick, Amblyomma americanum (Acari: Ixodidae), is consistent with ancestral range recovery. *Experimental and Applied Acarology*, 89(1):85–103.

Rochlin, I., Egizi, A., and Lindström, A. (2022). The original scientific description of the Lone Star tick (Amblyomma americanum, Acari: Ixodidae) and implications for the species' past and future geographic distributions. *Journal of Medical Entomology*, 59(2):412–420.

Rodrigues, A. S., Pilgrim, J. D., Lamoreux, J. F., Hoffmann, M., and Brooks, T. M. (2006). The value of the IUCN red list for conservation. *Trends in Ecology & Evolution*, 21(2):71–76.

Rosà, R., Andreo, V., Tagliapietra, V., Baráková, I., Arnoldi, D., Hauffe, H. C., Manica, M., Rosso, F., Blaňarová, L., Bona, M., et al. (2018). Effect of climate and land use on the spatio-temporal variability of tick-borne bacteria in Europe. *International Journal of Environmental Research and Public Health*, 15(4):732.

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

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference

for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sagurova, I., Ludwig, A., Ogden, N. H., Pelcat, Y., Dueymes, G., and Gachon, P. (2019). Predicted northward expansion of the geographic range of the tick vector Amblyomma americanum in North America under future climate conditions. *Environmental Health Perspectives*, 127(10):107014.

Sattler, T., Bontadina, F., Hirzel, A. H., and Arlettaz, R. (2007). Ecological niche modelling of two cryptic bat species calls for a reassessment of their conservation status. *Journal of Applied Ecology*, 44(6):1188–1199.

Savage, H. M., Burkhalter, K. L., Godsey Jr, M. S., Panella, N. A., Ashley, D. C., Nicholson, W. L., and Lambert, A. J. (2017). Bourbon virus in field-collected ticks, Missouri, USA. *Emerging Infectious Diseases*, 23(12).

Savage, H. M., Godsey Jr, M. S., Lambert, A., Panella, N. A., Burkhalter, K. L., Harmon, J. R., Lash, R. R., Ashley, D. C., and Nicholson, W. L. (2013). First detection of heartland virus (Bunyaviridae: Phlebovirus) from field collected arthropods. *The American Journal of Tropical Medicine and Hygiene*, 89(3):445.

Scheiner, S. M. and Willig, M. R. (2011). *The theory of ecology*. University of Chicago Press.

Serhiyenko, V. (2015). *Dynamic modeling of multivariate counts-fitting, diagnostics, and applications*. PhD thesis, University of Connecticut, Storrs, CT, USA.

Serhiyenko, V., Ravishanker, N., and Venkatesan, R. (2018). Multi-stage multivariate modeling of temporal patterns in prescription counts for competing drugs in a therapeutic category. *Applied Stochastic Models in Business and Industry*, 34(1):61–78.

Simpson, D. T., Teague, M. S., Weeks, J. K., Kaup, B. Z., Kerscher, O., and Leu, M. (2019). Habitat amount, quality, and fragmentation associated with prevalence of the tick-borne pathogen Ehrlichia chaffeensis and occupancy dynamics of its vector, Amblyomma americanum. *Landscape Ecology*, 34:2435–2449.

Sonenshine, D. E. (1979). Insects of Virginia no. 13. *Ticks of Virginia (Acari: Metastigmata). Research Division Bulletin*, 139.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Stein, K. J., Waterman, M., and Waldon, J. L. (2008). The effects of vegetation density and habitat disturbance on the spatial distribution of Ixodid ticks (Acari: Ixodidae). *Geospatial Health*, 2(2):241–252.

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability*, 7(5):893–899.

Sutherland, W. J., Pullin, A. S., Dolman, P. M., and Knight, T. M. (2004). The need for evidence-based conservation. *Trends in Ecology & Evolution*, 19(6):305–308.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, pages 1701–1728.

Tjøstheim, D. (2012). Some recent theory for autoregressive count time series. *Test*, 21(3):413–438.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for Bayesian inference with INLA. *Computational Statistics & Data Analysis*, 181:107–692.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer.

Vicente, G., Goicoa, T., and Ugarte, M. (2020). Bayesian inference in multivariate spatio-temporal areal models using INLA: analysis of gender-based violence in small areas. *Stochastic Environmental Research and Risk Assessment*, 34:1421–1440.

Wang, F., Li, H., Wang, H., and Li, Y. (2023). Spatial correlated incidence modeling with zero inflation. *Biometrical Journal*, 65(4):2200090.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts-a survey. *Advances in Statistical Analysis*, 92(3):319–341.

Willis, D., Carter, R., Murdock, C., and Blair, B. (2012). Relationship between habitat type, fire frequency, and Amblyomma americanum populations in east-central Alabama. *Journal of Vector Ecology*, 37(2):373–381.

Wimberly, M. C., Baer, A. D., and Yabsley, M. J. (2008). Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*, 7(1):1–14.

Wimms, C., Aljundi, E., and Halsey, S. J. (2023). Regional dynamics of tick vectors of human disease. *Current Opinion in Insect Science*, 55:101006.

Zannou, O. M., Ouedraogo, A. S., Biguezoton, A. S., Abatih, E., Coral-Almeida, M., Farougou, S., Yao, K. P., Lempereur, L., and Saegerman, C. (2021). Models for studying the distribution of ticks and tick-borne diseases in animals: A systematic review and a meta-analysis with a focus on africa. *Pathogens*, 10(7):893.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75(4):621–629.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8):1–25.

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.

Zipkin, E. F. (2012). *Hierarchical models for the anlaysis of species distributions and abundances: development and applications*. PhD thesis, University of Mayland, College Park, MD, USA.

Typeset using LaTeX

Appendix A

Simulation R-code for chapter 4

A.1 R code for simulations in section 4.4

```
##### libraries ####
library(MASS)
###### function to simulate LCM (with Poisson marginals) #####
sim.LCM=function(N=N,log.lambda,level.sigma){
  log.lambda=as.matrix(log.lambda)
  level.sigma=as.matrix(level.sigma)
  J=dim(log.lambda)[1]
  log.mean=matrix(rep(log.lambda,N),ncol=J)
  # simulate level alpha
  level.mu=rep(0,J)
  alpha=mvrnorm(n=N, mu=level.mu, Sigma=level.sigma)
  # simulate counts

  new.mean=alpha+log.mean
  mean=exp(new.mean)
```

```
  counts=matrix(rep(NA,N*J),ncol=J)
  for (j in 1:J) {
    counts[,j]=rpois(n=N, lambda=mean[,j])
  }
  results=list(counts=counts,alpha=alpha,mean=mean)
  return(results)
}
#A.2 Estimation of LCMs using R-INLA
###### libraries ####
library(INLA)
library(MCMCglmm)
library(tidyverse)
###### This section gives R code used for
###### simulation and estimation of LCMs
###### in Section 4.3.1
###### Results are given in
###### Tables 4.1, 4.2, and 4.3
###### Input for simulations ########
###### level.sigma
rho=-0.95
sigma11=0.4
sigma22=0.4
sigma12=rho*sqrt(sigma11*sigma22)
level.sigma.true=matrix(c(sigma11,sigma12,sigma12,sigma22),ncol=2)
###### log.lambda

mean=c(10,10)
log.mean=log(mean)
```

```
####################################
###### Simulation of counts
####################################
### Sample size n
num=c( 100, 500, 5000, 10000 )


out = c()
out[k]=c()


out.corr1=c()
out.corr2=c()
out.corr3=c()
out.corr4=c()


out.inla1=c()
out.inla2=c()
out.inla3=c()
out.inla4=c()


time.inla1=c()
time.inla2=c()
time.inla3=c()
time.inla4=c()


#fit.mcmc=list()
num.sim=length(num)
### Loop for n
for (k in 1:num.sim) {
```

```
N=num[k]
set.seed(1234567)
sim.data=sim.LCM(N=N,
                 log.lambda=log.mean,
                 level.sigma=level.sigma.true)
dim(sim.data$counts)
# True correlation
sigma=sigma11
sig12n=sigma12
corr.sim=(exp(sig12n)-1)/(exp(sigma)-1+
                            (mean[1]*exp(sigma/2))^(-1))
# Pearson Correlation
corr.p=cor.test(x=sim.data$counts[,1],y=sim.data$counts[,2],
                method = "pearson",
                alternative = "two.sided",conf.level=0.95)

  # Kendall's tau
corr.k=cor.test(x=sim.data$counts[,1],y=sim.data$counts[,2],
                method = "kendall",
                alternative = "two.sided",conf.level=0.95)

######### INLA set-up ############
y=matrix( NA, 2*N , 2 )
y[, 1] = c( sim.data$counts[ 1:N ], rep( NA, N ) )
y[, 2] = c( rep( NA, N ), sim.data$counts[ (N+1):(2*N)  ])
#y=c(sim.data$counts[,1],sim.data$counts[,2])
N= dim(sim.data$counts)[1]
N.all=2*N
```

```
index=1:N.all
n=N.all/2

int1=c(rep(1,n),rep(0,n))
int2=c(rep(0,n),rep(1,n))

 formula= y ~ int1+int2+f(index, model="iid2d", n=N.all)-1


 result4 = inla(formula,family= c( "poisson", "nbinomial"  ),
             data =data.frame(y,index,int1,int2),
             control.compute=list(dic=TRUE))

 result3 = inla(formula,family= c( "nbinomial", "poisson"  ),
                data =data.frame(y,index,int1,int2),
                control.compute=list(dic=TRUE))

 result2 = inla(formula,family= c( "nbinomial", "nbinomial"  ),
                data =data.frame(y,index,int1,int2),
                control.compute=list(dic=TRUE))

 result1 = inla(formula,family= c( "poisson", "poisson"  ),
                data =data.frame(y,index,int1,int2),
                control.compute=list(dic=TRUE))

 #summary(result)
# Results from INLA
t1=result1$summary.hyperpar[, c(1,3,5) ]
```

```
t1[1,]=1/t1[1,]

t1[2,]=1/t1[2,]

t2=exp(result1$summary.fixed[,c(1,3,5)])

hyper=rbind(t1,t2)

M.inla=exp(result1$summary.fixed[1,1])

M.inla.upp=exp(result1$summary.fixed[1,5])

M.inla.low=exp(result1$summary.fixed[1,3])

rho.inla=result1$summary.hyperpar[3,1]

sig11=1/result1$summary.hyperpar[1,1]

sig22=1/result1$summary.hyperpar[2,1]

sig12=rho.inla*sqrt(sig11*sig22)


corr.inla1=(exp(sig12)-1)/sqrt((exp(sig11)-1+

           (M.inla*exp(sig11/2))^(-1))*(exp(sig22)-1+

           (M.inla*exp(sig22/2))^(-1)))


t3=result2$summary.hyperpar[-c(1,2),c(1,3,5)]

t3[1,]=1/t3[1,]

t3[2,]=1/t3[2,]

t4=exp(result2$summary.fixed[,c(1,3,5)])

hyper2=rbind(t3,t4)

M.inla2=exp(result2$summary.fixed[1,1])

M.inla.upp2=exp(result2$summary.fixed[1,5])

M.inla.low2=exp(result2$summary.fixed[1,3])

rho.inla2=result2$summary.hyperpar[5,1]

sig112=1/result2$summary.hyperpar[3,1]

sig222=1/result2$summary.hyperpar[4,1]

sig122=rho.inla2*sqrt(sig112*sig222)
```

```
corr.inla2=(exp(sig122)-1)/sqrt((exp(sig112)-1+
           (M.inla2*exp(sig112/2))^(-1))*(exp(sig222)-1+
           (M.inla2*exp(sig222/2))^(-1)))

t5=result3$summary.hyperpar[-1,c(1,3,5)]
t5[1,]=1/t5[1,]
t5[2,]=1/t5[2,]
t6=exp(result3$summary.fixed[,c(1,3,5)])
hyper3=rbind(t5,t6)
M.inla3=exp(result3$summary.fixed[1,1])
M.inla.upp3=exp(result3$summary.fixed[1,5])
M.inla.low3=exp(result3$summary.fixed[1,3])
rho.inla3=result3$summary.hyperpar[4,1]
sig113=1/result3$summary.hyperpar[2,1]
sig223=1/result3$summary.hyperpar[3,1]
sig123=rho.inla3*sqrt(sig113*sig223)

corr.inla3=(exp(sig123)-1)/sqrt((exp(sig113)-1+
           (M.inla3*exp(sig113/2))^(-1))*(exp(sig223)-1+
           (M.inla3*exp(sig223/2))^(-1)))

t7=result4$summary.hyperpar[-1,c(1,3,5)]
t7[1,]=1/t7[1,]
t7[2,]=1/t7[2,]
t8=exp(result4$summary.fixed[,c(1,3,5)])
hyper4=rbind(t7,t8)
M.inla4=exp(result4$summary.fixed[1,1])
```

```
M.inla.upp4=exp(result4$summary.fixed[1,5])
M.inla.low4=exp(result4$summary.fixed[1,3])
rho.inla4=result4$summary.hyperpar[4,1]
sig114=1/result4$summary.hyperpar[2,1]
sig224=1/result4$summary.hyperpar[3,1]
sig124=rho.inla4*sqrt(sig114*sig224)

corr.inla4=(exp(sig124)-1)/sqrt((exp(sig114)-1+
           (M.inla4*exp(sig114/2))^(-1))*(exp(sig224)-1+
           (M.inla4*exp(sig224/2))^(-1)))

##### Set up output ########
out=data.frame(N,corr.p$estimate,
               corr.k$estimate,
               corr.inla1,
               corr.inla2,
               corr.inla3,
               corr.inla4,
               corr.sim)
out.corr1=rbind(out.corr1,out)
outA=data.frame(N,hyper)
out.inla1=rbind(out.inla1,outA)

out2=data.frame(N,corr.p$estimate,
                 corr.k$estimate,
                 corr.inla,
                 corr.inla2,
                 corr.inla3,
```

```
                corr.inla4,
                corr.sim)
out.corr2=rbind(out.corr2,out2)
outB=data.frame(N,hyper2)
out.inla2=rbind(out.inla2,outB)

out3=data.frame(N,corr.p$estimate,
                corr.k$estimate,
                corr.inla,
                corr.inla2,
                corr.inla3,
                corr.inla4,
                corr.sim)
 out.corr3=rbind(out.corr3,out3)
 outC=data.frame(N,hyper3)
 out.inla3=rbind(out.inla3,outC)

 out4=data.frame(N,corr.p$estimate,
                corr.k$estimate,
                corr.inla,
                corr.inla2,
                corr.inla3,
                corr.inla4,
                corr.sim)
 out.corr4=rbind(out.corr4,out4)
 outD=data.frame(N,hyper4)
 out.inla4=rbind(out.inla4,outD)
```

```
  ####  Set up output

  time.inla1=c(time.inla1,result1$cpu.used[4])
  time.inla2=c(time.inla2,result2$cpu.used[4])
  time.inla3=c(time.inla3,result3$cpu.used[4])
  time.inla4=c(time.inla4,result4$cpu.used[4])

  print(paste("Iteration number:", k))
  print(paste("Time INLA 1:",result1$cpu.used[4]))
  print(paste("Time INLA 2:",result2$cpu.used[4]))
  print(paste("Time INLA 3:",result3$cpu.used[4]))
  print(paste("Time INLA 4:",result4$cpu.used[4]))
}

# Correlation Table and Latex code:
library(xtable)
# 1. Table

out.corr1

# 2. Latex code
xtable( out.corr, digits = 3 )



t1=c(sigma11,sigma22,rho,mean[1],mean[2])
true=rep(t1,4)
out.inla1=data.frame(out.inla1,true);out.inla1
```

```
# Get Table for time (seconds) taken by INLA

t2=data.frame(N=out.corr1$N, time.inla1,time.inla2,
              time.inla3,time.inla4)

xtable(t2)



###### Output Tables
## Create data frames

out.inlaA = data.frame(out.inla1)
out.inlaB = data.frame(out.inla2)
out.inlaC = data.frame(out.inla3)
out.inlaD = data.frame(out.inla4,true)

w1 = round( out.inlaA, 3 )
w2 = round( out.inlaB, 3 )
w3 = round( out.inlaC, 3 )
w4 = round( out.inlaD, 3 )

f_table = cbind( w1[,-5], w2[,-1], w3[,-1], w4[,-1])

f_table = as.data.frame(f_table)

# Rearrange credible intervals for sigma11 and sigma22

f_table1 = f_table[c(1,2,6,7,11,12,16,17),
```

```
                    c(1,2,4,3,5,7,6,8,10,9,11,13,12,14)]

# Pull out rho, M-1 and M-2:

f_table2 =f_table[c(3,4,5,8,9,10,13,14,15,18,19,20),]

# Get latex code for the table

xtable::xtable(f_table1)
xtable::xtable(f_table2)



# View and get latex code for correlation's table

corr.table = round( out.corr1, digits = 3 ); corr.table

xtable::xtable(corr.table, digits = 3)
```

A.2 R code for computing MSEs

```
###### libraries ####
library( MASS )
library( xtable )

###### function to simulate LCM (with Poisson marginals) #####
```

```
sim.LCM = function( N = N , log.lambda , level.sigma )
  {
  level.sigma = as.matrix( level.sigma )
   log.lambda = as.matrix( log.lambda )
            J = dim( log.lambda )[1]
     log.mean = matrix(rep( log.lambda , N ), ncol = J )

   # simulate level alpha

  level.mu = rep( 0 , J )
     alpha = mvrnorm( n = N , mu = level.mu , Sigma = level.sigma)

  # simulate counts

  new.mean = alpha+log.mean
      mean = exp( new.mean )
    counts = matrix( rep( NA , N*J ) , ncol = J )
  for (j in 1:J) {
    counts[,j]=rpois(n=N, lambda=mean[,j])
  }
  results = list( counts = counts , alpha = alpha , mean = mean )
  return( results )
}



###### libraries ####
library( INLA )
library( MCMCglmm )
```

```
###### This section gives R code used for
###### simulation and estimation of LCMs
###### Input for simulations ########
###### level.sigma
rho = -0.95
sigma11 = 0.4
sigma22 = 0.4
sigma12 = rho*sqrt(sigma11*sigma22)
level.sigma.true = matrix(c(sigma11,sigma12,sigma12,sigma22),ncol=2)

###### log.lambda

mean = c( 10 , 10 )
log.mean = log( mean )

####################################
###### Simulation of counts
####################################

### Sample size n

out = c()
out.inla=c()
out.inla[k] = c()

fit.mcmc = c()
fit.mcmc[k] = c()
```

```
sig11 = c()

sig22 = c()

sig11[k] = c()

sig22[k] = c()


rho.inla = c()

rho.inla[k] = c()


M1 = c()

M2 = c()

M1[k] = c()

M2[k] = c()


mcmc.M1 = c()

mcmc.M2 = c()

mcmc.M1[k] = c()

mcmc.M2[k] = c()


rho.mcmc = c()

rho.mcmc[k] = c()


mcmc.sig11 = c()

mcmc.sig22 = c()

mcmc.sig11[k] = c()

mcmc.sig22[k] = c()


num= rep( 5000, 500 )

num.sim=length(num)
```

```
### Loop for n


for ( k in 1:num.sim ) {
  N = num[k]
  #set.seed(2468)
  sim.data=sim.LCM( N = N ,
                    log.lambda=log.mean,
                    level.sigma=level.sigma.true)
  dim(sim.data$counts)
  ######### INLA set-up ############


  y = c(sim.data$counts[,1],sim.data$counts[,2])
  N = dim(sim.data$counts)[1]
  N.all = 2*N
  index = 1:N.all
  n = N.all/2


  int1 = c( rep( 1 , n ),rep( 0 , n ))
  int2 = c( rep( 0 , n ),rep( 1 , n ))


  formula = y ~ int1 + int2 +
  f( index , model = "iid2d", n = N.all ) - 1
  result = inla( formula , family = "poisson",
                data = data.frame( y , index , int1 , int2 ),
                control.compute = list( dic = TRUE, config = TRUE))


  #summary(result)
```

```
# Results from INLA


t1 = result$summary.hyperpar[,c(1,3,5)]

t1[1,] = 1/t1[1,]

t1[2,] = 1/t1[2,]

t2 = exp(result$summary.fixed[,c(1,3,5)])

hyper = rbind( t1 , t2 )

M.inla = exp(result$summary.fixed[1,1])

M.inla.upp = exp(result$summary.fixed[1,5])

M.inla.low = exp(result$summary.fixed[1,3])

rho.inla[k] = result$summary.hyperpar[3,1]

sig11[k] = 1/result$summary.hyperpar[1,1]

sig22[k] = 1/result$summary.hyperpar[2,1]

sig12 = rho.inla*sqrt(sig11*sig22)

M1[k] = exp( result$summary.fixed[[1]][1] )

M2[k] = exp( result$summary.fixed[[1]][2] )


##### MCMC ##########

y1=sim.data$counts[, 1 ]

y2=sim.data$counts[, 2 ]

dat1=data.frame( y1 = y1 , y2 = y2 )


 # Start the clock

ptm <- proc.time(  )

#set.seed( 2468 )

fit1<- MCMCglmm(cbind(y1, y2) ~ trait-1,

                rcov = ~us(trait):units,

                data = dat1, family = c( "poisson" , "poisson" ),
```

```
                    nitt = 5001,burnin=1,thin=1,
                    verbose = FALSE)

  # Stop the clock
  runtime = proc.time() - ptm

   ##### Results from MCMC ########
  fit.mcmc[[k]]=fit1

  # Extracts fixed effects for y1 = M1
  mcmc.M1[k] = exp(summary(fit1)[[5]][1])
  mcmc.M2[k] = exp(summary(fit1)[[5]][2])

  # Extracts random effects  sigma11
  mcmc.sig11[k] = summary(fit1)[[8]][1]
  mcmc.sig22[k] = summary(fit1)[[8]][4]
  rho.mcmc[k] = summary(fit1)[[8]][3]/
  (sqrt(summary(fit1)[[8]][1]*summary(fit1)[[8]][4]))

  print(paste("Iteration number:",k))
  print(paste("Time INLA:",result$cpu.used[4]))
  print(paste("Time MCMC:",runtime[3]))
}

# INLA Averages

mean.sigma11.inla = sum( sig11 )/500
mean.sigma22.inla = sum( sig22 )/500
```

```
mean.rho.inla = sum( rho.inla )/500
mean.M1.inla = sum( M1 )/500
mean.M2.inla = sum( M2 )/500

mean.sigma11.mcmc = sum( mcmc.sig11 )/500
mean.sigma22.mcmc = sum( mcmc.sig22 )/500
mean.rho.mcmc = sum( rho.mcmc )/500
mean.M1.mcmc = sum( mcmc.M1 )/500
mean.M2.mcmc = sum( mcmc.M2 )/500

#MCMC Averages

library(Metrics) # for mse() function

inla.mse.sig11 = mse( sigma11 , sig11 )
inla.mse.sig22 = mse( sigma22 , sig22 )
inla.mse.rho   = mse( rho , rho.inla )
inla.mse.M1    = mse( mean[1] , M1 )
inla.mse.M2    = mse( mean[2] , M2 )

mcmc.mse.sig11 = mse( sigma11, mcmc.sig11 )
mcmc.mse.sig22 = mse( sigma22, mcmc.sig22 )
mcmc.mse.rho = mse( rho, rho.mcmc )
mcmc.mse.M1 = mse( mean[1], mcmc.M1 )
mcmc.mse.M2 = mse( mean[2], mcmc.M2 )
```

