Modelling the prevalence of wildlife diseases using simulated diagnostic test data

SUBMITTED BY ANNA FRANCES BUSH

TO THE UNIVERSITY OF EXETER AS A THESIS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGICAL SCIENCES,

AUGUST 2023

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Abstract

Bayesian Latent Class Models (BLCMs) are algorithms that are used to infer disease prevalence when true disease statuses and gold-standard diagnostic tests are not available. However, limited attention has been given to the specification and validation of BLCMs, which are necessary if credible estimates of diagnostic test performance and disease prevalence are to result.

Across six technical chapters, this thesis investigates the fundamental principles of specification and validation via a series of experiments that apply BLCMs to ante-mortem diagnostic test data. To achieve this, simulated arrays of diagnostic test data are generated to reflect the reality of the imperfect trapping and testing efforts that take place in nature. Moreover, the classic Hui-Walter algorithm is generalised within a Bayesian framework to unlock the capability of BLCMs to handle both varying prior information and varying hypotheses simultaneously.

Methods to validate BLCMs are developed and then scaled up across a wide range of possible diagnostic testing scenarios via the creation of procedures to explore high-dimensional parameter spaces. For the first time, it is demonstrated that the credibility of BLCM inferences is in fact predictable.

Among the key findings discovered are dependence structures that are critical to the identifiability of BLCMs; these structures are uncovered at the limits of parameter spaces, and between the means and variances of the inferred statistics. Accordingly, methods are explored to mitigate for these structures as a further prerequisite to obtaining credible estimates.

Attention then turns to testing the core assumptions used to specify the generalised Hui-Walter algorithm. The assumptions about where the true values of diagnostic test performance and disease prevalence exist are removed, and the resulting sensitivity analyses provide confirmation that the findings reported throughout the thesis are indeed generalisable, even to unusual testing scenarios.

With a rigorous validation protocol in place, a novel class of time-dependent BLCMs is specified, and then provided with data from one of the world's longest running wildlife studies. New and rigorously validated inferences of disease prevalence are revealed, and anecdotal trends are corroborated, highlighting the real-world applications of this thesis.

Preface

I spent many happy childhood summers on my Grandad's farm in Herefordshire, where there was a large sett of badgers in the woodland above his top field where he used to keep game birds. At dusk, if I lay very still on the grassy slope of the field before the cattle wire fence, I could experience the thrill of seeing and hearing the badgers move up close. Having recently lived on a cattle farm in Cambridgeshire, which this year experienced an inconclusive reactor, I have a first-hand understanding of the upset that even the suspicion of bovine TB in a herd can cause.

I am fortunate to have had a fascinating exposure to our natural world, which has inspired and shaped my career in conservation, but my specific interest in wildlife disease modelling unexpectedly began during my Master's degree in 2016. After my dissertation project based around the use of fixed-point photography fell through, I found myself with an unexpected alternative project, kindly supervised by Prof. Dave Hodgson, who supervised this present thesis. This alternative project required me needing to learn how to code, and to then be able to simulate disease flows between social animals. "Baptism of fire" somehow understates the experience. But gratifyingly, I was able to offer improvements on the well-known susceptible-infected-recovered model, and I became fascinated both by the subject, and coding, and—in combination—what I could possibly do next.

After an expedition to remote South Africa researching leopards, this part-time and self-funded PhD project started on 05 February 2018, on the same day as my first job at Natural England. In total, these past five and a half years have

brought with them five new jobs, four house moves—and a lot of work—and I'm now finishing this PhD thesis where it all began, back in Cambridge.

I'm proud to have established myself as an environmental planning professional, at a principal grade, at the same time as making my ambition to "discover" using wildlife disease models a reality—particularly in the aftermath of the COVID-19 pandemic.

Globally, the evidence seems clear: we're in a climate and nature crisis, and I'm excited to combine my skills from "work" and "PhD" to contribute to making a difference in the battle that lies ahead.

Acknowledgements

I would like to thank my primary supervisor Professor Dave Hodgson for his guidance, support, feedback, and technical input throughout the past five and half years of my PhD programme. From initially recommending that I read Professor Timothy Roper's book "Badger", to providing detailed commentary on final drafts, Dave's feedback has been crucial to the various pieces of research that this thesis has explored.

I would also like to thank my secondary supervisor Professor Robbie McDonald for his valuable perspectives on my findings at key project milestones, particularly at the gateway for acceptance into the writing-up phase of this programme. Thanks are also due to Dr Dave Hudson for helpful Teams chats over the years, and to the wider "Team Hodgson" for providing me with support and kind