Appendix B

Simulation R-code for chapter 5

```
###### libraries ####
library(MASS)
###### function to simulate bivariate counts (with Poisson marginals) #####
sim.biv.counts=function(N=N,level.sigma,log.lambda,level.b){

  log.lambda=as.matrix(log.lambda)
  level.sigma=as.matrix(level.sigma)
  level.b=as.matrix(level.b)

  J=dim(log.lambda)[1]
  log.mean=matrix(rep(log.lambda,N),ncol=J)

  # simulate level alpha

  level.mu=rep(0,J)
```

```
  alpha=mvrnorm(n=N, mu=level.mu, Sigma=level.sigma)

  b=mvrnorm(n=N,mu=level.mu,Sigma = level.b)

  # simulate counts
  new.mean = log.mean + alpha + b
  mean=exp(new.mean)
  counts=matrix(rep(NA,N*J),ncol=J)

  for (j in 1:J) {
    counts[,j]=rpois(n=N, lambda=mean[,j])
  }
  results=list(counts=counts,mean=mean,
               level.sigma=level.sigma, level.b = level.b)
  return(results)
}

###### libraries ####
library(INLA)
library(MCMCglmm)
###### This section gives R code used for
###### simulation and estimation of LCMs
###### in Section 5.3
###### for model in (5.3.4)-(5.3.5).
###### Results are given in
###### Tables 5.3.2, 5.3.3, and 5.3.4
###### Input for simulations ########
###### level.sigma
```

```
rho=0.90
sigma11=0.4
sigma22=0.4
sigma12=rho*sqrt(sigma11*sigma22)
level.sigma.true=matrix(c(sigma11,sigma12,sigma12,sigma22),ncol=2)
b.11 = 0.01
b.22 = 0.01
b.12 = b.21 = 0
level.b.true=matrix(c(b.11,b.12,b.21,b.22),ncol=2)

###### log.lambda

mean=c(5,5)
log.mean=log(mean)

####################################
###### Prior distrbutions
####################################

prec.prior.inla = list( prec = list( param = c( 1, 0.0005 ) ) )
prec.prior.bugs = list( prec = list( param = c( 0.001, 0.001 ) ) )
prec.prior.fong = list( prec = list( param = c( 0.5, 0.0162 ) ) )



####################################
###### Simulation of counts
####################################
```

```
### Sample size n
num = c( 50, 100, 500, 1000, 5000 )



out.inla=c()


#out.inla1=c()
#out.bugs1=c()
out.fong1=c()
num.sim=length(num)
#reps.sim=length(reps)
### Loop for n



for (k in 1:num.sim) {
  N=num[k]
  set.seed( 2468 )
  sim.data=sim.biv.counts(N=N,
                  level.sigma = level.sigma.true,
                  log.lambda = log.mean,
                  level.b = level.b.true)
  ######### set-up #############


  y=c(sim.data$counts[,1],sim.data$counts[,2])


  N=dim(sim.data$counts)[1]
  N.all=2*N
  index_jit = 1:N.all
```

```
#seas <- rep(c(1:4), each = (N)/2)
#re.index = rep(1:2, each = N)
#n.locations = rep(1:5, each = (2*N)/5)
m.locations.1 = c(sample( 12, N.all/2, replace = TRUE ),
                  rep(NA,N.all/2))
m.locations.2 = c(rep(NA,N.all/2), sample( 12, N.all/2,
                  replace = TRUE ))
n=N.all/2

int1=c(rep(1,n),rep(NA,n))
int2=c(rep(NA,n),rep(1,n))
# b.1=c(rep(1:(n/50),50),rep(NA,n))
# b.2=c(rep(NA,n),rep(1:(n/50),50))

#-----------------------------------------

formula1= y ~ - 1 +
  int1 +
  int2 +
  f(index_jit, model = "iid2d", n = N.all ) +
  f(m.locations.1, model="iid", hyper = prec.prior.inla ) +
  f(m.locations.2, model="iid", hyper = prec.prior.inla )

result1= inla(formula1,family="poisson",
              data =data.frame(y,index_jit,int1,int2,
                               m.locations.1,m.locations.2),
              control.compute=list(dic=TRUE))
```

```
# Set-up output -------------------------------------

  # 1. Results from inla prior

t1.inla=result1$summary.hyperpar[,c(1,5,3)]
t1.inla[1,]=1/t1.inla[1,]
t1.inla[2,]=1/t1.inla[2,]
t1.inla[4,]=1/t1.inla[4,]
t1.inla[5,]=1/t1.inla[5,]
t2.inla=exp(result1$summary.fixed[,c(1,3,5)])
hyper.fong=rbind(t2.inla,t1.inla)

  # 2. Output for all n and fixed m = 10:

#out1=data.frame(N,hyper.inla)
#out.inla=rbind(out.inla,out1)

  # 3. Output for fixed n = 5000 and all m: INLA prior

#out2=data.frame(N,hyper.inla)
#out.inla1=rbind(out.inla1,out2)

  # 4. Output for fixed n = 5000 and all m: BUGS prior
# out3=data.frame(N,hyper.bugs)
# out.bugs1=rbind(out.bugs1,out3)
```

```
    # 4. Output for fixed n = 5000 and all m: BUGS prior
  out4=data.frame(N,hyper.fong)
  out.fong1=rbind(out.fong1,out4)


}

# Set up posterior output table

t1_tab = c(mean[1],mean[2],sigma11,sigma22,rho,b.11,b.22)

true_value = rep( t1_tab, 5 )

out.inla1 = data.frame( out.inla, out.inla1, out.bugs1,
                        out.fong1, true_value )

round( out.fong1, 3 )

names( full_table ) = c("n", "Posterior Mean", "Lower CI","Upper CI" )

full_table = data.frame(round(cbind(out.bugs1[,c(1,2,4,3)],
                                    out.fong1[,c(1,2,4,3)]), 3),true_value)


# Produce table in as Latex code

xtable::xtable( full_table , digits = 3 )
```

```
hyper.inla

hyper.bugs

hyper.fong



summary(result1)



poisson.results = c(result1$dic$dic,result2$dic$dic,
                    result3$dic$dic)
nbinomila.results = c(result4$dic$dic,result5$dic$dic,
                    result6$dic$dic)
```

Appendix C

Journal articles

Multilevel dynamic multivariate modelling of tick life stages count data

Thabo Lephoto[1*], Henry Mwambi[1], Oliver Bodhlyera[1] and Holly Gaff[2]

[1]School of Mathematics Statistics and Computer Science, University of KwaZulu-Natal, KwaZulu-Natal Province, Private Bag X01, 3201, South Africa

[2]Department of Biological Sciences, Old Dominion University, Norfolk, United States

Abstract

The modelling described in this study enables adequate incorporation of the association in the response over time, and the association between the components of the response vector. We used a multilevel dynamic model to account for the association among tick life-stages in the response vector. The model for multivariate time series of counts account for over-dispersion and relaxes the assumption of a positive association between the components by allowing for the negative association. The flexibility of the model allows for a combination of different marginal count data distributions and builds a dynamic model for the vector time series. We model tick life stage

time series of counts as a function of habitat-type-specific and time-dependent covariates. The integrated nested Laplace approximation (INLA) approach is employed for fast-approximation. Simulation results have proven accurate parameter estimations of posterior parameters. Multilevel application to data have also successfully fit the data well and provided estimates. Results showed positive associations between the larvae and nymphs, larvae and adults, and nymphs and adult tick stages, with strong correlations between larva and nymph, followed by nymph and adult, and then larva and adult stages. The model estimated similar non-linear months effects across the three habitat types, while constant habitat type specific trends for larvae were estimated in edges, grass and woods. However, a declining nymph and adult trends were estimated in the edges and woods throughout the study period.

Key phrases: Dynamic modeling, Lone Star tick, INLA, Life stage trends.

Introduction

An increase in tick abundance and tick-borne pathogens constitute a growing threat to public health (Laaksonen et al., 2017; Paull et al., 2022; Wimms et al., 2023). A study by Nadolny et al. (2014) recorded more than 66 000 questing tick populations in the southeastern Virginia, comprising seven species from a variety of habitats, with *Amblyomma americanum* (Lone Star tick) constituting over 95% of the ticks collected. Our research is focused on modelling *A. americanum* tick life-stage abundance collected over time from different habitat types. The objective of this study is to develop a multilevel model for tick count data collected over time for tick life stages, namely: larvae, nymph and adult stages.

Models in literature for tick modelling involve GLMs, maximum entropy (MaxEnt), classification and regression tree (CART), species distribution modelling (SDM), ecological niche

factor analysis (ENFA), and Bayesian hierarchical models among many others. Generalized linear models have been employed to investigate tick aggregation at different spatial ranges, provide robust estimates of tick densities between landscapes, and to determine and disaggregate the drivers of tick density and presence probabilities in an area (De Clercq et al., 2015; Kaizer et al., 2015; Estrada-Peña et al., 2016; Requena-García et al., 2017; Sagurova et al., 2019; Fan et al., 2023). Maximum entropy methods have been used to explore the limits of potential distribution by extrapolating the environmental requirements of ticks. These methods were also utilised to analyse the possible spatial range of tick species, and to explore how climate change can shape their distribution (de Oliveira et al., 2017; Raghavan et al., 2019; Ma et al., 2021; Lippi et al., 2021; Rochlin et al., 2023). Studies using SDM techniques lead to precise boundaries of the range of tick presence based on computational map modelling and demonstrate the way in which local populations of ticks differ in abundance towards the boundaries of the range (Qviller et al., 2016; Estrada-Peña et al., 2016; Lado et al., 2020; Kopsco et al., 2022; Rochlin et al., 2022, 2023). The CART technique has been used to review data on tick distribution and prevalence of tick-borne disease (TBD) for a national TBD management approach using the ecological and epidemiological information of ticks and the related diseases they transmit (Lynen et al., 2008; De Clercq et al., 2015; Kjær et al., 2019). The ENFA technique has been used to measure the extent to which the preferences of a given species deviate from average conditions and the extent to which the species is selective over the range of environmental conditions available in a country (Estrada-Pena and Venzal, 2007; Lynen et al., 2008). Bayesian hierarchical models have been used in modelling prevalence of tick-borne pathogens and spatio-temporal distributions (Wimberly et al., 2008; Simpson et al., 2019; Lephoto et al., 2021; Clark and Wells, 2023).

Bayesian methods used in modelling tick abundances or related tick-borne diseases have em-

ployed the Markov chain Monte Carlo (MCMC) sampling technique (Zipkin, 2012; Liu et al., 2017; Neupane et al., 2021; Clark and Wells, 2023; Wang et al., 2023). The integrated nested Laplace approximation (INLA) technique can be used to conduct Bayesian inference as an alternative method to MCMC. The INLA technique approximates the posterior distribution based on the Laplace formulation, then it draws from the approximated posterior distribution to estimate the desired quantities and turns out to be more computationally efficient when compared to the MCMC approach (Arab, 2015; Ferkingstad et al., 2017; Khana et al., 2018; Van Niekerk et al., 2023).

We extend the application of a multilevel multivariate model proposed by Serhiyenko et al. (2018), also called the level correlated model (LCM), which they used to study temporal patterns in prescription drug counts for competing drugs in therapeutics. The model allows the flexibility of modelling multivariate hierarchical data at different cluster levels and employs the INLA algorithm for fast and accurate approximate Bayesian posterior distributions. The model is appealing in our case because it helps relax the assumption of positive associations between the components of the response vector (which is an assumption made when dealing with hierarchical life stages of the life cycle of a biological population) to include negative associations through the variance-covariance matrix, where components of the response vector are modeled jointly using univariate distributions and the components are correlated at the location level. This study focuses on models for multivariate time series of tick life stages count data collected at different habitat types to investigate covariate effects and determine dependence between the life stages. Univariate regression models cannot account for the dependence between tick life stages (Aitchison and Ho, 1989; Chib and Winkelmann, 2001; Ma et al., 2008). Since the dependence between tick life-stages may be due to biological processes, where eggs hatch to emerging larvae, larvae to nymphs and nymphs to adults, it

is clear that numbers in a later stage will depend on previous stages in a life cycle, and correlations can occur due to factors such as habitat type, sample sites and other factors which may simultaneously affect life stages. Multilevel models are used instead of the univariate regression models to account for the dependence between tick life stages. To our knowledge, the models have never been applied to ecological data, particularly tick life-stage count data. Our aim is to understand the causes and consequences of variation in the abundance as a function of environmental and temporal covariates.

Section 1 gives a detailed description of the structure of the multilevel model and its framework in modelling multivariate time series data. A brief description of model diagnostics is given in Section 2. Section 3 details the multilevel model simulation process and results. Section 4 describes the application of the model, where the first subsection gives a description of the tick life-stage data, then a detailed description of multilevel Poisson and zero-inflated Poisson model fitting, model application to data, model diagnostics, and lastly presents and interprets model results. The conclusion is presented in the last section.

1 Model framework

Let $\boldsymbol{Y}_{it} = (Y_{1,it}, \ldots, Y_{J,it})'$ be a $J-$dimensional vector of count responses observed over T regularly spaced time points and n locations, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The assumption of the Poisson or zero inflated Poisson (ZIP) marginal distributions for the multilevel time series model corresponds to

$$Y_{j,it}|\delta_j, \lambda_{j,it} \sim \begin{cases} \text{Poi}(\lambda_{j,it}); \text{ when } \delta_j = 1; \text{ or} \\ \text{ZIP}(\lambda_{j,it}, \delta_j); \text{ otherwise,} \end{cases} \tag{1}$$

and to include the effect for covariates we use the log-link regression model given by

$$\log(\lambda_{j,it}) = \beta_{j,i0} + \gamma_{j,t} + \boldsymbol{z}'_{j,it}\boldsymbol{\beta}_j + \xi_{j,it}, \tag{2}$$

where $\lambda_{j,it}$ and δ_j are the mean count and dispersion parameters for the jth component, respectively. $\beta_{j,i0}$ is a location-specific intercept for the jth component of the response vector, $\boldsymbol{z}'_{j,it}$ denotes a p_j-dimensional vector of predictors, $\boldsymbol{\beta}_j$ is a p_j-dimensional vector of coefficients corresponding to the predictors, and $\xi_{j,it}$ is the random effects term explaining the dependence between the components of the response vector through the variance covariance matrix $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3}$$

where $\mathbf{0}$ is a $J-$dimensional zero vector, $\boldsymbol{\Sigma}$ is a variance-covariance matrix with σ_{rs} elements, such that $1 \leq r \leq s \leq J$, and under Bayesian estimation will be part of hyperparameters. The dependence between components of the response vector can be postulated either though $\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ or $\boldsymbol{\xi}_{it} \ \sim \ \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. Postulating correlations through the matrix $\boldsymbol{\Sigma}$ means that the correlations between the J tick-life cycle developmental stages (which are the same as the response vector components) are location independent whereas $\boldsymbol{\Sigma}_i$ is a location-specific variance covariance matrix. Consider $1 \leq r \leq s \leq J$, and suppose $r \neq s$, then the off-diagonal elements of $\boldsymbol{\Sigma}$ are given as follows:

$$\sigma_{rs} = \rho_{rs}\sqrt{\sigma_{rr}\sigma_{ss}},$$

where σ_{rr} and σ_{ss} are the marginal variances and ρ_{rs} is the correlation coefficient in the matrix. To introduce location specificity we need to introduce an index i in the equation above. The parameter, $\gamma_{j,t}$, is the jth component specific dynamic time effect modeled as a random walk such that:

$$\gamma_{j,t} = \gamma_{j,(t-1)} + \omega_{j,t}, \tag{4}$$

where $\omega_{j,t} \sim \mathrm{N}(0, 1/W_j)$ denotes the error term for component j of the response vector assumed to follow the normal distribution. W is the precision parameter based on the inverse-gamma distribution, and the LogGamma prior is specified for $\log(W_j) \sim \mathrm{LogGamma}(a, b)$ in the application, where $a = 1$ and $b = 0.00005$. The time effect in Equation 4 can be modified such that:

$$\gamma_t = \gamma_{(t-1)} + \omega_t, \tag{5}$$

which means that the model in Equation 2 is a model where the time effect is not stage (or response component) specific, and $\omega_t \sim \mathrm{N}(0, 1/W)$. This implies a model assuming shared time effects across all locations and all components of the response vector. The second case is

$$\gamma_{it} = \gamma_{i(t-1)} + \omega_{it}, \tag{6}$$

and this implies that Equation 2 is a model where the time effect is location specific and life-cycle stage (or response component) independent, and $\omega_{it} \sim \mathrm{N}(0, 1/W_i)$ (an assumption where the errors share the same distribution over all components of the response vector at a given location i). Thus, the model in Equation 2 modifies into a model assuming shared

time effects across all components of the response vector. Lastly, Equation 4 can be modified to

$$\gamma_{j,it} = \gamma_{j,i(t-1)} + \omega_{j,it}, \tag{7}$$

which is a model where the time effect is assumed to have separate time evolution on each of the components of the response vector and each location, where $\omega_{j,it} \sim \mathrm{N}(0, 1/W_i)$ or $\omega_{j,it} \sim \mathrm{N}(0, 1/W_j)$ (an assumption that state errors share the same distribution for all components of the response vector or that state errors follow response component specific normal distributions). Thus, the model in Equation 2 modifies into a model assuming different time effects across all locations and all components of the response vector. The dispersion parameter $\delta_j = 1$ in Equation 1 implies the assumption of the Poisson distribution for the jth component, and $\delta_j \neq 1$ implies the ZIP distribution for the jth component. The variance-covariance matrix, $\boldsymbol{\Sigma}$, plays a crucial role in understanding the dependence between the components of the vector of counts where each component of the response vector follows a univariate distribution. Consequently, the model in Equations 1 to 3 accounts for overdispersion through the use of the ZIP model and in this case the diagonal terms of the matrix $\boldsymbol{\Sigma}$; for $r = s$ in $\boldsymbol{\Sigma}_{1 \leq r \leq s \leq J}$, $\sigma_{jj} > 0$, so that $\mathrm{Var}[Y_{j,it}] > \mathrm{E}[Y_{j,it}]$. Note that

$$\mathrm{E}[Y_{j,it}] = \exp(z'_{j,it}\beta_j \exp(\sigma_{jj}/2) = m_{j,it},$$

$$\mathrm{Var}[Y_{j,it}] = m_{j,it} + m^2_{j,it}(\exp(\sigma_{jj}) - 1),$$

$$\text{Cov}[Y_{r,it}, Y_{s,it}] = m_{r,it} m_{s,it} (\exp(\sigma_{rs}) - 1).$$

Moreover, the model can allow for the positive or negative dependence between the components of the response vector (Chib and Winkelmann, 2001). This follows from the sign of the covariance in Equation 3 which depends directly on the value of σ_{rs}, and a negative value of σ_{rs} yields a negative value of the association, while a positive value results in a positive association between the components Y_{ri} and Y_{si}.

2 Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \bar{D}(\theta) + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter et al. (2002) state that p_D values less than zero indicate substantial conflict between the prior distribution and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of deviance suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter et al. (2002).

3 Simulation study

Suppose $\boldsymbol{Y}_i = (Y_{1i}, Y_{2i})$ represent a bivariate ($J = 2$) count data vector that is observed at $i = 1, \ldots, n$ locations when there is no over-dispersion in the data. The data is simulated

according to the simplified model described in Equations 8 and 9, and does not assume any temporal dependence. That is, the marginal Poisson multilevel model is given by:

$$Y_{ji}|\lambda_{ji} \sim \text{Poi}(\lambda_{ji}), \tag{8}$$

$$\log(\lambda_{ji}) = \beta_{j,0} + \xi_{ji}, \tag{9}$$

where $\beta_{j,0}$ and $\xi_{j,i}$ are the intercept and random effects parameters for the jth component, respectively. The bivariate multilevel model data simulation comprises $\boldsymbol{\beta}_0 = (\beta_{1,0}, \beta_{2,0})'$ and $\boldsymbol{\xi}_i \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with σ_{11}, σ_{22} and ρ_{12}, the elements of $\boldsymbol{\Sigma}$. We assume that the true parameter values were $\sigma_{11} = \sigma_{22} = 0.4$, $\rho_{12} = -0.95$, and $\beta_{1,0} = \beta_{2,0} = 10$ and run the simulation for $n = 100$, $n = 500$, $n = 5\ 000$ and $n = 10\ 000$ using the R-INLA package in R software proposed by Rue et al. (2009) and fitted the four different bivariate multilevel model. The bivariate multilevel model is formulated through joint modelling of the four models below.

Model 1:

$$Y_{1i}|\lambda_{1i} \sim \text{Poi}(\lambda_{1i}); Y_{2i}|\lambda_{2i} \sim \text{Poi}(\lambda_{2i}). \tag{10}$$

Model 2:

$$Y_{1i}|\lambda_{1i}, \delta_1 \sim \text{ZIP}(\lambda_{1i}, \delta_1); Y_{2i}|\lambda_{2i}, \delta_2 \sim \text{ZIP}(\lambda_{2i}, \delta_2). \tag{11}$$

Model 3:

$$Y_{1i}|\lambda_{1i}, \delta_1 \sim \text{ZIP}(\lambda_{1i}, \delta_1); Y_{2i}|\lambda_{2i} \sim \text{Poi}(\lambda_{2i}). \tag{12}$$

Model 4:

$$Y_{1i}|\lambda_{1i} \sim \text{Poi}(\lambda_{1i}); Y_{2i}|\lambda_{2i}, \delta_2 \sim \text{ZIP}(\lambda_{2i}, \delta_2). \tag{13}$$

λ_{ji} is modeled as a function of ξ_{ji} and $\beta_{j,0}$ as in Equation 9. We assume a Normal prior for $\beta_{j,0}$ and the matrix $\boldsymbol{\Sigma}$ has a Wishart prior with $r = 2 \times J + 1$ degrees of freedom and identity matrix as a prior precision matrix. We also use default hyperparameter specifications in the `inla()` function from the R-INLA package. In particular we assume $\beta_{j,0} \sim \text{N}(0, 10^3)$ and the precision matrix $\boldsymbol{\Sigma}^{-1} \sim W(r, \boldsymbol{\Sigma}^{-1})$. For computational time comparison, a fully Bayesian inference model using Markov Chain Monte Carlo (MCMC) was fitted using the `MCMCglmm()`function in the MCMCglmm R package (Hadfield, 2010). Here, we assume the default Normal prior for $\beta_{j,0}$, with a zero mean vector and a diagonal variance matrix with large variances $\text{N}(0, 1 \times 10^{10})$ (Hadfield, 2010). We assumed a default inverse Wishart prior for each component in the covariance matrix of the random effects.

Estimated parameters, CPU (Central Processing Unit) times between INLA and fully Bayesian inference using MCMC, and estimated averages of posterior means are given in the subsections that follow.

3.1 Estimation using INLA

The CPU time comparisons between a fully Bayesian inference using MCMC and an approximate Bayesian inference using INLA are presented in Table 1. The MCMC sampling was run with 105 000 iterations, where 5 000 iterations were used for burn-in, and 100 iterations were for thinning. As a result, only 1 000 posterior samples were retained from which posterior summaries were calculated. It can be seen from Table 1 that the computational times quickly increase with the increase in sample size for the MCMC runs, and that INLA runs

faster than the MCMC, as expected in all cases. The INLA technique is approximately 20 times faster for Poisson distributions, 6 times faster for ZIP, and between 10 and 17 times faster for Poisson and ZIP mixed distributions of counts when n is large ($\geq$ 5 000) compared to the MCMC, and its posterior estimates are closer to those obtained using the MCMC (Table 2). Note that the speed under M2 - M4 is approximately double the speed under M1. This indicates that the distributions involving the inclusion of ZIP models are expected to run twice as slow as models including Poisson distributions.

Table 1: CPU time of model estimation using INLA and MCMC (seconds)

n	INLA				MCMC
	(M1)	(M2)	(M3)	(M4)	
100	1	2	2	2	11
500	2	4	3	3	48
1 000	5	10	7	7	95
5 000	21	80	27	49	467

Table 2: Estimated parameters in simulated models.

n	Model parameter	M1			M2			M3			M4			True value
		Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI		
			2.5%	97.5%		2.5%	97.5%		2.5%	97.5%		2.5%	97.5%	
100	σ_{11}	0.48	0.34	0.69	0.47	0.34	0.69	0.48	0.34	0.70	0.47	0.34	0.69	0.40
	σ_{22}	0.43	0.31	0.62	0.43	0.31	0.62	0.43	0.31	0.62	0.43	0.31	0.62	0.40
	ρ_{12}	-0.86	-0.92	-0.77	-0.86	-0.92	-0.77	-0.86	-0.92	-0.77	-0.86	-0.92	-0.78	-0.95
	$\beta_{1,0}$	9.93	8.48	11.56	9.98	8.53	11.61	9.97	8.51	11.61	9.93	8.49	11.57	10.00
	$\beta_{2,0}$	9.18	7.89	10.63	9.18	7.88	10.63	9.18	7.89	10.64	9.18	7.88	10.63	10.00
500	σ_{11}	0.39	0.34	0.46	0.38	0.34	0.43	0.40	0.35	0.47	0.40	0.35	0.47	0.40
	σ_{22}	0.42	0.36	0.50	0.42	0.38	0.48	0.43	0.38	0.51	0.43	0.37	0.50	0.40
	ρ_{12}	-0.88	-0.92	-0.84	-0.88	-0.90	-0.85	-0.89	-0.92	-0.84	-0.89	-0.92	-0.86	-0.95
	$\beta_{1,0}$	9.72	9.12	10.35	9.73	9.13	10.36	9.72	9.12	10.35	9.72	9.12	10.35	10.00
	$\beta_{2,0}$	9.75	9.13	10.40	9.74	9.12	10.39	9.76	9.14	10.41	9.75	9.12	10.40	10.00
5 000	σ_{11}	0.42	0.37	0.47	0.40	0.37	0.43	0.41	0.37	0.46	0.42	0.39	0.47	0.40
	σ_{22}	0.42	0.38	0.48	0.42	0.38	0.45	0.43	0.38	0.48	0.42	0.38	0.47	0.40
	ρ_{12}	-0.90	-0.93	-0.87	-0.91	-0.93	-0.90	-0.90	-0.93	-0.87	-0.90	-0.93	-0.88	-0.95
	$\beta_{1,0}$	9.70	9.26	10.15	9.71	9.28	10.16	9.72	9.29	10.17	9.70	9.26	10.15	10.00
	$\beta_{2,0}$	10.09	9.63	10.56	10.11	9.66	10.57	10.08	9.63	10.55	10.11	9.66	10.58	10.00
10 000	σ_{11}	0.40	0.38	0.42	0.41	0.39	0.42	0.40	0.38	0.42	0.40	0.38	0.42	0.40
	σ_{22}	0.40	0.38	0.42	0.40	0.38	0.42	0.40	0.38	0.43	0.40	0.38	0.42	0.40
	ρ_{12}	-0.94	-0.95	-0.93	-0.95	-0.95	-0.94	-0.95	-0.96	-0.93	-0.94	-0.95	-0.93	-0.95
	$\beta_{1,0}$	9.86	9.67	10.06	9.87	9.67	10.07	9.87	9.67	10.06	9.85	9.65	10.05	10.00
	$\beta_{2,0}$	10.05	9.85	10.25	10.05	9.85	10.25	10.05	9.85	10.26	10.06	9.86	10.26	10.00

For brevity, 500 replications were run for $n = 5\ 000$, and M1 model parameters were estimated using MCMC and INLA. The model averages of posterior means and MSEs (mean square errors) of 500 replicates are provided in Table 3. It is clear from the table that INLA average posterior estimates are reasonably close to MCMC average posterior estimates in all cases. As a result, this study will use the INLA method because of its fast computation time when compared to the MCMC as well as the accuracy of its estimates.

Table 3: Averages of posterior means and MSEs of 500 replicates for $n = 5\ 000$

Parameters	MCMC		INLA		True value
	Posterior mean	MSE	Posterior mean	MSE	
σ_{11}	0.401	0.0001037147	0.401	0.0001042946	0.40
σ_{22}	0.401	0.0001043847	0.401	0.0001062165	0.40
ρ_{12}	-0.947	0.0000516038	-0.941	0.0001146017	-0.95
$\beta_{1,0}$	10.002	0.0094999830	9.997	0.0094975680	10.0
$\beta_{2,0}$	9.992	0.0092616110	9.987	0.0093305050	10.0

4 Tick life-stage modelling using multilevel model

In order to understand the dynamics and habitat type variation in the abundance of tick life-stages, we describe the statistical analysis pertaining to tick life-stage count data using multilevel dynamic models. In addition, the study focuses on determining factors contributing to tick life-stage abundances across different habitat types. Most of the existing research focuses on location level counts for the different life stages of the tick cycle by investigating the tick aggregation at different spatial ranges, determining and disaggregating the drivers of tick density and probability of presence, and providing robust estimates of tick densities (Zannou et al., 2021). The association between tick life-stages and their locations is of

interest. Furthermore, there is interest in knowing temporal effects on the distribution of counts.

4.1 Data description

We used data collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, Virginia, United States. Ticks were collected using standard flagging techniques (CDC, 2020) along established transects and identified species and life-stage according to (Sonenshine, 1979). Larvae-, nymph- and adult-stage count data were recorded for each location, habitat and habitat type through the four seasons of the year. Eight counties and independent cities were sampled in this state, namely; City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County. For brevity, all will be referred to as counties. Data was collected at least once a month on varying days of the week and at 12 different sites in southeast Virginia from May 2009 through December 2018. The data were collected from multiple areas referred to as habitats, and each area was designated by a unique number ranging from 1 to 5 for different habitats. The habitat type was used to designate the kind of area (woods, edges or grass) where the data were collected. The number of the week (from 1 to 53) was also recorded during data collection. Week 1 is the first week of the year, and while fewer ticks were observed in winter, the adult stage of *I. scapularis* is active in winter. To help align the information from year to year, data collected during the last week of December were recorded as week 53, which was also the first week of January of the following year. Track records for the month and year were also kept. This study makes use of habitat types (edges, woods and grass), time in months, and three tick life stage counts for modelling. After removing unused variables from the raw dataset, we

were left with 103 months out of the 116 total monthly data points. Note that the removal of certain data points was due to life stages being collected at different months during the first 13 months. The retained dataset consisted of three habitat types, within which there were three tick life stages each of length 103 (months).

4.2 Exploratory data anlysis

Figure 1 depicts the monthly tick life stage counts by habitat type from June 2010 through December 2018. The data shows strong seasonality throughout the years, indicated by peaks during warmer seasons. The plots show that larvae counts were higher in all habitat types relative to nymph and adult counts. The woods habitat type shows that higher tick counts were recorded in 2012, with the larvae counts reaching approximately 30 000. However, we observe a decline in the abundances for all life stages over the study period. The grass habitat type had lower nymphal tick numbers from 2010 through 2014, after which there were considerable peaks during summer seasons. We also observe a steady increase over the study period in all stages. In the edges habitat type we observe a decline over time for nymph and adult abundances, while larvae abundances were constant throughout the study period.

4.3 Model framework for tick life stage data

Let $\boldsymbol{Y}_{it} = (Y_{L,it}, Y_{N,it}, Y_{A,it})'$ be a vector of count responses for tick life stage $j = L, N, A$ in habitat type $i = E, G, W$ (E = Edges, G = Grass, W = Woods) at equally spaced monthly times $t = 1, \ldots, 103$. The multivariate vector of tick life-stage responses $\boldsymbol{Y}_{it}$ is modelled

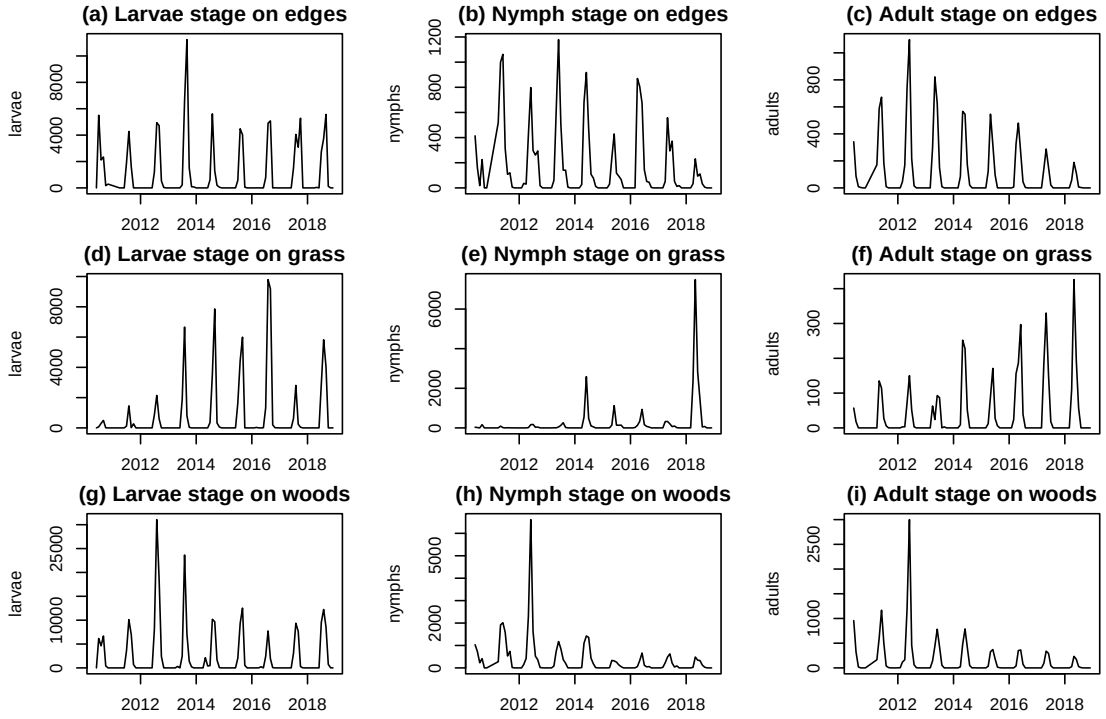


Figure 1: Plots showing time series of tick life stage counts in each habitat type, 2009 through 2018.

using marginal Poisson and/or zero-inflated Poisson distributions as follows:

$$Y_{j,it}|\delta_j, \lambda_{j,it} \sim \begin{cases} \text{Poi}(\lambda_{j,it}); \text{ when } \delta_j = 1; \text{ or} \\ \text{ZIP}(\lambda_{j,it}, \delta_j); \text{ otherwise,} \end{cases} \tag{14}$$

where $\lambda_{j,it}$ is the jth tick life-stage mean count and δ_j is a dispersion parameter at stage j. The log-link is given by

$$\log(\lambda_{j,it}) = \beta_{j,it,0} + \gamma_t + \alpha_i + z'_{j,it}\beta_j + \xi_{j,it}, \tag{15}$$

where $\beta_{j,it,0}$ is the dynamic intercept, γ_{it} is the month time effect in the ith habitat type, $z'_{j,it}$ denotes a p_j-dimensional vector corresponding to the fixed effects of seasonal changes and sinusoidal predictors $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$ included to handle the seasonality with period 12 (in months), $\xi_{j,it}$ is the random effects term explaining the dependence between tick stages. The random vector parameter, $\boldsymbol{\xi}_{it} = (\xi_{L,it}, \xi_{N,it}, \xi_{A,it})'$, is the vector of the unobserved random variables introduced to capture the tick-stage dependence through the multivariate normal distribution as follows:

$$\boldsymbol{\xi}_{it} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{16}$$

where $\mathbf{0}$ is a $J-$dimensional zero vector, $\boldsymbol{\Sigma}$ is the variance-covariance matrix for the random effect terms such that for $r \neq s$ where $\{r \text{ and } s \in j = 1, ..., J\}$, $\boldsymbol{\rho}_{rs} = (\rho_{LN}, \rho_{LA}, \rho_{NA})'$ denotes the tick life stage correlation vector. The month time effects vector $\boldsymbol{\gamma}_t = (\gamma_{L,t}, \gamma_{N,t}, \gamma_{A,t})$ and the dynamic intercept $\beta_{j,it,0}$ are assumed to be independent and evolve according to a random walk process in the state equation of the above multilevel model. The terms also capture

the autoregressive time series nature of the data and hence can be specified as

$$\gamma_t = \gamma_{(t-1)} + \omega_t,$$

while

$$\beta_{j,it,0} = \beta_{j,i(t-1)+\varpi_{j,it}},$$

where $\omega_t \sim \mathrm{N}(0, 1/W)$ and $\varpi_{j,it} \sim \mathrm{N}(0, 1/\eta_j)$; that is, the errors are assumed to follow the normal distribution, where W and η_j are the precision parameters based on the inverse-gamma distribution. The reference level for the seasonal changes categorical predictor variable was taken to be winter since it is the season expected to have lower tick activity, and hence lower abundance. It therefore makes sense to contrast other seasons to the winter season. The type of habitat was also included in the model to assess the influence of habitat type on tick abundance and the baseline was taken to be the edge habitat type. Different multilevel mixed effects models taking the form in Equation 15 were fitted and compared using the DIC values. Model 1 denotes the multivariate model which assumes Poisson marginal distributions for all life stages in the response vector Y_{it}. Models 2 to 7 assume combinations of ZIP and Poisson mixed effects for tick life stages in the response vector. Model 8 denotes the multivariate ZIP model for tick life stage components in the response vector. The multivariate multilevel model is formulated by introducing the variance covariance matrix and jointly modelling the response vector assuming marginal count distributions as shown below:

Model 1:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it}); \\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poi}(\lambda_{N,it}); \\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 2:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it}); \\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poisson}(\lambda_{N,it}); \\
Y_{A,it}|\lambda_{A,it}, \delta_A &\sim \text{ZIP}(\lambda_{A,it}, \delta_A)
\end{aligned}
$$

Model 3:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it}); \\
Y_{N,it}|\lambda_{N,it}, \delta_N &\sim \text{ZIP}(\lambda_{N,it}, \delta_N); \\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 4:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it}, \delta_L &\sim \text{ZIP}(\lambda_{L,it}, \delta_N); \\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poi}(\lambda_{N,it}); \\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poisson}(\lambda_{A,it})
\end{aligned}
$$

Model 5:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it} &\sim \text{Poi}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

Model 6:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it});\\
Y_{N,it}|\lambda_{N,it} &\sim \text{Poi}(\lambda_{N,it});\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

Model 7:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it},\delta_N);\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it} &\sim \text{Poi}(\lambda_{A,it})
\end{aligned}
$$

Model 8:

$$
\begin{aligned}
Y_{L,it}|\lambda_{L,it},\delta_L &\sim \text{ZIP}(\lambda_{L,it},\delta_N);\\
Y_{N,it}|\lambda_{N,it},\delta_N &\sim \text{ZIP}(\lambda_{N,it},\delta_N);\\
Y_{A,it}|\lambda_{A,it},\delta_A &\sim \text{ZIP}(\lambda_{A,it},\delta_A)
\end{aligned}
$$

where $j = L, N$ or A, and δ_j the dispersion parameter for the jth life stage. All models were implemented through an approximate sampling based framework which provide a mechanism for Bayesian inference based on accurate approximations to the posterior distributions of the parameters. The approximate sample Bayesian framework requires the usual prior specification on the parameters. We assume a Wishart prior for $\boldsymbol{\Sigma}$, and a Log-Gamma prior for $\log(W)$ and $\log(\eta_j)$. The default hyperparameter specifications assume the default prior from the R-INLA package in R. The matrix $\boldsymbol{\Sigma}$ has a Wishart prior with $2 \times J + 1$ degrees of freedom and identity matrix as a prior precision matrix. The DIC values for the fitted model given by Equation 15 are presented in Table 4.

Table 4: DIC values for the fitted models 1 to 8.

Model	$\text{MDC}_j(\theta_{j,it})$	DIC	p_{DIC}
1	Poi Poi Poi	5007.80	798.47
4	ZIP Poi Poi	7513432.10	-332850.08
7	ZIP ZIP Poi	7539545.18	-308560.64

Even though we do not compare the INLA model with the MCMC model, we report that the INLA approach took 32.3 seconds (approx half a minute) to run using R-INLA on Intel(R) Core(TM) i7-8550U CPU 3.80 GHz with 8GB of RAM.

Models 2, 3, 5, 6 and 8 crashed when fitted to the data and therefore we show the DIC values for model 1, 4 and 7 in Table 4. It is clear from the table that model 1 fits the data well compared to models 4 and 7 since its DIC value (5007.80) is smaller compared to the DIC values for the other models.

4.4 Results

4.4.1 Fixed effects and correlations

Table 5 gives the posterior means of parameters of the multivariate multilevel model. We discover that there is a significant, strong and positive association between larvae and nymphs, larvae and adults, and nymphs and adults with posterior correlations $\rho_{LN} = 0.903$, $\rho_{LA} = 0.833$ and $\rho_{NA} = 0.858$, and 95% credible intervals (CI) $(0.806, 0.950)$, $(0.7163, 0.891)$, and $(0.668, 0.860)$, respectively.

Table 5: Estimated coefficients for covariates.

Covariates	Mean	SD	2.5%	97.5%
Habitat type (**ref:** Edge)				
Grass	1.042	1.033	0.949	1.137
Woods	0.755	0.926	0.551	0.970
Season (**ref:** Winter)				
Spring.L	0.927	2.000	-2.997	4.865
Spring.N	4.631	1.745	1.208	8.078
Spring.A	4.613	1.745	1.190	8.059
Summer.L	4.049	2.057	0.007	8.097
Summer.N	4.441	1.927	0.648	8.232
Summer.A	3.113	1.927	-0.680	6.904
Fall.L	4.631	1.964	0.768	8.488
Fall.N	4.205	1.888	0.489	7.911
Fall.A	-0.683	1.917	-4.459	3.076
Cosine term	-0.111	0.174	-0.453	0.231
Sine term	0.206	0.165	-0.118	0.530
Correlations				
ρ_{LN}	0.903	0.037	0.806	0.950
ρ_{LA}	0.833	0.042	0.716	0.891
ρ_{NA}	0.858	0.063	0.668	0.933

Variation in habitat type affects tick life cycle stages abundance in a similar manner. Comparing the grass habitat type with the edges and woods habitat types, the model estimated

higher and significant [1.042 and 0.755; CI: $(0.949, 1.137)$ and $(0.926, 0.970)$] tick life cycle stage abundances in the edges and woods habitat types, respectively. Variation in seasons affects variation in tick abundances in a life stage specific manner during fall, but it affect the abundance in a similar manner during summer and spring seasons. In comparison to winter season as the reference category, the model yields significant and higher [4.631 and 0.613; CI: $(1.208, 8.078)$ and $(1.190, 8.059)$] log mean abundances for nymphs and adults and non-significant [0.927; CI: $(-2.997, 4.865)$] log mean abundances for larvae in the spring season. Higher and significant [4.049 and 4.441; CI: $(0.007, 8.097)$ and $(0.648, 8.232)$] and non-significant [3.113; CI: $(-0.680, 6.904)$] in the summer season relative to the winter season. Higher and significant [4.631 and 4.205; CI: (0.768,8.488) and (0.489,7.911)] log mean abundances for larvae and nymphs and non-significant [-0.683; CI: $(-4.459, 3.076)$] log mean abundances for adults during the fall season as compared to the winter season. However, seasonal variations in the data were not significantly [-0.111 and 0.206; CI: $(-0.453, 0.231)$ and $(-0.118, 0.530)$] strengthened by the cosine and sine terms in the model.

4.4.2 Habitat type specific non-linear month effects

Figure 2 shows the model estimated non-linear month effects, modelled as the random walk of order one (RW1). The model estimated similar months effects across the three habitat types. The graphs show the posterior mean (solid line) and 95% credible intervals (dotted lines). The curves show an expected gradual increase in tick life cycle stages abundances from January to June, and a steady decrease between June and August, slight increase between August and September, sharp decrease from September to November, and a gradual increase from November to December.

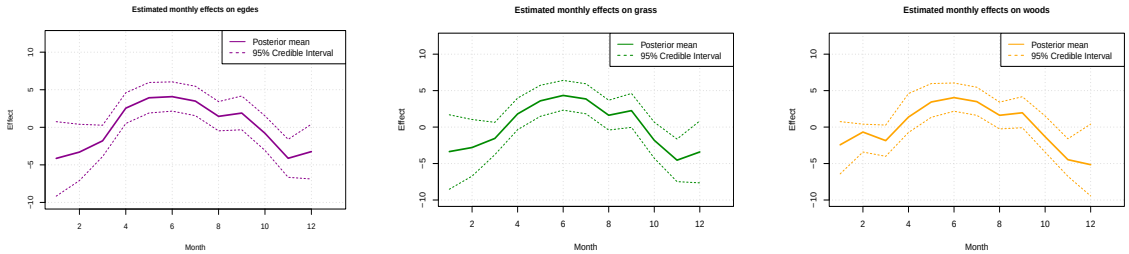


Figure 2: Model estimated non-linear month effects in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types.

4.4.3 Habitat type specific trends

Figure 3 shows the model estimated dynamic intercept ($\beta_{j,it,0}$), showing trends in the edges (purple curves), grass (green curves) and woods (brown curves) habitat type for larvae, nymph and adult tick life cycle stages. The fitted model poorly tracked the stochastic pattern across the three habitat types for larvae. This suggest that the larvae abundance is affected by environmental characteristics beyond those measured in this study. The top panel (row) are the estimated larvae trends showing constant fluctuations throughout the study period of 2009 through 2018 across the habitat types. The middle and bottom panels shows a declining nymph and adult tick trends since 2012 in the edges and woods habitat types, and an increasing trend in the grass habitat type.

5 Discussion

This study described the multilevel model for the vector of time series of counts using the R-INLA technique, which is effective and computationally feasible for large datasets. The aim of this study was to investigate the influence of environmental and temporal predictors of tick life cycle stages abundances, determine trends and demonstrate the computational speed of the INLA technique proposed by Rue et al. (2009) relative to the MCMC tech-

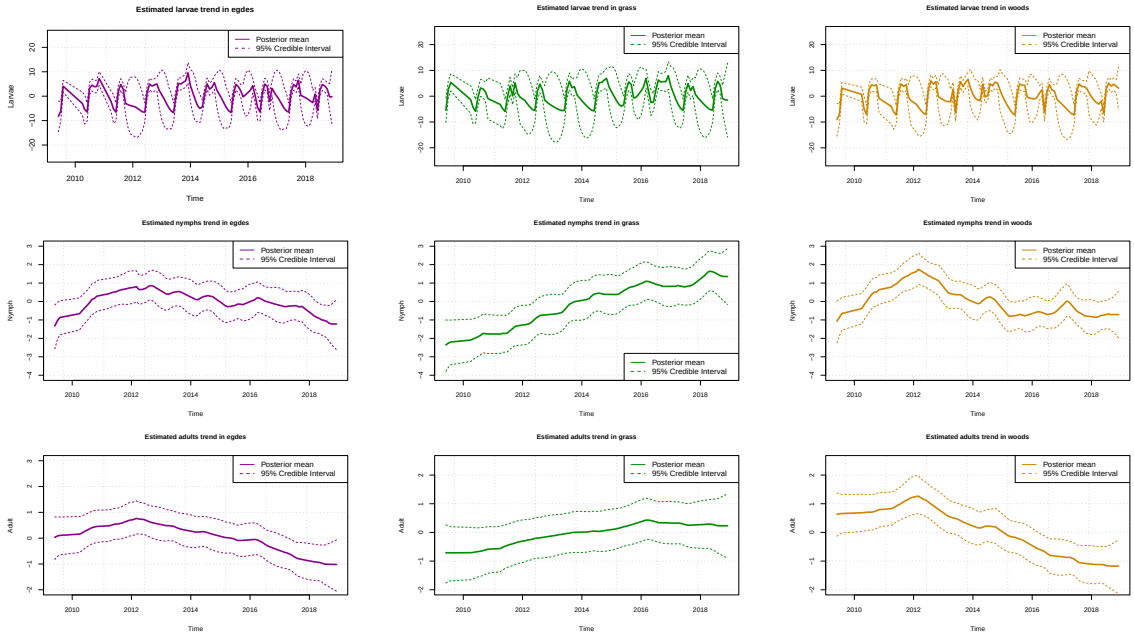


Figure 3: Estimated model trends in the edges (purple curve), grass (green curve) and woods (orange curve) habitat types.

nique. The multilevel model used in this study has seen applications in prescription counts of physicians in a large pharmaceutical company in Serhiyenko (2015); Serhiyenko et al. (2018) and ridesourcing data in NYC by Ravishanker et al. (2022) among other applications. To our knowledge, these models have not been applied to ecological data, specifically tick life stage count data. The MCMC is a large class of methods that enables inference in highly dimensional problems with unknown quantities and is able to handle complicated distributions. The literature on the methods is well documented in Gamerman and Lopes (2006); Robert et al. (1999); Chen et al. (2012), and these methods are commonly used for computing posterior quantities through the popular Gibbs sampling methods including the Metropolis-Hastings algorithm (Geman and Geman, 1984; Metropolis et al., 1953; Hastings, 1970). However, the MCMC technique is computationally intensive when complex models are involved, or when the data is large. As an alternative to the computationally intensive

MCMC Bayesian technique, the INLA technique for approximating posterior parameters was suggested and its feasibility demonstrated by simulating the data and estimating parameters, while taking into account the time taken to complete in INLA and in MCMC before models were applied to tick data. This study assumed tick life stage specific effects of the covariates on the distribution of tick stages. That is, we assumed that each life stage was affected differently from another tick life stage. However, the results in this study showed that monthly effects are similar across the three habitat types. Seasonal changes were different for each life stage and abundances were estimated to be higher in the grass and woods habitat types compared to the edges habitat type. The model estimated trends revealed an increasing trend for nymphs and adults in grassy habitat types. The results in this chapter are similar to those in Willis et al. (2012) and Lephoto et al. (2021).

6 Conclusion

The multilevel model successfully fitted tick life stage data and revealed that assuming common covariates effects on tick life stages may be unrealistic. However, this seems to be the case with monthly effects as similar trends were estimated. The assumption was unrealistic in the sense that factors such as sampling variation and different relationships across regions contribute to different responses to the same stimuli as one moves across regions and over a period of time. Spring and fall seasons affected life stages differently, though the effects of summer were similar for the three tick life-stages. The model is also beneficial in allowing for a mixture of count distributions in the vector of responses.

Declarations

Funding and/or Conflict of Interest/Competing Interests

This research was funded by NIH grant 1R01AI136035 as part of the joint National Institutes of Health, National Science Foundation and the United States Department of Agriculture (NIH-NSF-USDA) Ecology and Evolution of Infectious Diseases program. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The authors would like to declare that there is no conflict of interests.

Bibliography

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653.

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health*, 12(9):10536–10548.

CDC (2020). Guide to the surveillance of metastriate ticks (Acari: Ixodidae) and their pathogens in the United States.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Chib, S. and Winkelmann, R. (2001). Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435.

Clark, N. J. and Wells, K. (2023). Dynamic generalised additive models (DGAMs) for forecasting discrete ecological time series. *Methods in Ecology and Evolution*, 14(3):771–784.

De Clercq, E., Leta, S., Estrada-Peña, A., Madder, M., Adehan, S., and Vanwambeke, S. O. (2015). Species distribution modelling for Rhipicephalus microplus (Acari: Ixodidae) in Benin, West Africa: comparing datasets and modelling algorithms. *Preventive Veterinary Medicine*, 118(1):8–21.

de Oliveira, S. V., Romero-Alvarez, D., Martins, T. F., Dos Santos, J. P., Labruna, M. B., Gazeta, G. S., Escobar, L. E., and Gurgel-Gonçalves, R. (2017). Amblyomma ticks and future climate: range contraction due to climate warming. *Acta Tropica*, 176:340–348.

Estrada-Peña, A., de la Fuente, J., and Cabezas-Cruz, A. (2016). A comparison of the performance of regression models of Amblyomma americanum (L.)(Ixodidae) using life cycle or landscape data from administrative divisions. *Ticks and Tick-borne Diseases*, 7(4):624–630.

Estrada-Pena, A. and Venzal, J. M. (2007). Climate niches of tick species in the Mediterranean region: modeling of occurrence data, distributional constraints, and impact of climate change. *Journal of Medical Entomology*, 44(6):1130–1138.

Fan, X., Ma, R., Yue, C., Liu, J., Yue, B., Yang, W., Li, Y., Gu, J., Ayala, J. E., Bunker, D. E., et al. (2023). A snapshot of climate drivers and temporal variation of Ixodes ovatus abundance from a giant panda living in the wild. *International Journal for Parasitology: Parasites and Wildlife*, 20:162–169.

Ferkingstad, E., Held, L., and Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat*, 6(1):331–344.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Hadfield, J. (2010). MCMCglmm: Markov chain Monte Carlo methods for generalised linear mixed models. *Tutorial for MCMCglmm package in R*, 125.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97.

Kaizer, A., Foré, S., Kim, H.-J., and York, E. (2015). Modeling the biotic and abiotic factors that describe the number of active off-host Amblyomma americanum larvae. *Journal of Vector Ecology*, 40(1):1–10.

Khana, D., Rossen, L. M., Hedegaard, H., and Warner, M. (2018). A Bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-INLA. *Journal of Data Science*, 16(1):147.

Kjær, L. J., Soleng, A., Edgar, K. S., Lindstedt, H. E. H., Paulsen, K. M., Andreassen, Å. K., Korslund, L., Kjelland, V., Slettan, A., Stuen, S., et al. (2019). Predicting and mapping human risk of exposure to Ixodes ricinus nymphs using climatic and environmental data, Denmark, Norway and Sweden, 2016. *Eurosurveillance*, 24(9):1800101.

Kopsco, H. L., Smith, R. L., and Halsey, S. J. (2022). A scoping review of species distribution modeling methods for tick vectors. *Frontiers in Ecology and Evolution*, 10:893016.

Laaksonen, M., Sajanti, E., Sormunen, J. J., Penttinen, R., Hänninen, J., Ruohomäki, K., Sääksjärvi, I., Vesterinen, E. J., Vuorinen, I., Hytönen, J., et al. (2017). Crowdsourcing-based nationwide tick collection reveals the distribution of Ixodes ricinus and I. persulcatus and associated pathogens in Finland. *Emerging Microbes & Infections*, 6(1):1–7.

Lado, P., Smith, M. L., Carstens, B. C., and Klompen, H. (2020). Population genetic structure and demographic history of the Lone Star tick, Amblyomma americanum (Ixodida: Ixodidae): new evidence supporting old records. *Molecular Ecology*, 29(15):2810–2823.

Lephoto, T., Mwambi, H., Bodhlyera, O., and Gaff, H. (2021). Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients. *Geospatial Health*, 16(2).

Lippi, C. A., Gaff, H. D., White, A. L., St. John, H. K., Richards, A. L., and Ryan, S. J. (2021). Exploring the niche of Rickettsia montanensis (Rickettsiales: Rickettsiaceae) infection of the american dog tick (Acari: Ixodidae), using multiple species distribution model approaches. *Journal of Medical Entomology*, 58(3):1083–1092.

Liu, Y., Lund, R. B., Nordone, S. K., Yabsley, M. J., and McMahan, C. S. (2017). A Bayesian spatio-temporal model for forecasting the prevalence of antibodies to Ehrlichia species in domestic dogs within the contiguous United States. *Parasites & Vectors*, 10(1):1–14.

Lynen, G., Zeman, P., Bakuname, C., Di Giulio, G., Mtui, P., Sanka, P., and Jongejan, F. (2008). Shifts in the distributional ranges of boophilus ticks in tanzania: evidence that a parapatric boundary between boophilus microplus and b. decoloratus follows climate gradients. *Experimental and Applied Acarology*, 44:147–164.

Ma, D., Lun, X., Li, C., Zhou, R., Zhao, Z., Wang, J., Zhang, Q., and Liu, Q. (2021). Predicting the potential global distribution of Amblyomma americanum (Acari: Ixodidae) under near current and future climatic conditions, using the maximum entropy model. *Biology*, 10(10):1057.

Ma, J., Kockelman, K. M., and Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, 40(3):964–975.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nadolny, R. M., Wright, C. L., Sonenshine, D. E., Hynes, W. L., and Gaff, H. D. (2014). Ticks and spotted fever group rickettsiae of southeastern Virginia. *Ticks and Tick-borne Diseases*, 5(1):53–57.

Neupane, N., Goldbloom-Helzner, A., and Arab, A. (2021). Spatio-temporal modeling for confirmed cases of lyme disease in virginia. *Ticks and Tick-borne Diseases*, 12(6):101822.

Paull, S. H., Thibault, K. M., and Benson, A. L. (2022). Tick abundance, diversity and pathogen data collected by the national ecological observatory network. *Gigabyte*, 2022:1–11.

Qviller, L., Viljugrein, H., Loe, L. E., Meisingset, E. L., and Mysterud, A. (2016). The influence of red deer space use on the distribution of Ixodes ricinus ticks in the landscape. *Parasites & Vectors*, 9(1):1–9.

Raghavan, R., Barker, S., Cobos, M. E., Barker, D., Teo, E., Foley, D., Nakao, R., Lawrence, K., Heath, A., and Peterson, A. T. (2019). Potential spatial distribution of the newly introduced long-horned tick, Haemaphysalis longicornis in North America. *Scientific Reports*, 9(1):498.

Ravishanker, N., Raman, B., and Soyer, R. (2022). *Dynamic time series models using R-INLA: an applied perspective*. CRC Press.

Requena-García, F., Cabrero-Sañudo, F., Olmeda-García, S., González, J., and Valcárcel, F. (2017). Influence of environmental temperature and humidity on questing ticks in central Spain. *Experimental and Applied Acarology*, 71:277–290.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Rochlin, I., Egizi, A., and Ginsberg, H. S. (2023). Modeling of historical and current distributions of Lone Star tick, Amblyomma americanum (Acari: Ixodidae), is consistent with ancestral range recovery. *Experimental and Applied Acarology*, 89(1):85–103.

Rochlin, I., Egizi, A., and Lindström, A. (2022). The original scientific description of the Lone Star tick (Amblyomma americanum, Acari: Ixodidae) and implications for the species' past and future geographic distributions. *Journal of Medical Entomology*, 59(2):412–420.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sagurova, I., Ludwig, A., Ogden, N. H., Pelcat, Y., Dueymes, G., and Gachon, P. (2019). Predicted northward expansion of the geographic range of the tick vector Amblyomma

americanum in North America under future climate conditions. *Environmental Health Perspectives*, 127(10):107014.

Serhiyenko, V. (2015). *Dynamic modeling of multivariate counts-fitting, diagnostics, and applications*. PhD thesis, University of Connecticut, Storrs, CT, USA.

Serhiyenko, V., Ravishanker, N., and Venkatesan, R. (2018). Multi-stage multivariate modeling of temporal patterns in prescription counts for competing drugs in a therapeutic category. *Applied Stochastic Models in Business and Industry*, 34(1):61–78.

Simpson, D. T., Teague, M. S., Weeks, J. K., Kaup, B. Z., Kerscher, O., and Leu, M. (2019). Habitat amount, quality, and fragmentation associated with prevalence of the tick-borne pathogen Ehrlichia chaffeensis and occupancy dynamics of its vector, Amblyomma americanum. *Landscape Ecology*, 34:2435–2449.

Sonenshine, D. E. (1979). Insects of Virginia no. 13. *Ticks of Virginia (Acari: Metastigmata). Research Division Bulletin*, 139.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for Bayesian inference with INLA. *Computational Statistics & Data Analysis*, 181:107–692.

Wang, F., Li, H., Wang, H., and Li, Y. (2023). Spatial correlated incidence modeling with zero inflation. *Biometrical Journal*, 65(4):2200090.

Willis, D., Carter, R., Murdock, C., and Blair, B. (2012). Relationship between habitat type, fire frequency, and Amblyomma americanum populations in east-central Alabama. *Journal of Vector Ecology*, 37(2):373–381.

Wimberly, M. C., Baer, A. D., and Yabsley, M. J. (2008). Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*, 7(1):1–14.

Wimms, C., Aljundi, E., and Halsey, S. J. (2023). Regional dynamics of tick vectors of human disease. *Current Opinion in Insect Science*, 55:101006.

Zannou, O. M., Ouedraogo, A. S., Biguezoton, A. S., Abatih, E., Coral-Almeida, M., Farougou, S., Yao, K. P., Lempereur, L., and Saegerman, C. (2021). Models for studying the distribution of ticks and tick-borne diseases in animals: a systematic review and a meta-analysis with a focus on Africa. *Pathogens*, 10(7):893.

Zipkin, E. F. (2012). *Hierarchical models for the anlaysis of species distributions and abundances: development and applications*. PhD thesis, University of Mayland, College Park, MD, USA.

Multivariate Bayesian dynamic tick life stage time series modelling and assessing impacts of sampled sites

Thabo Lephoto[1*], Henry Mwambi[1], Oliver Bodhlyera[1] and Holly Gaff[2]

[1]School of Mathematics Statistics and Computer Science, University of KwaZulu-Natal, KwaZulu-Natal Province, Private Bag X01, 3201, South Africa

[2]Department of Biological Sciences, Old Dominion University, Norfolk, United States

Abstract

Background: Understanding the influence and consequences of variation in the abundance of organisms has been a long-standing goal in ecological studies. An increase in tick abundance and tick-borne diseases constitute a growing threat to public health. Studies defining species' variations within counties and addressing hierarchical dynamic time series of tick life stage abundances remain limited. Our objective was to assess the influence of temporal and environmental effects and identify tick hotspot areas in sampled counties in Virginia, United States.

Methods: We proposed extending the Poisson and negative binomial Bayesian hi-

erarchical dynamic time series models that enable modelling tick life stage abundances, finding correlations between them, determining temporal effects and identifying tick hotspot areas. Our novel hierarchical dynamic time series model, which includes sample sites as random effects, is the first to identify tick hotspot areas and assess the temporal and environmental influence of tick stage abundances.

Results: The results showed that higher tick abundance was expected in summer compared to winter. The model-estimated month trend increases between January and September, while the model-estimated year trend decreases over the study period from 2009 to 2018. The analysis also revealed tick hotspot areas in Chesapeake, York, Portsmouth, and Northampton counties in Virginia.

Conclusion: Although the model showed susceptibility due to small sample sizes, we conclude that the model was successful in fitting and estimating posterior parameters when applied to the tick data.

Keywords: Lone Star tick stages, multivariate analysis,tick life-stages, environmental and temporal effects, tick hotspot areas.

Background

Increasing tick abundance and tick-borne pathogens constitute a growing threat to public health (Laaksonen et al., 2017; Paull et al., 2022; Wimms et al., 2023). For example in the USA, Nadolny et al. (2014) recorded more than 66 000 questing tick populations in the southeastern Virginia, comprising 7 species from a variety of habitats, with *Amblyomma americanum* (Lone Star tick) constituting over 95% of the ticks collected. Our research focused on modeling *A. americanum* tick life-stage abundance collected over time from three different habitat types namely: edges, woods and grass, along 12 random sample sites. The

objective of this study was to identify abundance levels of *A. americanum* variations within sample sites from eight counties in the state of Virginia, United States. We also wanted to investigate the influence of environmental and temporal predictors and estimate time trends.

Five major tick species were reported to be active and responsible for transmitting a variety of pathogens of both human and animals. These were *Ixodes scapularis*, *A. americanum*, *Dermacentor variabilis*, *Amblyomma maculatum*, and *Haemaphysalis longicornis* (Nadolny and Gaff, 2018). The *H. longicornis* is of high veterinary importance for transmission of *Theileria orientalis* Ikeda in cattles even though it is not a key human-biting species (Oakes et al., 2019). The most common human-biting tick in Virginia is the *A. americanum*, and it transmits several pathogens of medical and veterinary significance, including *Ehrlichia chaffeensis* (the causative agent of human monocytic ehrlichiosis), Heartland virus (HRTV) and Bourbon virus (BRBV) among others (Goddard and Varela-Stokes, 2009; Savage et al., 2013, 2017)

Due to the effects of climate change and transportation by hosts, tick ranges remain ever-changing. In recent years the *A. americanum* species has expanded its range into north-eastern and midwestern United States and nothward into Canada (Molaei et al., 2019). A recent study in Kentucky reported detection of *Ehrlichia chaffeensis* in 32 counties out of the 77 sampled counties in 2019 through 2020 (Pasternak and Palli, 2023). A 2011 Centers for Disease Control and Prevention (CDC) report indicated that there were 863 human cases of ehrlichiosis in the United States. An increase has been observed over the seven-year period, with human cases steadily increasing to 1 832, reporting 1 799 as a result of *E. chaffeensis* infection (Adams et al., 2013; CDC, 2021).

A. americanum ticks have long been identified as nuisance biters because of their aggressive

questing behaviour (Hair and Howell, 1970; Merten et al., 2000; Otálora-Luna et al., 2022). Despite its prominence as a nuisance biter and vector of human pathogens, efforts to define the species' variations within counties remain limited (Benham et al., 2021). Therefore, one of the strategies to help reduce tick-borne diseases is to have an in-depth analysis of the life stages abundance patterns and factors contributing in each of the randomly sampled locations. The aim of this study was to unmask the heterogeneity within southeastern counties in Virginia. Thus, we modelled tick life stage abundance variations within the randomly sampled locations. The multivariate time series of tick counts was modelled as a function of season, habitat type, non-linear effects of month and year, and sampling site random effects.

In the next Section we describe the data for this study, give a modelling framework in Section , and simulate the data in Section . We describe model diagnostics in Section , give model results in Section , and conclude in Section .

Methods

Data description

This study used data collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, Virginia, United States. Data collection was done at least once a month on varying days of the week in southeastern Virginia from May 2009 through December 2018. Ticks were collected using standard flagging techniques (CDC, 2020) along established transects and species and life-stages were identified according to (Sonenshine, 1979). Larvae-, nymph- and adult-stage count data were recorded for each edge, wooded and grassy habitat type in York, Chesapeake, Norfolk, Hampton, Isle of Wight,

Virginia Beach, Northampton and Portsmouth counties through the four seasons of the year. The datafile included sample sites/locations which designate the area where the data were collected. Sample sites were abbreviated JC and ST (Jacobson Track and Stephens Tract) in Chesapeake, LA (Langley) in Hampton, BW (Blackwater Ecological Preserve) in Isle of Wight, WS (Weyanoke) in Norfolk, KP (Kiptopeke State Park) in Northampton, HC and PC (Hoffler Creek Wildlife Preserve and Paradise Creek Nature Park) in Portsmouth, BB and OD (Back Bay NWR and Oceana/Dam Neck) in Virginia Beach, and CA and NN (Naval Supply Center, Cheatham Annex and Newport News Park) in York.

Statistical modelling

For practical situations where responses arise as a vector of counts that vary across different observational sample sites/locations, univariate Poisson regression models for each of the components in the response vector cannot account for the association or dependence among the components of the response vector. The dependence may be due to omitted variables which simultaneously affect the response vector Aitchison and Ho (1989); Chib and Winkelmann (2001); Ma et al. (2008), and therefore multivariate modelling approach is needed. However, in the case of tick stages which define the components of the response vector, dependence between the stages is expected, because individuals in stage $k+1$ depend on the survival of individuals from stage k. Suppose the response is a $J-$variate vector of counts from n habitat types at $t = 1, \ldots, T$ time points. A hierarchical multivariate Poisson model is defined as follows for $j = 1, \ldots, J$; $i = 1, \ldots, n$; $t = 1, \ldots, T$:

$$\boldsymbol{Y}_{it}|\boldsymbol{\lambda}_{it} \ \sim \ \mathrm{MVP}_J(\boldsymbol{\lambda}_{it}), \tag{1}$$

$$\log(\lambda_{j,it}) = \mathbf{D}'_{j,it}\boldsymbol{\gamma}_{j,it} + \boldsymbol{S}'_{j,it}\boldsymbol{\beta}_j + \upsilon_{j,it}, \tag{2}$$

where $\boldsymbol{D}_{j,it} = (D_{j,it,1}, \ldots, D_{j,it,a})'$ is an $a-$dimensional vector of exogenous (dynamic) location-time varying predictors with the corresponding dynamic coefficients $\boldsymbol{\gamma}_{j,it} = (\gamma_{j,it,1}, \ldots, \gamma_{j,it,a})$, $\boldsymbol{S}_{j,it} = (S_{j,it,1}, \ldots, S_{j,it,c})$ is an $c-$dimensional vector of exogenous predictors with static coefficients $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,c})$, and the error term $\upsilon_{j,it}$ incorporates the dependence between the components of the response vector by assuming a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{V}$ for each habitat type i at time t. For identifiability, we assume that if the model includes the habitat type and time varying intercept $\gamma_{j,it,1} = 1$, then there is no intercept $\beta_{j,1} = 1$ in the static part.

For all i and t, we assume that the errors $\upsilon_{j,it} \sim \mathrm{N}(0, V_{jj})$, and that the correlation $Corr(\upsilon_{j,it}, \upsilon_{\ell,it}) = 1$ if $j = \ell$ and $\rho_{j\ell}$ if $j \neq \ell$, j, $\ell = 1, \ldots, J$. Suppose $\mathrm{Cov}(\boldsymbol{\upsilon}_{it}) = \boldsymbol{V} = \{V_{j\ell}\}$ and $\mathrm{Corr}(\upsilon_{it}) = \boldsymbol{R} = \{\rho_{j\ell}\}$. Let $\boldsymbol{M}_\upsilon = \mathrm{diag}(V_{11}, \ldots, V_{JJ})$. Note that $\boldsymbol{V}, \boldsymbol{R}$ and $\boldsymbol{M}_\upsilon$ are $J \times J$ symmetric positive definite matrices, and $\boldsymbol{R} = \boldsymbol{M}_\upsilon^{-1/2}\boldsymbol{V}\boldsymbol{D}_\upsilon^{-1/2}$. For example, if $j = 2$, then the covariance matrix is:

$$\boldsymbol{V} = \begin{pmatrix} \upsilon_{11} & \upsilon_{12} \\ \upsilon_{12} & \upsilon_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\tau_{11}} & \frac{R_{12}}{\sqrt{\tau_{11}\tau_{22}}} \\ \frac{R_{12}}{\sqrt{\tau_{11}\tau_{22}}} & \frac{1}{\tau_{22}} \end{pmatrix}$$

where the marginal precisions of $\upsilon_{1,t}$ and $\upsilon_{2,t}$ are $\tau_{11} = 1/V_{11}$ and $\tau_{22} = 1/V_{22}$, the correlation between $\upsilon_{1,t}$ and $\upsilon_{2,t}$ is $\rho_{12} = V_{12}/\sqrt{V_{11}V_{22}}$. We assume that the precision matrix $\boldsymbol{V}^{-1}$ follows an inverse Wishart distribution with r degrees of freedom and scale matrix $\boldsymbol{\Sigma}^{-1}$; that is $\boldsymbol{V}^{-1} \sim \mathrm{IW}(r, \boldsymbol{\Sigma}^{-1})$.

A detailed exposition of the models described in 2 can be found in Fokianos (2015) and Ravishanker et al. (2022) among many others. We extend the model in 2 to include the random effect, b_k, which quantify the variation in abundance that is accounted for by the m sample sites' variance. The resulting multivariate dynamic Poisson model is defined as follows for $j = 1, \ldots, J$; $i = 1, \ldots, n$; $k = 1, \ldots, m$; $t = 1, \ldots, T$:

$$\boldsymbol{Y}_{it} | \boldsymbol{\lambda}_{it} \ \sim \ \text{MVP}_J(\boldsymbol{\lambda}_{it}), \tag{3}$$

$$\log(\lambda_{j,ikt}) = \mathbf{D}'_{j,ikt}\boldsymbol{\gamma}_{j,ikt} + \boldsymbol{S}'_{j,ikt}\boldsymbol{\beta}_j + \upsilon_{j,ikt} + b_k, \tag{4}$$

where $\boldsymbol{D}, \boldsymbol{\gamma}, \boldsymbol{S}$, $\boldsymbol{\beta}$ and $\upsilon_{j,it}$ are defined as in 2, and b_k is the random effects term which follows independent and identical normal distribution (iid) such that, $b_k \ \sim \ \text{N}(\mathbf{0}, \boldsymbol{W}_{b_k})$. For $j = 2$, the diagonal covariance matrix $\boldsymbol{W}_{b_k}$ is given as:

$$\boldsymbol{W}_{b_k} = \begin{pmatrix} \omega_{11} & 0 \\ 0 & \omega_{22} \end{pmatrix}$$

where the diagonal precisions of ω_{11} and ω_{22} are $\varrho_{11} = 1/\omega_{11}$ and $\varrho_{22} = 1/\omega_{22}$.

Simulations

This section presents a simulation study to mimic the observed data generation process. We assume a two-component outcome vector for simplicity. Let $\boldsymbol{Y}_{ik} = (Y_{1ik}, Y_{2ik})$ represent a bivariate ($J = 2$) vector of counts data observed at $k = 1, \ldots, m$ random sample sites that are within $i = 1, \ldots, n$ locations. We simulate the data according to the simplified model

described as follows:

$$\boldsymbol{Y}_{ik}|\boldsymbol{\lambda}_{ik} \sim \text{MVP}_J(\boldsymbol{\lambda}_{ik}), \tag{5}$$

$$\log(\lambda_{j,ik}) = \boldsymbol{S}_{j,ik}\boldsymbol{\beta}_{j,ik} + \upsilon_{j,i} + b_k, \tag{6}$$

where $\boldsymbol{S}$ and $\boldsymbol{\beta}$ represent vectors of static predictors and coefficients respectively, each simulated from a multivariate normal distribution, such that $\boldsymbol{S} \sim \text{N}(\boldsymbol{0},\boldsymbol{V}_s)$ and $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{0},\boldsymbol{V}_\beta)$. The prior distribution for $\boldsymbol{S}_i \sim \text{N}(\boldsymbol{0}, \text{diag}(10^3))$ and $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{0}, \text{diag}(10^3))$. The error term $\upsilon_{j,i}$ is simulated from $\text{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \text{diag}(V_{11}, V_{22})$ and $\rho_{12} = \rho_{21}$ for $i = 1, \ldots, n$. We let the precisions terms be $V_{11} = V_{22} = 0.4$, and $\rho_{12} = 0.90$.

The errors in the random effects term b_k are simulated from a multivariate normal i.i.d, that is $b_k \sim \text{N}(\boldsymbol{0}, \boldsymbol{W}_{b_k})$ with $\boldsymbol{W}_{b_k} = \text{diag}(\omega_{11}, \omega_{22})$, where the precision parameters are based on the inverse-gamma distribution; that is $\varrho_{11} \sim \text{IG}(a_1, b_1)$ and $\varrho_{22} \sim \text{IG}(a_2, b_2)$ (Grilli et al., 2015). The hyperpriors a and b are the shape and rate parameters in the gamma distribution respectively. Since the choice of priors can have a significant impact on the posterior distributions of the model parameters and model performance can be sensitive to the choice of sample site-specific random effects variance priors Gelman (2006), we considered three hyperpriors for ω_{11} and ω_{22} in $\boldsymbol{W}_{b_k}$ through the precision parameters ϱ_{11} and ϱ_{22} in the precision matrix $\boldsymbol{\Omega}_{b_k}^{-1}$. The hyperpriors are as follows:

- $\Gamma(1, 0.0005)$, the default choice of the `inla()` function in the R-INLA package (Rue et al., 2009);

- $\Gamma(0.001, 0.001)$, the default choice of the BUGS software (Lunn et al., 2012) and
- $\Gamma(0.5, 0.0164)$ proposed by Fong et al. (2010), corresponding to random effects b_k with marginal Cauchy distribution such that $e^{b_k} \in [0.1, 10]$ with probability 0.95.

Bayesian estimation was carried out using the `inla()` function from the R-INLA package by (Rue et al., 2009). The aim was to investigate the influence of n and m on the posterior estimates. Since the tick data under study has $n = 3$ habitat types and $m = 8$ or 9 random sample sites, we consider two scenarios in estimating the posterior parameters: In the first scenario we allow the location type of size n to change, while the random sample site of size is fixed at $m = 10$ while $n = 3, 10, 20, 50, 100$ and 1 000. In the second scenario, $n = 5\ 000$ and $m = 10, 50, 100, 500, 1\ 000$ and 2 500. The second scenario investigates the effects in the posterior parameters using the hyperprior in the `inla()` function by Rue et al. (2009), BUGS software by (Lunn et al., 2012) and that by Fong et al. (2010). We set a large value of $n = 5\ 000$ and observe the change in the posterior parameter as $m = 10, 50, 100, 500, 1\ 000$ and 2 500. The results of the simulation are presented in Table 1.

Simulation results

Table 1 presents the results of the simulations for different values of n when m is fixed in scenario 1, and simulations for different values of m when n is fixed in scenario 2. In scenario 1 we note that posterior estimates are close to the true values when $n \geq 50$ and $m = 10$, and that credible intervals become narrower as n increases. In scenario 2, posterior estimates of the sample site-specific random effects are close to the true values for $n \geq 1\ 000$ for INLA and BUGS hyperpriors as compared to the BUGS hyperpriors. Credible intervals are narrower throughout the values of m, and the results show that posterior estimates are

approximately equal to the true values when there is $\frac{m}{n} = \frac{1\ 000}{5\ 000} \approx 20\%$ of m random sample sites that are within n habitat types locations. The results of the simulation reveal that posterior estimates are not accurate for $n \leq 20$ and $m = 10$, and these sizes include the sizes for our tick data.

Table 1: Estimated posterior parameters in the simulated model.

Parameter	**SCENARIO 1** ($m = 10$)				**SCENARIO 2** ($n = 5000$)									**True value**
	INLA					INLA			BUGS			FONG		
	n	Posterior mean	95% CI		m	Posterior mean	95% CI		Posterior mean	95% CI		Posterior mean	95% CI	
			2.5%	97.5%			2.5%	97.5%		2.5%	97.5%		2.5%	97.5%	
β_1		4.773	1.995	11.007		5.028	4.632	5.459	5.024	4.743	5.323	5.026	4.686	5.394	5.0
β_2		3.585	1.458	8.577		4.967	4.609	5.351	4.964	4.628	5.326	4.968	4.590	5.375	5.0
V_{11}		0.232	0.087	1.089		0.410	0.386	0.435	0.410	0.385	0.435	0.411	0.387	0.437	0.4
V_{22}	3	0.220	0.083	1.002	10	0.424	0.400	0.449	0.425	0.399	0.451	0.424	0.400	0.451	0.4
ρ_{12}		-0.039	-0.661	0.597		0.876	0.858	0.895	0.877	0.857	0.894	0.877	0.858	0.896	0.9
ω_{11}		0.000	0.000	0.004		0.019	0.014	0.027	0.009	0.006	0.015	0.013	0.005	0.036	0.01
ω_{22}		0.000	0.000	0.004		0.015	0.011	0.022	0.014	0.009	0.020	0.018	0.011	0.031	0.01
β_1		3.145	1.936	4.980		5.031	4.917	5.147	5.032	4.915	5.150	5.032	4.915	5.150	5.0
β_2		2.785	1.249	5.424		4.969	4.855	5.087	4.974	4.857	5.093	4.974	4.857	5.093	5.0
V_{11}		0.201	0.084	0.695		0.409	0.386	0.432	0.408	0.387	0.432	0.408	0.387	0.432	0.4
V_{22}	10	0.613	0.235	2.366	50	0.425	0.400	0.449	0.423	0.401	0.448	0.423	0.401	0.448	0.4
ρ_{12}		0.304	-0.394	0.824		0.877	0.859	0.895	0.878	0.860	0.894	0.878	0.860	0.894	0.9
ω_{11}		0.000	0.000	0.004		0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.01
ω_{22}		0.000	0.000	0.004		0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.01
β_1		3.667	2.454	5.292		5.033	4.920	5.148	5.032	4.914	5.152	5.030	4.903	5.159	5.0
β_2		3.337	2.220	4.831		4.972	4.859	5.087	4.972	4.855	5.091	4.974	4.850	5.100	5.0
V_{11}		0.383	0.185	0.974		0.407	0.384	0.433	0.405	0.382	0.429	0.407	0.385	0.431	0.4
V_{22}	20	0.370	0.178	0.984	100	0.423	0.400	0.449	0.426	0.401	0.451	0.419	0.398	0.444	0.4
ρ_{12}		0.596	0.161	0.872		0.878	0.859	0.895	0.876	0.858	0.895	0.880	0.863	0.895	0.9
ω_{11}		0.001	0.000	0.004		0.001	0.000	0.003	0.002	0.001	0.004	0.005	0.003	0.007	0.01
ω_{22}		0.000	0.000	0.004		0.001	0.000	0.002	0.001	0.001	0.003	0.004	0.002	0.007	0.01
β_1		4.843	3.907	5.925		5.030	4.917	5.145	5.029	4.916	5.144	5.033	4.917	5.150	5.0
β_2		4.189	3.240	5.317		4.975	4.863	5.087	4.972	4.861	5.084	4.966	4.851	5.082	5.0
V_{11}		0.293	0.170	0.551		0.405	0.382	0.429	0.407	0.384	0.432	0.400	0.378	0.427	0.4
V_{22}	50	0.463	0.277	0.841	500	0.421	0.398	0.446	0.422	0.400	0.447	0.421	0.397	0.447	0.4
ρ_{12}		0.724	0.484	0.879		0.884	0.864	0.901	0.880	0.862	0.897	0.886	0.868	0.904	0.9
ω_{11}		0.000	0.000	0.006		0.005	0.003	0.009	0.001	0.001	0.002	0.008	0.005	0.011	0.01
ω_{22}		0.000	0.000	0.006		0.003	0.002	0.005	0.001	0.000	0.001	0.005	0.004	0.008	0.01
β_1		4.383	3.640	5.225		5.017	4.906	5.131	5.029	4.916	5.144	5.038	4.921	5.155	5.0
β_2		4.610	3.963	5.325		4.963	4.853	5.076	4.973	4.860	5.087	4.962	4.849	5.077	5.0
V_{11}		0.517	0.359	0.775		0.402	0.378	0.426	0.405	0.382	0.431	0.397	0.374	0.430	0.4
V_{22}	100	0.297	0.199	0.472	1000	0.416	0.393	0.442	0.421	0.399	0.446	0.418	0.395	0.443	0.4
ρ_{12}		0.738	0.555	0.867		0.895	0.871	0.926	0.882	0.858	0.904	0.890	0.872	0.907	0.9
ω_{11}		0.000	0.000	0.004		0.010	0.007	0.013	0.003	0.002	0.008	0.010	0.007	0.014	0.01
ω_{22}		0.000	0.000	0.004		0.009	0.006	0.012	0.002	0.001	0.005	0.010	0.007	0.013	0.01
β_1		5.028	4.632	5.459		5.029	4.920	5.140	5.041	4.930	5.153	5.024	4.913	5.138	5.0
β_2		4.967	4.609	5.351		4.963	4.853	5.075	4.971	4.861	5.083	4.970	4.859	5.083	5.0
V_{11}		0.410	0.386	0.435		0.403	0.374	0.428	0.400	0.378	0.429	0.404	0.376	0.429	0.4
V_{22}	1000	0.424	0.400	0.449	2500	0.418	0.391	0.444	0.418	0.395	0.447	0.415	0.391	0.440	0.4
ρ_{12}		0.876	0.858	0.895		0.889	0.870	0.908	0.887	0.868	0.904	0.894	0.873	0.911	0.9
ω_{11}		0.019	0.014	0.027		0.008	0.005	0.011	0.007	0.004	0.012	0.010	0.007	0.016	0.01
ω_{22}		0.015	0.011	0.022		0.007	0.005	0.011	0.004	0.002	0.010	0.009	0.005	0.015	0.01

Model application on tick life stage data

In order to understand the dynamic causes and consequences of variation in the abundance of ticks, multivariate hierarchical dynamic and static count data model, which include sample site random effects terms, were used to model the tick life-stage data. In the application, we let $\boldsymbol{Y}_{ikt} = (larvae, nymphs, adults)'$ be a $3-$dimensional time series vector of tick counts collected from location $i = 1, \ldots, 3$, comprising $k = 1, \ldots, 12$ random sampling sites over $t = 1, \ldots, 10$ years time points. Suppose that counts $\boldsymbol{Y}_{ikt}$ follow a multivariate Poisson (MVP) or the negative binomial distribution (MVNB). The model can be written as:

$$\boldsymbol{Y}_{ikt}|\boldsymbol{\delta}, \boldsymbol{\lambda}_{ikt} \sim \begin{cases} \text{MVP}_J(\boldsymbol{y}_{ikt}|\boldsymbol{\lambda}_{ikt}); \text{ when } \boldsymbol{\delta} = \mathbf{1}; \text{ and/or} \\ \text{MVNB}_J(\boldsymbol{\lambda}_{ikt}, \boldsymbol{\delta}); \text{ otherwise,} \end{cases} \tag{7}$$

where $\boldsymbol{\lambda}_{ikt}$ and $\boldsymbol{\delta}$ are the vectors of mean and dispersion parameters, respectively. The mean count vector of parameters $\boldsymbol{\lambda}_{ikt}$ is linked to the dynamic and static covariates through the log link function as follows:

$$\begin{aligned} \log(\lambda_{j,ikt}) &= \boldsymbol{\beta}_0 + \boldsymbol{Month}_{it} + \boldsymbol{Year}_{it} \\ &\quad + \boldsymbol{\beta}_1 I(Spring = 1)_{it} + \boldsymbol{\beta}_2 I(Summer = 2)_{it} + \boldsymbol{\beta}_3 I(Fall = 3)_{it} \\ &\quad + \boldsymbol{\beta}_4 I(Grass = 2)_i + \boldsymbol{\beta}_5 I(Woods = 3)_i \\ &\quad + \boldsymbol{\beta}_6 Cosine_t + \boldsymbol{\beta}_7 Sine_t + \upsilon_t + b_k \\ &= \boldsymbol{D}'_{j,ikt}\boldsymbol{\gamma}_{j,ikt} + \boldsymbol{S}'_{j,ikt}\boldsymbol{\beta}_j + \upsilon_{j,t} + b_k, \end{aligned} \tag{8}$$

where the $\boldsymbol{D}$, $\boldsymbol{\gamma}$, $\boldsymbol{S}$, $\boldsymbol{\beta}$, $\boldsymbol{\upsilon}$ and b are defined as in 2 and 4. Suppose that the number of coefficients in the dynamic and static vectors are given by $p_d = \sum_{j=1}^{J} a_j$ and $p_s = \sum_{j=1}^{J} b_j$

respectively, and let $\boldsymbol{\pi}_{it}$ be a p_d-dimensional vector constructed by stacking the a_j coefficients, $\boldsymbol{\gamma}_{it}$. That is,

$$\boldsymbol{\pi}_{it} = \boldsymbol{\theta}_t + \boldsymbol{\upsilon}_t \tag{9}$$

where $\boldsymbol{\theta}_t$ is the state parameter and $\boldsymbol{\upsilon}_t \sim \text{N}(\mathbf{0}, \boldsymbol{V})$. The state equation is given as

$$\boldsymbol{\theta}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\alpha}_t \tag{10}$$

where $\boldsymbol{G}$ is a $p_d \times p_d$ state transition matrix, and the state errors $\boldsymbol{\alpha}_t$ are assumed to be i.i.d $\text{N}(\mathbf{0}, \boldsymbol{\Lambda})$. Thus, the hierarchical equation in 8 can be written in a more compact form as follows:

$$\log(\lambda_{j,ikt}) = \boldsymbol{I}\boldsymbol{\pi}_{j,ikt} + \boldsymbol{\beta}\boldsymbol{S}_{j,ikt} + b_k. \tag{11}$$

We assume multivariate normal priors for the static predictor $\boldsymbol{\beta}_j$ as $\text{N}(\mathbf{0}, \text{diag}(10^3))$, and inverse-Wishart priors in $\boldsymbol{V}_i$, $\boldsymbol{\Lambda}_j$, and $\boldsymbol{W}_{b_k}$ for $\boldsymbol{\upsilon}_t$, $\boldsymbol{\alpha}_t$ and b_k respectively.

The reference category for the habitat type covariate is taken to be 1 corresponding to the edge habitat type, while the season variable is taken to be the winter season corresponding to low tick activity. Month and year covariates entered the model as continuous variables and we assume that they evolve according to a random walk process, and the sample sites' covariates were taken as random effects.

We employed Bayesian estimation using the R-INLA package by Rue et al. (2009) given its fast computation time compared to the MCMC counterpart. The Bayesian approach treats parameters as unknown random variables having prior probability distributions which are updated with data leading to posterior distributions in contrast to how likelihood methods treat parameters as fixed constants. Here, the likelihood of the model describes the data

generating process given the parameters, while the prior usually reflects any previous information about the model parameters. Consequently, when the prior knowledge is scarce then we assume vague or non-informative priors so that the posterior distribution is driven by the observed data (Gelman, 2006).

Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \bar{D}(\theta) + p_D$ where $\bar{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter et al. (2002) state that p_D values less than zero indicate substantial conflict between the prior and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of deviance suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter et al. (2002).

Results

To identify factors associated with tick life stage abundances, several models were fitted assuming the Poisson and the negative binomial multivariate distributions. Since the choice of priors can have an important impact on the posterior distributions of the model parameters and the performance of the model can be sensitive to the choice of the location-specific random effects variance priors Gelman (2006), we consided three inverse-gamma distribution priors for sample site random effects term $b_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}_{b_k})$, postulated through the preci-

sion matrix $\boldsymbol{\Omega}_{b_k}^{-1}$, that is $\boldsymbol{W}_{b_k} \sim IW(r_{b_k}, \boldsymbol{\Omega}_{b_k}^{-1})$. The fitted model included the intercept, sinusoidal terms, season and habitat type fixed effects, month and year non-linear effects, and the abundance sample site random effects. The model was fitted using three priors of the hyperparmeter for the variance of sample site random effects. Figure 1 displays tick abundance variations in the posterior densities of the hyperpriors. The plots suggest that the Rue et al. (2009) prior (black curve) was better in modelling the tick life stages since its log-likelihood curve was higher than the other curves. However, the DIC value of the Lunn et al. (2012) prior was smaller for the NB model compared to other models (Table 2), suggesting the results of this prior to be preferred for our multivariate data. We note that the posterior estimates of the models are similar and are presented in Table 3.

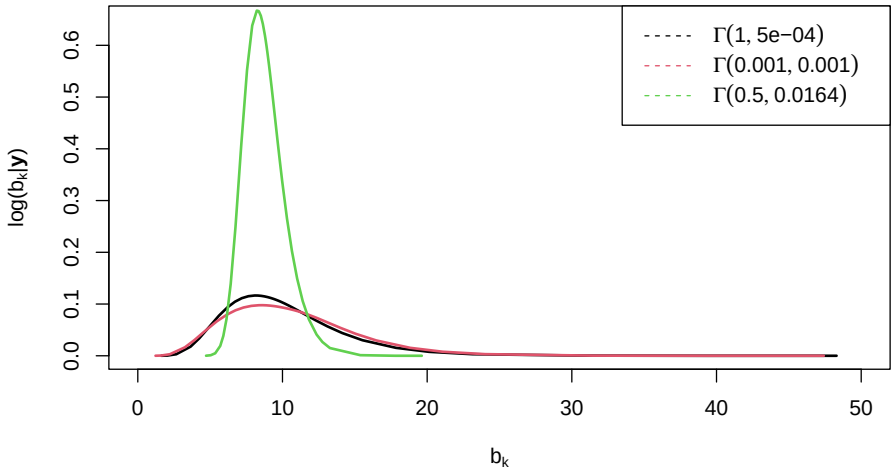


Figure 1: Sensitivity analysis on the priors of precision parameter of fitted negative binomial model. The black, red and green curves are for $\Gamma(1, 0.0005)$, $\Gamma(0.001, 0.001)$ and $\Gamma(0.5, 0.0164)$, respectively.

Table 3 displays the posterior means and the 95% credible interval estimates from the NB model assuming three sample sites random effects priors. Since the results from the fitted

Table 2: Summary of Poisson and negative binomial model DIC values for different priors.

	DIC	
Prior	Poisson	Neg.Bin
$\Gamma(1, 0.0005)$	-206894.20	41717.78
$\Gamma(0.001, 0.001)$	-206887.00	41706.62
$\Gamma(0.5, 0.0164)$	-206891.10	41716.74

models are similar, we only interpret the the results from the first model M1.

The estimated effects of the predictors of the tick abundance reveal no significant effects due to the type of habitat. That is, the model estimated non-significant and higher [0.33; CI:(-0.04,0.69) and 0.12; CI:(-1.03, 1.28)] log mean overall tick abundances in the grass and woods habitat types compared to the edges. However, the model estimated significant and higher [4.24; CI:(0.92,7.56)] log mean overall tick abundances in summer season, while non-significant [3.34; CI:(-0.14,7.96) and 2.74, (-1.14,6.50)] abundances were estimated during spring and fall seasons. The model estimated non-significant [-0.77; CI:(-2.99,1.36) and 0.02; CI:(-2.11)] cosine and sine sinusoidal terms. This indicates the model's inability to capture the seasonal patterns in the data. The model estimated a significant [5.09; CI:(2.61,12.65)] overall sample site random effect, which indicates differing overall tick abundances within sample sites with similar environmental and temporal conditions. The model estimated non-significant [-0.02; CI:(-0.41,0.33) and -0.04; CI:(-0.46,0.38))] correlations between larvae and nymph and nymph and adult ticks, while non-significant [0.03; CI:(-0.44,0.54)] correlation was estimated between larvae and adults.

Sample sites random effects

Figure 2 shows the caterpillar plot for the estimated sampling sites random effects posterior means (the dots) and 95% credible intervals. A negative posterior mean value for the

Table 3: Posterior estimates of the negative binomial models assuming **M1**: $\Gamma(0.1, 0.0005)$, **M2**: $\Gamma(0.001, 0.001)$ and **M3**: $\Gamma(0.5, 0.0164)$ priors in the precision matrix $\boldsymbol{\Omega}$.

Covariates	**M1**		**M2**		**M3**	
Fixed Effects	Mean	95% CI	Mean	95% CI	Mean	95% CI
(Intercept)	-4.87	(-7.67, -2.07)	-4.88	(-7.91, -1.86)	-4.90	(-8.14, -1.65)
Season (**ref**: Winter)						
Spring	3.91	(-0.40, 8.13)	3.89	(-0.68, 8.37)	3.84	(-1.17, 8.68)
Summer	4.24	(0.73, 7.75)	4.25	(0.51, 7.99)	4.25	(0.16, 8.34)
Fall	2.70	(-1.42, 6.66)	2.66	(-1.71, 6.90)	2.60	(-2.23, 7.19)
Habitat type (**ref**: Edge)						
Grass	0.33	(-0.03, 0.69)	0.33	(-0.03, 0.70)	0.33	(-0.03, 0.70)
Woods	0.12	(-1.03, 1.28)	0.12	(-1.03, 1.29)	0.14	(-1.03, 1.33)
Cosine term	-0.77	(-2.99, 1.36)	-0.84	(-3.35, 1.57)	-0.89	(-3.70, 1.71)
Sine term	0.02	(-2.12, 2.11)	-0.02	(-2.46, 2.36)	-0.05	(-2.75, 2.55)
Random Effects						
b_k	5.09	(2.61, 12.65)	4.82	(2.75, 8.77)	5.34	(2.04, 12.37)
Correlations						
ρ_{LN}	-0.02	(-0.41, 0.33)	0.00	(-0.24, 0.25)	0.08	(-0.35, 0.54)
ρ_{LA}	0.03	(-0.44, 0.54)	-0.01	(-0.26, 0.23)	0.04	(-0.39, 0.50)
ρ_{NA}	-0.04	(-0.46, 0.38)	0.03	(-0.38, 0.46)	0.02	(-0.41, 0.46)

sampling site random effects implies that lower variations are estimated, while a positive value implies higher variations are estimated within the sampling site. The model estimated significant higher tick variation in Stephens tract (ST), Jacobson tract (JC), Newport News Park (NN), Kiptopek (KP) and Hoffer Creek Wildlife Preserve (HC) and significant lower abundances in Paradise Creek Nature Park (PC) and Weyanoke (WS). Non-significant lower variations were estimated in Blackwater Ecological Preserve (BW), Oceana Dam/Neck (OD), Back Bay (BB) and Langley (LA), while non-significant higher abundances were estimated in Cheatham Annex (CA). The results from the model suggest that ST and JC sampled sites in Cheasapeake county, CA in York county, and KP Portsmouth county and KP in Northampton county had higher tick variations, while lower variations were estimated in PC sampling site in Portsmouth county and in WS sampling site in Norfolk county. On the other hand, non-significant lower variations were estimated in Virginia Beach, Isle of Wight, Hampton and York counties. The *A. americanum* is found throughout the southeastern

United States, and its populations extend west to central Texas and north to Iowa (Childs and Paddock, 2003).

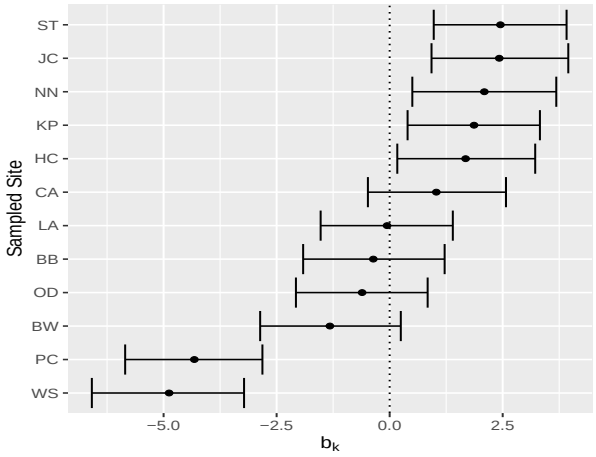


Figure 2: Caterpillar plots for predictors of the random effects of sampling locations with posterior means (dots) and 95% credible intervals in Virgnia, USA.

Non-linear effects

Figure 3 shows the shared trend effects of months and years over the study period, 2009 - 2018. The month trend plot shows that the log of mean tick abundances increase during the first six months (January to June), then decrease between June and July, increase between July and September, decrease between September and November, and increase from November to December. The year trend plot shows a steady decline in the log of mean tick abundances between 2009 through 2012, followed by a sharp decrease from 2012 through 2015, and then a steady decline from 2015 to 2017, after which a steady increase from 2017 through 2018 can be observed.

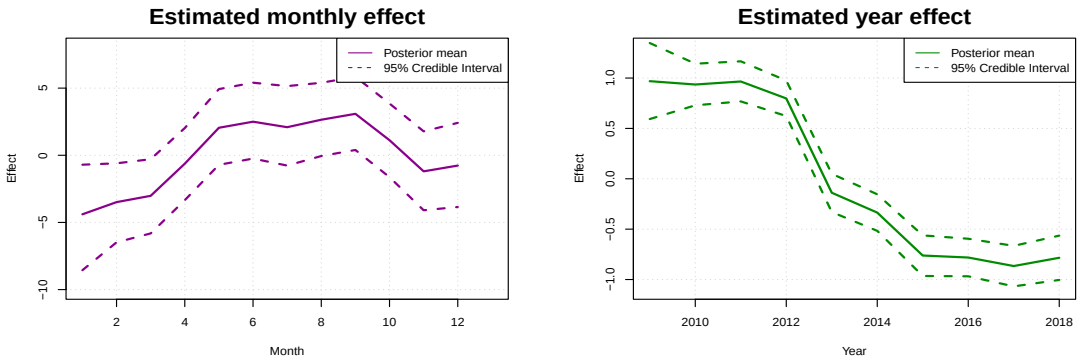


Figure 3: Estimated shared month and year time trends for larvae, nymph and adult tick life stages.

Discussion

This study used hierarchical dynamic Poisson and negative binomial count data models to investigate the influence of temporal and location specific effects on the distribution of *A. americanum* abundance in Virginia, United States. The model included the sampling sites random effects term to assess tick abundance variations within the sample sites. The proposed model was investigated through simulations at different locations and sample site sizes. It is worth noting that for small samples sizes the random effects model proposed may be susceptible to bias (Grilli et al., 2015). Furthermore, the model was fitted to the tick data and results showed that there were no significant effects of habitat types. We also noted that credible intervals were wider as a result of small $n = 3$ and $m = 12$ values. However, summer season effects were significant and positive indicating an expected increase in tick abundance during summer seasons. This aligns with the results in Childs and Paddock (2003); Kennedy et al. (2021); Lephoto et al. (2021); **?**, where the larvae life stage abundance peaks in August, the nymph stage in July, and the adult stage in June. The evidence of sample site abundance variations was detected in Chesapeake, York, Portsmouth, Northampton, and Norfolk coun-

ties suggesting potential disease hotspots. These results confirm the results in (Gaines et al., 2014; Cohen et al., 2010). The estimate of non-linear effects of the model showed an upward monthly tick abundance trend from January to June, a dip between June and July, steady peak between July and September, a sharp decrease between September and November, and a steady peak from November to December. However, the year tick abundance trend showed a steady decrease between 2009 and 2011, a steep decrease between 2011 and 2012, and a sharp decrease from 2012 through 2018. These trends are well documented in Davidson et al. (1994); Kollars Jr et al. (2000); Mangan et al. (2018). Similar non-linear effects of month and year were noted from our previous work using the same dataset (Lephoto et al., 2021).

Conclusion

In this study, we investigated the effects of dynamic, static and random sample site effects on the distribution of Lone Star tick abundance using the hierarchical dynamic Poisson and negative binomial count data models. The simulation study showed that the proposed model accurately captured the data for larger sample sizes. We conclude that application of the proposed model to data did not fit well, and wide credible intervals were obtained. Although the model showed insignificant relationships between tick abundance and the habitat type, the model was able to reveal that significant higher tick abundances were expected in summer compared to winter seasons. The model revealed that there were significant variations due to the random sample sites.

Declarations

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TL performed data analysis, interpreted the results and drafted the manuscript. HM and OB were responsible for steering the statistical part of the manuscript, while HG provided the data and assisted with data description. All authors discussed the results and implications and commented on the manuscript at all stages. All authors contributed extensively to the work presented in this paper. All authors read and approved the final manuscript.

Funding

This research was funded by NIH grant 1R01AI136035 as part of the joint NIH-NSF-USDA Ecology and Evolution of Infectious Diseases program. The funding body had no role in the desgin of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Bibliography

Adams, D. A., Gallagher, K. M., Jajosky, R. A., Kriseman, J., Sharp, P., Anderson, W. J., Aranas, A. E., Mayes, M., Wodajo, M. S., Onweh, D. H., et al. (2013). Summary of notifiable diseases–united states, 2011. *Morbodity and Mortality Weekly Report*, 60(53):1–17.

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653.

Alkishe, A. and Peterson, A. T. (2022). Climate change influences on the geographic distributional potential of the spotted fever vectors amblyomma maculatum and dermacentor andersoni. *PeerJ*, 10:e13279.

Benham, S. A., Gaff, H. D., Bement, Z. J., Blaise, C., Cummins, H. K., Ferrara, R., Moreno, J., Parker, E., Phan, A., Rose, T., et al. (2021). Comparative population genetics of Amblyomma maculatum and Amblyomma americanum in the mid-Atlantic United States. *Ticks and Tick-borne Diseases*, 12(1):101600.

CDC (2020). Guide to the surveillance of metastriate ticks (Acari: Ixodidae) and their pathogens in the United States.

CDC (2021). National notifiable diseases surveillance system, 2019 annual tables of infectious disease data.

Chib, S. and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435.

Childs, J. E. and Paddock, C. D. (2003). The ascendancy of Amblyomma americanum as a vector of pathogens affecting humans in the United States. *Annual Review of Entomology*, 48(1):307–337.

Cohen, S. B., Yabsley, M. J., Freye, J. D., Dunlap, B. G., Rowland, M. E., Huang, J., Dunn, J. R., Jones, T. F., and Moncayo, A. C. (2010). Prevalence of Ehrlichia chaffeensis and Ehrlichia ewingii in ticks from Tennessee. *Vector-Borne and Zoonotic Diseases*, 10(5):435–440.

Davidson, W. R., Siefken, D. A., and Creekmore, L. H. (1994). Seasonal and annual abundance of Amblyomma americanum (Acari: Ixodidae) in central Georgia. *Journal of Medical Entomology*, 31(1):67–71.

Fokianos, K. (2015). *Handbook of discrete-valued time series, handbooks of modern statistical methods*. Chapman & Hall.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Gaines, D. N., Operario, D. J., Stroup, S., Stromdahl, E., Wright, C., Gaff, H., Broyhill, J., Smith, J., Norris, D. E., Henning, T., et al. (2014). Ehrlichia and spotted fever group Rickettsiae surveillance in Amblyomma americanum in Virginia through use of a novel six-plex real-time PCR assay. *Vector-Borne and Zoonotic Diseases*, 14(5):307–316.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Goddard, J. and Varela-Stokes, A. S. (2009). Role of the Lone Star tick, Amblyomma americanum (L.), in human and animal diseases. *Veterinary Parasitology*, 160(1-2):1–12.

Grilli, L., Metelli, S., and Rampichini, C. (2015). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 85(13):2718–2726.

Hair, J. A. and Howell, D. E. (1970). *Lone Star ticks: their biology and control in Ozark recreation areas*. Oklahoma Agricultural Experiment Station, Volume 679 of Bulletin B.

Kennedy, A. C., Marshall, E., et al. (2021). Lone Star ticks (amblyomma americanum): an emerging threat in Delaware. *Delaware Journal of Public Health*, 7(1):66.

Kollars Jr, T. M., Oliver Jr, J. H., Durden, L. A., and Kollars, P. G. (2000). Host associations and seasonal activity of Amblyomma americanum (Acari: Ixodidae) in Missouri. *Journal of Parasitology*, 86(5):1156–1159.

Laaksonen, M., Sajanti, E., Sormunen, J. J., Penttinen, R., Hänninen, J., Ruohomäki, K., Sääksjärvi, I., Vesterinen, E. J., Vuorinen, I., Hytönen, J., et al. (2017). Crowdsourcing-based nationwide tick collection reveals the distribution of Ixodes ricinus and I. persulcatus and associated pathogens in Finland. *Emerging Microbes & Infections*, 6(1):1–7.

Lephoto, T., Mwambi, H., Bodhlyera, O., and Gaff, H. (2021). Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients. *Geospatial Health*, 16(2).

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press.

Ma, J., Kockelman, K. M., and Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, 40(3):964–975.

Mangan, M. J., Foré, S. A., and Kim, H.-J. (2018). Ecological modeling over seven years to describe the number of host-seeking Amblyomma americanum in each life stage in northeast Missouri. *Journal of Vector Ecology*, 43(2):271–284.

Merten, H. A., Durden, L. A., et al. (2000). A state-by-state survey of ticks recorded from humans in the United States. *Journal of Vector Ecology*, 25(1):102–113.

Molaei, G., Little, E. A., Williams, S. C., and Stafford, K. C. (2019). Bracing for the worst-range expansion of the Lone Star tick in the northeastern United States. *New England Journal of Medicine*, 381(23):2189–2192.

Nadolny, R. M. and Gaff, H. D. (2018). Natural history of Amblyomma maculatum in Virginia. *Ticks and Tick-borne Diseases*, 9(2):188–195.

Nadolny, R. M., Wright, C. L., Sonenshine, D. E., Hynes, W. L., and Gaff, H. D. (2014). Ticks and spotted fever group rickettsiae of southeastern Virginia. *Ticks and Tick-borne Diseases*, 5(1):53–57.

Oakes, V. J., Yabsley, M. J., Schwartz, D., LeRoith, T., Bissett, C., Broaddus, C., Schlater, J. L., Todd, S. M., Boes, K. M., Brookhart, M., et al. (2019). Theileria orientalis Ikeda genotype in cattle, Virginia, USA. *Emerging Infectious Diseases*, 25(9):1653.

Otálora-Luna, F., Dickens, J. C., Brinkerhoff, J., and Li, A. Y. (2022). Geotropic, hydrokinetic and random walking differ between sympatric tick species: the deer tick Ixodes scapularis and the Lone Star tick Ambylomma americanum. *Journal of Ethology*, 40(2):133–143.

Pasternak, A. R. and Palli, S. R. (2023). County-level surveillance for the lone star tick, Amblyomma americanum, and its associated pathogen, Ehrlichia chaffeensis, in Kentucky. *Ticks and Tick-borne Diseases*, 14(1):102072.

Paull, S. H., Thibault, K. M., and Benson, A. L. (2022). Tick abundance, diversity and pathogen data collected by the national ecological observatory network. *Gigabyte*, 2022:1–11.

Ravishanker, N., Raman, B., and Soyer, R. (2022). *Dynamic time series models using R-INLA: an applied perspective*. CRC Press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Savage, H. M., Burkhalter, K. L., Godsey Jr, M. S., Panella, N. A., Ashley, D. C., Nicholson, W. L., and Lambert, A. J. (2017). Bourbon virus in field-collected ticks, Missouri, USA. *Emerging Infectious Diseases*, 23(12).

Savage, H. M., Godsey Jr, M. S., Lambert, A., Panella, N. A., Burkhalter, K. L., Harmon, J. R., Lash, R. R., Ashley, D. C., and Nicholson, W. L. (2013). First detection of heartland virus (Bunyaviridae: Phlebovirus) from field collected arthropods. *The American Journal of Tropical Medicine and Hygiene*, 89(3):445.

Sonenshine, D. E. (1979). Insects of Virginia no. 13. *Ticks of Virginia (Acari: Metastigmata). Research Division Bulletin*, 139.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Wimms, C., Aljundi, E., and Halsey, S. J. (2023). Regional dynamics of tick vectors of human disease. *Current Opinion in Insect Science*, 55:101006.

pagepress

Geospatial Health 2021; volume 16:1004

Spatio-temporal modelling of tick life-stage count data with spatially varying coefficients

Thabo Lephoto,[1] Henry Mwambi,[1] Oliver Bodhlyera,[1] Holly Gaff[2]

[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, KwaZulu-Natal Province, South Africa; [2]Department of Biological Sciences, Old Dominion University, Norfolk, VA, USA

Abstract

There is a vast amount of geo-referenced data in many fields of study including ecological studies. Geo-referencing is usually by point referencing; that is, latitudes and longitudes or by areal referencing, which includes districts, counties, states, provinces and other administrative units. The availability of large geo-referenced datasets for modelling has necessitated the development and application of spatial statistical methods. However, spatial varying coefficients models exploring the abundance of tick counts remain limited. In this study we used data that was collected and prepared by researchers in the Department of Biological Sciences from the Old Dominion University, Virginia, USA. We modelled tick life-stage counts and abundance variability from 12 sampling locations, with 5 different habitats (numbered 1-5), three habitat types; namely: woods, edges and grass; collected monthly from May 2009 through December 2018. Spatio-temporal Poisson and spatio-temporal negative binomial (NB) count data models were fitted to the data and compared using the deviance information criteria (DIC). The NB model outperformed the Poisson models with all its DIC values being smaller than those of the Poisson model. Results showed that the covariates varied spatially across counties. There was a decreasing time (in years) effect over the study period. However, even though the time effect was decreasing over the study period, space-time interaction effects were seen to be increasing over time in York County.

Correspondence: Thabo Lephoto, University of KwaZulu-Natal, KwaZulu-Natal Province, Private Bag X01, Pietermaritzburg 3201, South Africa.
Tel.: +27.33.260.5785 /+27.74.704.8528.
E-mail: LephotoT@ukzn.ac.za

Key words: Bayesian modelling; tick-counts; spatial and spatio-temporal effects.

Funding: this research was funded by NIH grant 1R01AI136035 as part of the joint NIH-NSF-USDA Ecology and Evolution of Infectious Diseases program. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Received for publication: 6 April 2021.
Revision received: 20 May 2021.
Accepted for publication: 20 May 2021.

Geospatial Health 2021; 16:1004
doi:10.4081/gh.2021.1004

Introduction

Infectious agents that are transmitted by tick bites cause tick-borne diseases, such as Rocky Mountain spotted fever, Lyme disease, Ehrlichiosis, tularaemia, babesiosis, Colorado tick fever, and relapsing fever (Tälleklint and Jaenson, 1998; Bowman *et al.*, 2004; Bratton and Corey, 2005). Ticks are a highly specialized group of obligate, bloodsucking and non-permanent ectoparasitic arthropods that feed on mammals, birds and reptiles in most regions of the Earth. They are characterized by having relatively large body sizes among the acari; ingesting enormous quantities of vertebrate blood, digested tissues, or lymph; laying 200 to 23,000 eggs; and having moulting and reproduction regulated by blood ingestion (Anderson, 2002).

In Europe, ticks have spread and become established in areas that were previously not considered to be favourable for them. The reason is not only directly due to climate change, but also due to continuous changes in vegetation, landscape features and human social habits which are leading to new areas of contact between ticks, their pathogens and the interface between animals and humans (Salman, 2012). Tälleklint and Jaenson (1998) attribute these changes to climate change, which has led to the *Ixodes ricinus* species being detected in northern Sweden where they have now colonized relatively high altitude ranges in the mountains. In African countries, Asia, the Near East and parts of Europe, serious outbreaks of Crimean-Congo haemorrhagic fever have been reported. This fever is said to mainly be the result of tick-bites by *Hyalomma marginatum*, which is the most common tick species in Mediterranean-type environments and in the African steppes, but neither found in America nor Australia with absence of reported cases there to date (Salman, 2012).

Amblyomma americanum (lone star tick) is a major human-biting tick in eastern, southern and mid-western U.S. (Goddard and Varela-Stokes, 2009). As in much of the world, tick-borne diseases are an emerging public health threat in this part of the U.S. (Sayler *et al.*, 2017). According to Stromdahl and Hickling (2012), people who live and usually visit state parks and national forests in North America are at a higher risk of exposure to tick bites and their associated pathogens than those whose activities keep them outside of these areas. In this part of the U.S., there are only five

tick species, *i.e. A. americanum* L., *A. maculatum* Koch, *Dermacentor variabilis* Say, *Ixodes scapularis* Say, and *Rhipicephalus sanguineus*, that commonly bite humans (Stromdahl and Hickling, 2012; Nathavitharana and Mitty, 2015). According to Childs and Paddock (2003), *A. americanum* is the primary vector for *Ehrlichia chafleensis* and *E. ewingii*, and it is associated with southern tick-associated rash illness and a number of other diseases. Despite its prominence as a nuisance biter and vector of human pathogens in the U.S., there is a lack of efforts to define the species' geographical range. This deprive us of information not only about identification of human populations at risk for *A. americanum* and its associated pathogens, but also in determining the landscape, host assemblage and environmental conditions that are favourable for the establishment and rapid reproduction of this tick species. It has emerged as one of the most important tick vectors in the U.S. transmitting pathogens to both humans and domestics animals (Childs and Paddock, 2003; Mixson *et al.*, 2006; Goddard and Varela-Stokes, 2009; Fritzen *et al.*, 2011). Hendricks *et al.* (2017) state that domestic dogs and cats are potentially effective sentinel populations for monitoring occurrence and spread of Lyme disease.

The focus of this study was to develop statistical models to model tick count data that include areal referencing. We wished to know whether the lone star tick life stage counts vary over time, across, and between, counties in south-eastern Virginia in the U.S. The lone star tick abundance is largely influenced by the availability of suitable animal hosts for the life stages and by the availability of habitats with physiographic features that offer protection for hosts and guard against desiccation of the tick (Childs and Paddock, 2003). Spatial and spatiotemporal statistics can reveal important environmental and temporal characteristics, for example, by incorporating time, space, and space time covariate interactions into the model. The main objective of this research was to use spatiotemporal analysis to explore environmental and temporal relationships with tick life-stage count data using data collected from 2009 through 2018. The knowledge and information unpacked are as essential in tick-related disease surveillance, as it is for effective planning and decision-making. This study could also be of interest with respect to the use of the INLA package in the R statistical software as compared to the Markov Chain Monte Carlo (MCMC) in the WinBUGS statistical software package.

Materials and methods

Data source and study area

We used data collected and prepared by researchers in the Department of Biological Sciences from Old Dominion University, Virginia. U.S. Ticks were collected using standard flagging techniques (CDC, 2020) along established transects and identified species and life-stage according to (Sonenshine, 1979). Larvae-, nymph- and adult-stage count data were recorded for each different location, habitat and habitat type throughout the four seasons of the year. Eight counties and independent cities were sampled in this state, namely; City of Chesapeake, City of Hampton, Isle of Wight County, City of Norfolk, Northampton County, City of Portsmouth, City of Virginia Beach and York County (Figure 1). For brevity, all will be referenced as counties. Data collection was done at least once a month on varying days of the week and at 12 different sites in southeast Virginia from May 2009 through December 2018 (H. Gaff unpublished data). The data were collected at random from multiple areas referred to as habitats, and each area was designated by a unique number ranging from 1 to 5 for different habitats. The habitat type was used to designate the kind of area (woods, edges or grass) where the data were collected. The number of the week (from 1 to 53) was also recorded during data collection. Week 1 is the first week of the year, and while fewer ticks were observed in winter, the adult stage of *I. scapularis* is active in winter. To help align the information from year to year, we recorded data collected during the last week of December as week 53, which was also the first week of January of the following year. We also kept track of the month and year for the data collection. The ten-year study period was grouped into two-year successive periods such that 2009 and 2010 were grouped together, 2011 and 2012, 2013 and 2014, 2015 and 2016 and 2017 and 2018. This predictor variable was used to capture residual spatial effects on the abundance of tick counts. The paired time segmentation was to help find out if there were times when counts were significantly high or low compared with the rest of the data. The above-mentioned variables were used as predictors of tick life-stage count outcome data to find relationships using count regression models.

Model descriptions

Various statistical models have been developed to model count data. In this study, we applied the Poisson and the negative binomial distribution models. The classical Poisson model, which assumes that mean and variance of the count responses are equal, is often of limited use when the empirical data sets exhibit over-dispersion or have more zeros than expected. This can be addressed by introducing a dispersion parameter in this model or by extension to models that can account for excessive zeros in the data (Zeileis *et al.*, 2008). As mean and variance are identical in the standard Poisson model, this means that the dispersion parameter is fixed to 1. In the presence of over-dispersion, the Poisson model underestimates the variance and render all model-based tests more conservative. Violation of the equal mean and variance assumption indicates correlation in the data, which may affect both standard errors of the parameter estimates and the model. In this study, the tick count data showed greater proportion of zeros than that of positive counts, which is an indication of zero-inflation and over-dispersion.

The negative binomial (NB) distribution is commonly used to model over-dispersed count data. A dispersion parameter in the NB model caters for over-dispersion allowing the variance to be greater than the mean while also accommodating the unobserved heterogeneity in the data. In addition to over-dispersion, it is common that many empirical count data sets exhibit more zero observations than would be expected by the classical Poisson model. A model capable of capturing both properties is the zero-inflated Poisson (ZIP), which assumes that zero counts occur with some probability, while a Possion (λ) random variable is observed with probability $1 - p$. The ZIP distribution model approaches the classical Poisson distribution when $p \rightarrow 0$. It is worth noting that zero observations arise from both the zero-component distribution and the classical Poisson distribution. The zero-component distribution accounts for the inflated zeros that are observed in addition to zeros that are expected to be observed under the classical Poisson distribution. A more detailed account of the development of zero-inflated models can be found in (Lambert, 1992).

Let y_{ijkm} be a tick count for life-stage m (m=1: larvae; m=2: nymphs; and m=3: adults) in habitat-type k, habitat j and location i modelled by the multi-hierarchical Poisson model. Under the Poisson model, we assume that the dependent variable y_{ijkm} is

Poisson-distributed, *i.e.*:

$$y_{ijkm} \mid \mu_{ijkm} \sim \text{Poisson}(\mu_{ijkm})$$

where μ_{ijkm} is the mean tick count in the respective life-stage, habitat type, habitat and location.

Let $X_{ijkm} = (\alpha_m; x_{ijkm1}, x_{ijkm2},\ldots, x_{ijkmp})'$ be a vector of p continuous predictors with the first component accounting for the constant and $W_{ijkm} = (w_{ijkm1}, w_{ijkm2},\ldots, w_{ijkmr})'$ be a vector of r categorical predictors. The link function $h(\cdot)$ relates the mean μ_{ijkm} to the predictors as follows:

$$h(\mu_{ijkm}) = X'\beta_m + W'\gamma_m \quad (1)$$

where $W = W_{ijkm}$ and $X = X_{ijkm}$; while $h(\cdot)$ is the log link function, β_m an m-dimensional vector of regression coefficients for continuous predictors; and γ_m a vector of the categorical predictors.

In order to cater for non-linear effects of the continuous covariates, the spatial autocorrelation, and temporal effects in the data, we incorporated the random walk model of order 2 (RW2); the convolution models; the linear trend over years; and the space-time interactions into the model such that (Eq. 1) becomes:

$$h(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f_i(S_i, x_{ijkm}) + f_{month}(m) + f_{spat}(S_{im}) + f_{year}(t) + f_{it}(S_i, t) \quad (2)$$

where the function $f_i(S_i, x_{ijkm})$ represents the space-covariate interaction; $f_{month}(m)$ a non-linear twice differentiable smooth function for the continuous month covariate effect; and the functions $f_{spat}(S_{im})$, $f_{year}(t)$ and $f_{it}(S_i, t)$ functions for space, time (years) and space-time interaction, respectively. The convolution model assumes that the spatial effect can be decomposed into two; namely, the spatially structured and spatially unstructured components. This means that $f_{spat}(S_{im}) = f_{str}(S_{im}) + f_{unstr}(S_{im})$ where $m = 1$, 2 or 3 (Manda and Leyland, 2007; Ngesa *et al.*, 2014; Okango *et al.*, 2015).

The spatially structured random effects account for the unobserved covariates, which vary spatially across counties, while the spatially unstructured random effects account for the unobserved covariates that are inherent within counties or correlations within counties, for example, the climate and common cultural practices among other things. The spatially structured random effects are the spatial autocorrelations and they are technically defined as the correlation computed among the values of a single geo-referenced variable that is attributable to the geographic proximity of the objects, to which the values are attached (Cliff and Ord, 1981). Moreover, $f_{year}(t)$ represents random time effects which can be modeled as a first-order random walk (RW1) or a first-order autoregressive process (AR1), while $f_{it}(S_i, t)$is a space-time interaction.

In the presence of over-dispersion, the Poisson model is replaced by the negative binomial model, where the variance depends on the mean as $\mu(1 + \mu\Psi)$, and where Ψ is an over-dispersion parameter, which measures the extent at which variance deviates from the mean. We assumed that a tick count variable y_{ijkm} follows a negative binomial distribution, such that

$$y_{ijkm} \mid \mu_{ijkm} \sim \text{NB}(\mu, \Psi)$$

The mean function $E(y_{ijkm}) = \mu_{ijkm}$ relates to the predictors in the same way as that of the Poisson distribution model in Eq. 1. In this study, a full Bayesian estimation approach was used, where parameters were assigned prior distributions.

Model diagnostics

The models were compared using the deviance information criterion (DIC), which is obtained by adding the posterior mean of the deviance that measures the goodness of fit to the number of effective parameters as: $DIC = \overline{D}(\theta) + p_D$ where $\overline{D}$ is the posterior mean deviance and p_D is the effective number of parameters in the model, which penalizes the fit for complexity of the model. Spiegelhalter *et al.* (2002) state that p_D values less than zero indicate substantial conflict between the prior and the data or that the posterior mean is a poor estimator. The best model is said to be the one with the smallest DIC value. Low values of $\overline{D}$ suggest a better fit, while small values of p_D suggest model parsimony as discussed in Spiegelhalter *et al.* (2002).

Fitted models

Four spatio-temporal models were fitted in the R statistical software, version 3.5.1 using the '*INLA*' package to predict the effects of the ecological covariates on the distribution of larvae, nymph and adult tick counts in the eight previously mentioned counties of Virginia, U.S. Preparation, organization and merging of the data with the map was done using the QGIS software, version 3.6.3-Noosa (QGIS, 2009). The first approach modelled the space-covariate effects $f(S_i, x_{ijkm})$, non-linear effects of the month covariate, $f_{month}(m)$, the temporal time effects $f_{year}(t)$ and the space-time effects $f_{it}(S_i, t)$. This model (M1) does not consider the spatially structured and spatially unstructured random effects and the three life-stage counts were modelled independently as follows:

$$\text{M1: } \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f(S_i, x_{ijkm}) + f_{month}(m) + f_{year}(t) + f_{it}(S_i, t) \quad (3)$$

where m equals 1 for larvae; 2 for nymphs; and 3 for adults with respect to counts. The second model (M2) is the same as M1 but with spatially unstructured effects to cater for the unobserved covariates that are inherent within the counties. The spatially unstructured effects were specified by the identically and independent distributed (iid) with the normal distribution.

$$\text{M2: } \log(\mu_{ijkm}) = \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f(S_i, x_{ijkm}) + f(month) + f_{unstr}(S_i) + f_{year}(t) + f_{it}(S_i, t) \quad (4)$$

The third model (M3) is the same as M1 but with spatially structured effects which cater for any unobserved covariates which vary spatially across counties, specified by the conditional autoregressive *model* (CAR).

$$\text{M3: } \log(\mu_{ijkm}) = \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f(S_i, x_{ijkm}) + f(month) + f_{str}(S_i) + f_{year}(t) + f_{it}(S_i, t) \quad (5)$$

The fourth model (M4) is the same as M1 with a convolution of spatially unstructured and spatially structured effects, which are specified by the iid normal distribution and CAR model respectively.

$$\text{M4: } \log(\mu_{ijkm}) = \log(\mu_{ijkm}) = \beta_{0m} + \sum_{i=1}^{p} f(S_i, x_{ijkm}) + f(month) + f_{unstr}(S_i) + f_{str}(S_i) + f_{year}(t) + f_{it}(S_i, t) \quad (6)$$

We used the Poisson and the negative binomial count data distributions and compared them using the DIC. We then interpreted the results from the best performing models based on the DIC.

Assessment

Before presenting the study results, we first performed an exploratory data analysis and also compared all the models using the DIC. We compared spatio-temporal Poisson models with their corresponding spatio-temporal negative binomial models.

Exploratory data analysis

Table 1 shows environmental and time predictor variables used in this study. Table 2 shows the number of positive and zero counts in the data. It is obvious that the data have more zero counts than positive counts for all tick life stages.

Our records show that the larvae were more abundant in the environment than the nymph and adult stage ticks at all times. As seen in Figure 2, the blue bars for larvae are on average, approximately 7 and 14 times taller when compared with the orange and grey bars for the nymph and adults, respectively, during the study period from 2009 through 2018.

Model comparisons

Table 3 shows the DICs for four spatio-temporal Poisson and NB models, where the model with the smallest DIC is the one with the best fit. As seen, all the spatio-temporal NB models have lower DICs compared to the corresponding spatio-temporal Poisson models. M1 showed the best fit for the tick larvae count data compared to the all the other models, which also suggests that unobserved covariates vary significantly over time. Similarly, M4 showed the best fit for the nymph and adult tick count data. The differences in the DIC values suggest that the spatio-temporal NB model would be the best model compared to the spatio-temporal Poisson model. This outcome also suggests that unobserved covariates vary spatially across counties and over time.

Results

Space-covariate effects

Spatio-temporal NB models out-performed the spatio-temporal Poisson models in our case. We show the choropleth maps (Figures 2-4) of all the models with the smallest DICs for the fitted larvae, nymph and adult tick count data, respectively. The choropleth maps show the space-covariate interaction effects for the selected counties in Virginia. A yellow shade was used if the effects are greater and black or dark shade if the effects were lower.

Larvae

The effect of location on the log mean tick counts was highest in Hampton County followed by York and Chesapeake counties (Figure 3). The effect was lower in Norfolk, Portsmouth and Isle of Wight counties. Virginia Beach County showed the lowest effect of the location variable. The effect of habitat was almost the same across all the counties except for Isle of Wight County. The effect of change in weeks was evident in York and Norfolk counties. The effect of habitat type on the log of mean larvae tick counts was high in Chesapeake County. The effect of change in season was very low in Isle of Wight County compared to other counties.

Table 1. Predictor variables and their types.

Predictor	Type
Habitat	Environmental
Habitat type	Environmental
Location	Environmental
Year	Time
Month	Time
Season	Time

Table 2. Tick counts in the data, 2009-2018.

Variable	Positive counts	Nil counts
Larvae	726	3767
Nymphs	1770	2723
Adults	1365	3128

Table 3. Deviance information criteria values for spatio-temporal Poisson and negative binomial models.

Response	Total counts	Model	Spatio-temporal models			
			Poisson		NB	
			DIC	p_D	DIC	p_D
Larvae	145020		281147.18	71.79	12563.79	13.31
Nymphs	20637	1	29982.53	79.52	-	-
Adult	10509		13990.35	62.55	-	-
Larvae	145020		277483.97	73.81	12591.82	9.64
Nymphs	20637	2	30176.78	67.67	14203.41	49.58
Adult	10509		14308.39	62.55	10101.87	47.52
Larvae	145020		295800.95	71.16	12600.72	10.82
Nymphs	20637	3	34093.74	65.24	14518.67	45.97
Adult	10509		14317.57	62.60	10101.87	47.52
Larvae	145020		277468.27	73.88	12591.73	9.64
Nymphs	20637	4	30176.53	67.64	14202.29	49.36
Adult	10509		14308.23	62.55	10100.97	47.13

DIC, deviance information criteria; NB, negative binomial.

Nymphs

The effect of location on the log mean nymph counts was high in Hampton County and very low in Norfolk County compared to the others (Figure 4). The effect of habitat was very low in Isle of Wight County, and low in Northampton County followed by Chesapeake, Hampton and Portsmouth counties. However, there was high effect of habitat in Norfolk County. The effect of change in week was very low in York County and Northampton County but high in Portsmouth County. The effect of habitat type on the log nymphal counts was evident in Chesapeake County but low in Isle of Wight County. The effect of change in seasons was very low in Isle of Wight County, while a higher effect can be observed in York, Hampton and Norfolk counties (Figure 4).

Adults

The effect of location on the log mean adult counts was high in York County, while it was generally low in other counties. The effect of habitat was high in Norfolk and Hampton counties, but it was a very low in Isle of Wight County. The effect of change in week was also evident in the latter. A very low effect in this respect can be observed in York County (Figure 5). The effect of habitat type was very high in Chesapeake County and very low in York County compared with other counties. The effect of change in season was very high in Norfolk County and very low in Isle of Wight County. Other counties had low effects as compared to the effect in the latter.

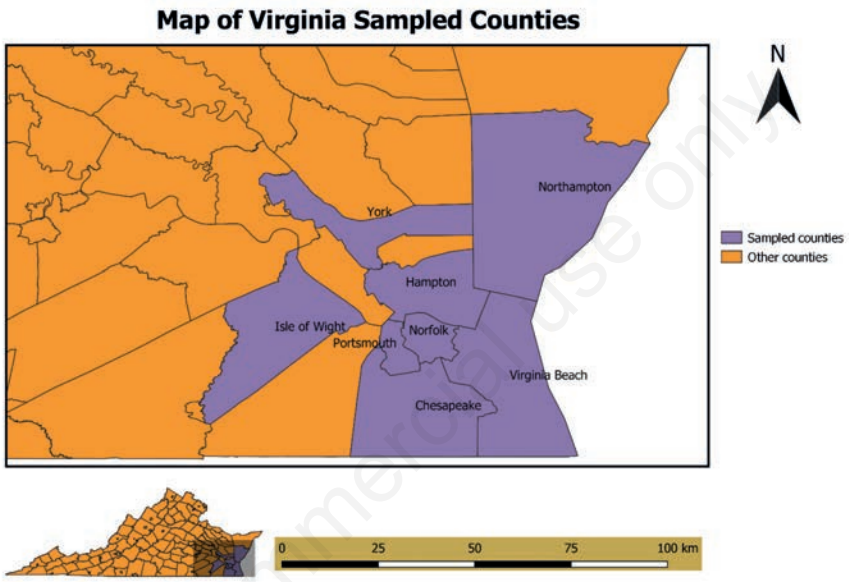
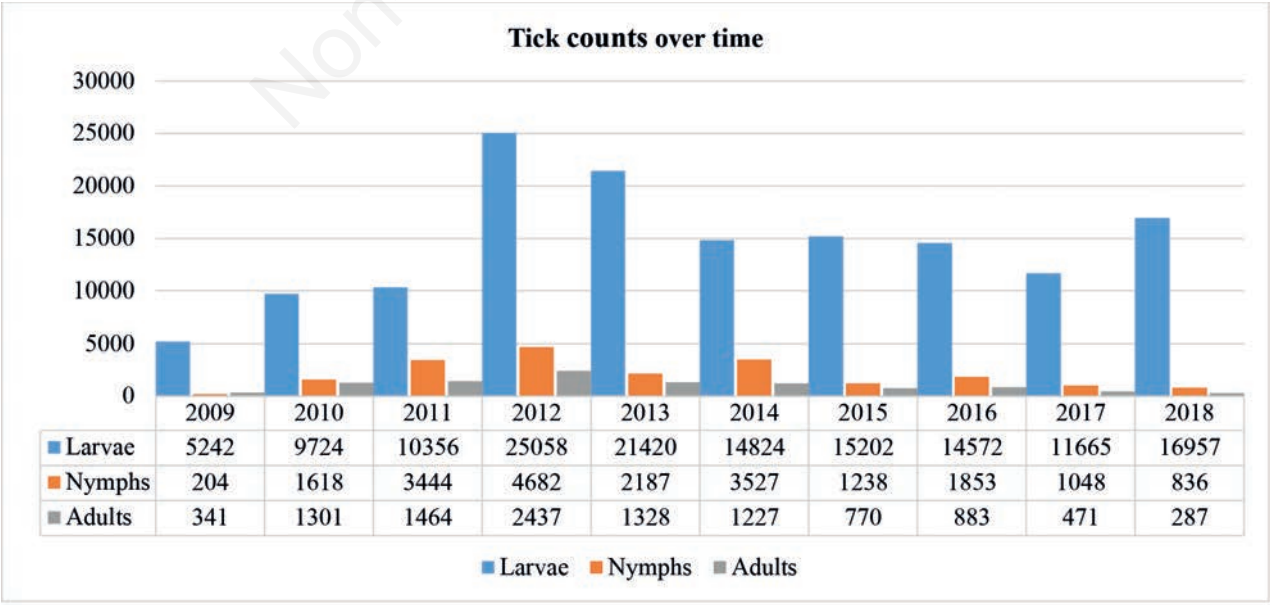


Figure 1. Map of Virginia with names of counties where ticks were sampled.



	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Larvae	5242	9724	10356	25058	21420	14824	15202	14572	11665	16957
Nymphs	204	1618	3444	4682	2187	3527	1238	1853	1048	836
Adults	341	1301	1464	2437	1328	1227	770	883	471	287

Figure 2. Tick counts of larvae, nymph and adult species in Virginia, USA 2009-2018.

The non-linear effects of the month

Figure 6 shows non-linear associations between time (month) and larvae, nymph and adult tick counts. These figures give the posterior mean (black line in the figure) of the smooth function and their corresponding 95% credible intervals (red and blue lines). From the figures it is evident that there is a non-linear relationship between month and tick counts.

We could confirm the general changes along the seasons in the study area. Thus we report the probability of observing larvae increases between January and April, fluctuates between April and June and then starts to increase even further between June and August. After those chances of observing larvae decrease until December. The chances of observing nymphs increase slightly between January and February, the number of larvae increases abruptly between February and June, then decreases again between June and August followed by a steady decrease between August and September. From there, it decreases sharply between September and December. The chances of observing adults increases steadily between January and February, then increases

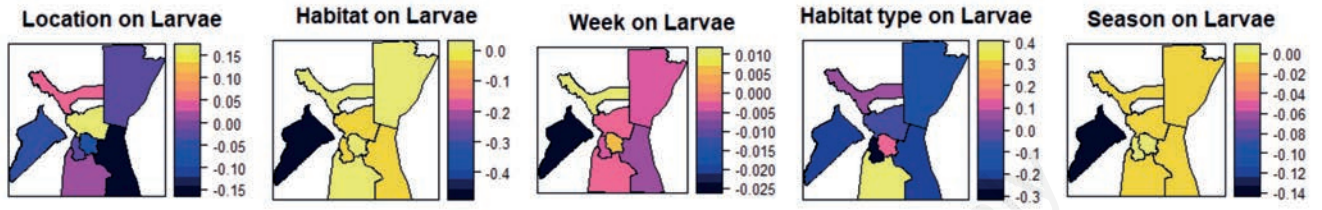


Figure 3. Choropleth maps showing the space-covariate effects on larvae.

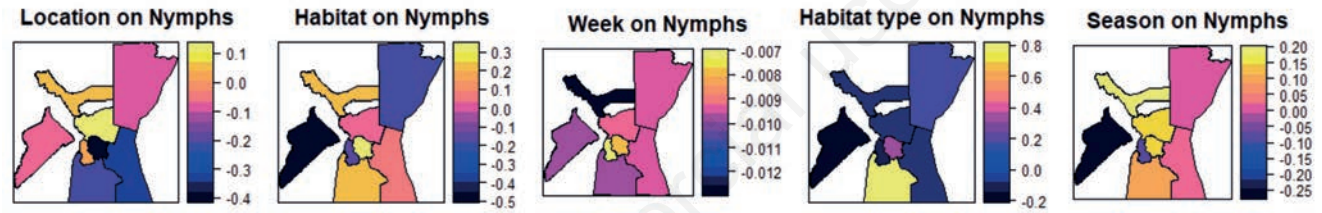


Figure 4. Choropleth maps showing the space-covariate effects on nymphs.

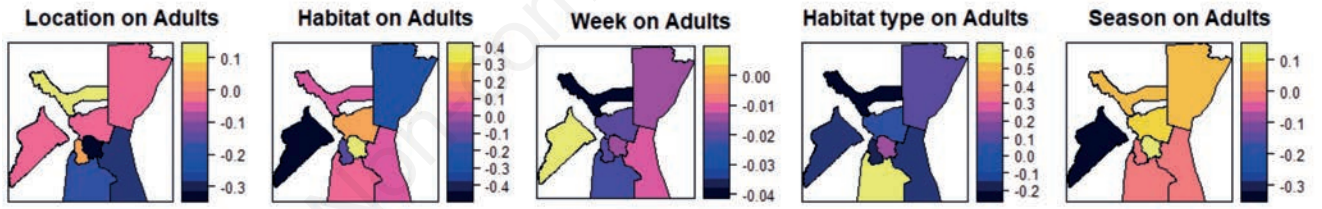


Figure 5. Choropleth maps showing the space-covariate effects on adults.

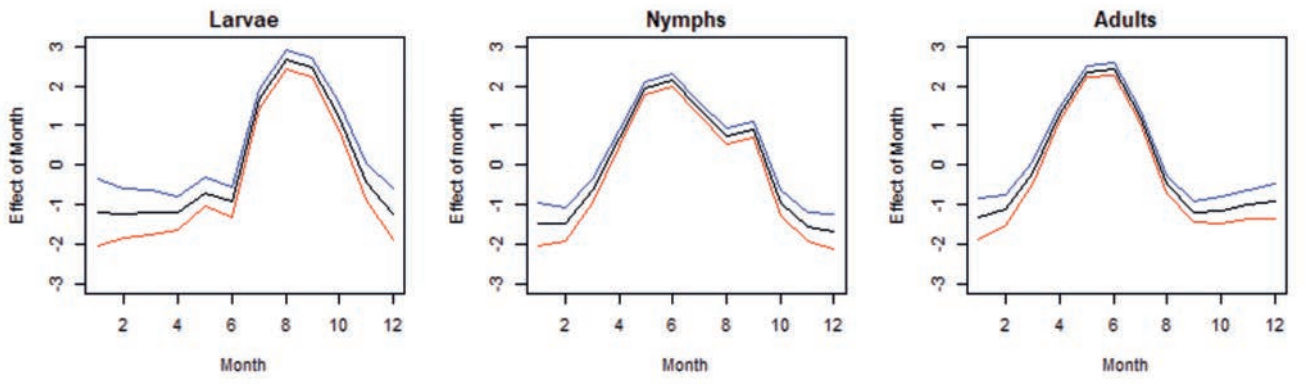


Figure 6. Non-linear effects of month variable on larvae, nymphs and adults 2009-2018.

abruptly from February to May, steadily between May and June and decreases sharply between June to September. The probability then starts to increase steadily from September to December.

Space-time effects

Larvae

Figure 7 shows the mapped estimated residual spatial effects on the abundance of larvae counts between 2009 and 2018. The residual spatial effects that represent unobserved spatial factors either not measured in the surveys or abducting the effects of cultural patterns are evident. High effects were observed in Portsmouth, Chesapeake and Hampton counties in 2009/2010 while the other counties showed low residual spatial effects. Three counties showed decreased residual spatial effects in the period 2011/2012 and these effects were evident in Norfolk and Northampton counties. In the period 2013/2014 the effects were evident in Hampton, Portsmouth, Chesapeake and York counties. The effects were high in Chesapeake, Norfolk and York counties 2015/2016, while they were very low in Portsmouth County. During 2017/2018 all counties, except for Isle of Wight and Portsmouth counties, showed high effects on the abundance of larvae.

Nymphs

The effects on the nymph abundance decreased over the whole study period, particularly in York County (Figure 8).

Adults

Residual spatial effects on the distribution of abundance of adult ticks increased between 2009 and 2016 in York County, while all other counties showed a decrease of spatial residual effects throughout the study period 2019 to 2018. In York, a decrease in residual spatial effects was only observed after 2015/2016 study period. Higher residual spatial effects were evident in Norfolk County during the periods 2009/2010, 2011/2012 and 2013/2014, respectively (Figure 9).

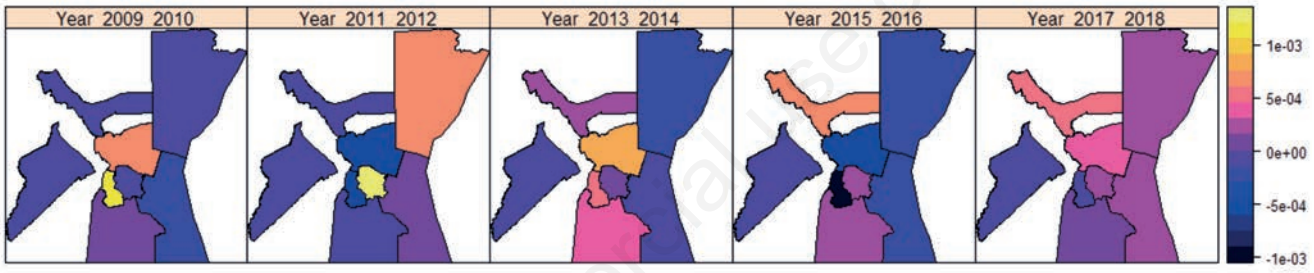


Figure 7. Choropleth maps showing residual spatial effect of larvae tick abundance 2009-2018.

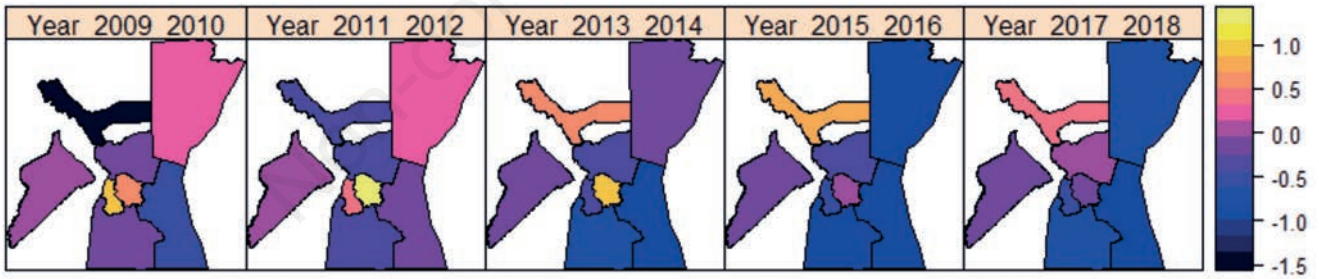


Figure 8. Choropleth maps showing residual spatial effect of nymph tick abundance 2009-2018.

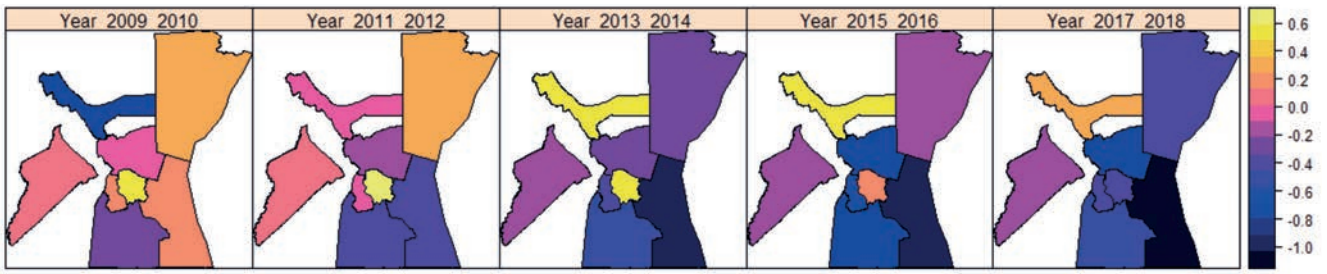


Figure 9. Choropleth maps showing residual spatial effect of adult tick abundance 2009-2018.

Temporal effects

For larvae, there was a steady decreased in the mean count between periods 1 and 5, then a steady increase from period 4. The graph indicates that larvae abundance declined between 2009 and 2016, after which there was a steady increase in the abundance larvae ticks. The graph indicates that nymph abundance increased between periods 2009 and 2012 then declined between 2012 and 2016, after which there was a steady increase in the abundance of nymph ticks. For adults, the abundance was constant from 2009 through 2012, after which a gradual decline was observed from 2012 until 2018 (Figure 10).

Discussion

This study applied the Poisson and NB count data models to tick count data with the aim to explore the influence of environmental and temporal predictors on the distribution of tick counts in Virginia, U.S. We relaxed the assumption that the relationship between the predictors and the response variables in a regression model are constant across the study region and over time. This assumption is unrealistic for spatial processes as factors such as sampling variation and different relationships across regions, for example, attitudes, preferences, culture and others, contribute to different responses to the same stimuli as one moves across regions and over time. A frequent approach to spatial modelling dates back to the work by Besag *et al.* (1991) which was extended by Bernardinelli *et al.* (1995) to include a linear term for space-time interaction. Many studies have relaxed this assumption, *e.g.*, Assunçao *et al.* (1999) introduced a Bayesian space-varying parameter model to examine micro-region factor productivity and the degree of factor substitution in the Brazilian agriculture; Gamerman *et al.* (2003) developed a flexible modelling approach for space-varying regression models; and Okango *et al.* (2015) modelled the HIV and HSV-2 viruses using spatially varying coefficient models. The Bayesian spatio-temporal process model was used to allow covariates to vary spatially and over time. We specified the CAR prior for the structured random effects; autoregressive of order one (AR1) prior for the temporal random effects; and normal iid priors for the unstructured random effects. Non-linear effects of the month variable on tick counts were also evident. As a result, an assumption of linear relationship would have led to misleading results and consequently to wrong interpretations.

The exploratory data analysis showed that more larvae counts were observed compared with nymphs and adult tick counts. We found that the effects of covariates on tick counts varied spatially across counties and over time. Spatio-temporal models were powerful, in a sense that they were capable of revealing county specific effects of each covariate, county space-time effects and the effects of change in time on the distribution of life-stages of lone star ticks. We were able to show that tick abundance has been increasing over the study time in Virginia, which confirms the previous results by Lantos *et al.* (2015) that observed the significant expansion of Lyme disease distribution in Virginia between 2000 and 2014, particularly southward into the Virginia mountain ranges.

It is clear that change in temperature affect tick numbers, such that they decrease in winter (Linske *et al.*, 2020). We confirmed this unsurprising fact and noted also that larvae count remained low up to May. Non-linear effects of the month showed that nymphs and adults were observed in spring and summer, meaning that the distribution of larvae would be expected to increase later that summer and thus adults the following spring. This happens in locations early summer thus determining the distribution of larvae and adults in late summer and spring the following year, respectively (Stein *et al.*, 2008).

Our study also revealed that the effects of habitat type were high in Chesapeake County. This could be because of the abundant white-tailed deer population found in forests, farms, parks and backyards throughout the Chesapeake Bay Watershed. The lone star tick is said to be very aggressive and specific when looking for hosts (Goddard and Varela-Stokes, 2009), but they are unspecific during each life-stage, as this species is found on humans, domesticated animals, ground-dwelling birds as well as on small and large wild animals (Sonenshine and Stout, 1971; Kollars, 1993). The white-tailed deer feeds on fruits and vegetation that are available to them each season, which makes it simple for ticks to attach and feed from these animals. During warm seasons (summer and spring), these animals feed on green plants, during fall they feed on nuts, acorns and crops and in winter they feed on woody vegetation, such as bark, twigs and buds of hardwood and pine trees, where indeed ticks are found (Willis *et al.*, 2012).

Limitations

The failure to account for excess zeros in the discrete distribu-

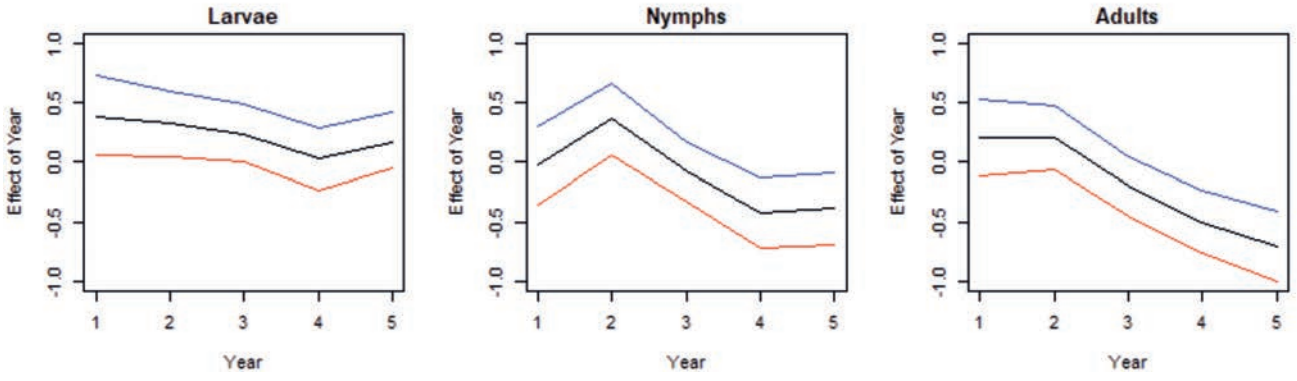


Figure 10. The temporal year random effect for the cumulative best fitting model. Black line = the estimated posterior mean; blue and red lines = the upper and lower 95% credible interval.

tions was a problem. The study focused only on the univariate independent models rather than joint multivariate modelling of counts. In the future, zero-inflated models should be developed and applied to the data. Covariate-time-space interaction effects could also be incorporated. This interaction is able to show the effects of predictors of tick life-stage counts in space and over time. Due to limited data, the study only looked at three environmental and two time predictor variables.

Conclusions

Tick counts are influenced by environmental factors and seasonal changes. There is no dominant weekly influence or observable change in the number of ticks due to the changing weeks. We conclude that tick numbers depend on the type of habitat where they are closer to their hosts and the time when their hosts are more likely to be targeted. Grassy and wooded places are the most liked by ticks as hosts feed and live in such places. Larvae counts are to be expected during summer between June and August, while nymphs are found in abundance between February and May, while adult counts appear mainly between February and May.

References

Anderson JF, 2002. The natural history of ticks. Med Clin North Am 86:205-18.

Assunçao J, Gamerman D, Assunçao R, 1999. Regional differences in factor productivities of Brazilian agriculture: a space varying parameter approach. In Proceedings of the XV Latin American Meeting of the Econometric Society.

Bernardinelli L, Clayton D, Montomoli C, 1995. Bayesian estimates of disease maps: how important are priors? Stat Med 14:2411-31.

Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43:1-20.

Bowman AS, Nuttall PA, Chappell L, 2004. Ticks: biology, disease and control. Parasitol 129:S1-S1.

Bratton RL, Corey GR, 2005. Tick-borne disease. Am Fam Physician 71:2323-30.

CDC, 2020. Guide to the surveillance of metastriate ticks (Acari: Ixodidae) and their pathogens in the U.S. Division of Vector-Borne Diseases, CDC, Atlanta & Ft. Collings.

Childs JE, Paddock CD, 2003. The ascendancy of Amblyomma americanum as a vector of pathogens affecting humans in the United States. Annu Rev Entomol 48:307-37.

Cliff AD, Ord JK, 1981. Spatial processes: models & applications. Taylor & Francis, Boca Raton, FL, USA.

Fritzen CM, Huang J, Westby K, Freye JD, Dunlap B, Yabsley MJ, Schardein M, Dunn JR, Jones TF, Moncayo AC, 2011. Infection prevalences of common tick-borne pathogens in adult lone star ticks (Amblyomma americanum) and American dog ticks (Dermacentor variabilis) in Kentucky. Am J Trop Med Hyg 85:718-23.

Gamerman D, Moreira AR, Rue H, 2003. Space-varying regression models: specifications and simulation. Comput Stat Data Anal 42:513-33.

Goddard J, Varela-Stokes AS, 2009. Role of the lone star tick, Amblyomma americanum (L.), in human and animal diseases. Vet Parasitol 160:1-12.

Hendricks B, Mark-Carew M, Conley J, 2017. Evaluating the utility of companion animal tick surveillance practices for monitoring spread and occurrence of human Lyme disease in West Virginia, 2014-2016. Geospat Health 12:582.

Kollars TM, 1993. Ticks (Acari: Ixodidae) infesting medium-sized wild mammals in southwestern Tennessee. J Med Entomol 30:896-900.

Lambert D, 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34:1-14.

Lantos PM, Nigrovic LE, Auwaerter PG, Fowler Jr VG, Ruffin F, Brinkerhoff RJ, Reber J, Williams C, Broyhill J, Pan WK, 2015. Geographic expansion of Lyme disease in the southeastern United States, 2000-2014. In Open forum infectious diseases, pp. ofv143. Oxford University Press.

Linske MA, Williams SC, Stafford KC, Lubelczyk CB, Henderson EF, Welch M, Teel PD, 2020. Determining effects of winter weather conditions on adult Amblyomma americanum (Acari: Ixodidae) survival in Connecticut and Maine, USA. Insects 11:13.

Manda SO, Leyland A, 2007. An empirical comparison of maximum likelihood and Bayesian estimation methods for multivariate disease mapping: theory and methods. S Afr Stat J 41:1-21.

Mixson TR, Campbell SR, Gill JS, Ginsberg HS, Reichard MV, Schulze TL, Dasch GA, 2006. Prevalence of Ehrlichia, Borrelia, and Rickettsial agents in Amblyomma americanum (Acari: Ixodidae) collected from nine states. J Med Entomol 43:1261-8.

Nathavitharana RR, Mitty JA, 2015. Diseases from North America: focus on tick-borne infections. Clin Med (Northfield Il) 15:74.

Ngesa O, Mwambi H, Achia T, 2014. Bayesian spatial semi-parametric modeling of HIV variation in Kenya. PLoS One 9:e103299.

Okango E, Mwambi H, Ngesa O, Achia T, 2015. Semi-parametric spatial joint modeling of HIV and HSV-2 among women in Kenya. PLoS One 10:e0135212.

QGIS DT, 2009. QGIS Geographic Information System, Open Source Geospatial Foundation.

Salman MD, 2012. Ticks and tick-borne diseases: geographical distribution and control strategies in the Euro-Asia region; CABI. Available from: https://www.cabi.org/bookshop/book/9781845938536/

Sayler K, Rowland J, Boyce C, Weeks E, 2017. Borrelia burgdorferi DNA absent, multiple Rickettsia spp. DNA present in ticks collected from a teaching forest in North Central Florida. Ticks Tick Borne Dis 8:53-9.

Sonenshine, D. E. (1979) Insects of Virginia: No. 13: Ticks of virginia (Acari, Metastigmata). Research Division Bulletin 139, Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Sonenshine DE, Stout IJ, 1971. Ticks infesting medium-sized wild mammals in two forest localities in Virginia (Acarina: Ixodidae). J Med Entomol 8:217-27.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A, 2002. Bayesian measures of model complexity and fit. J Roy Stat Soc Ser B (Stat Method) 64:583-639.

Stein KJ, Waterman M, Waldon JL, 2008. The effects of vegetation density and habitat disturbance on the spatial distribution of ixodid ticks (Acari: Ixodidae). Geospat Health 2:241-52.

Stromdahl E, Hickling G, 2012. Beyond Lyme: aetiology of tick-borne human diseases with emphasis on the South-Eastern United States. Zoonoses Public Health 59:48-64.

Tälleklint L, Jaenson TG, 1998. Increasing geographical distribu-

tion and density of Ixodes ricinus (Acari: Ixodidae) in central and northern Sweden. J Med Entomol 35:521-6.
Willis D, Carter R, Murdock C, Blair B, 2012. Relationship between habitat type, fire frequency, and Amblyomma americanum populations in east-central Alabama. J Vector Ecol 37:373-81.
Zeileis A, Kleiber C, Jackman S, 2008. Regression models for count data in R. J Stat Softw 27:1-25.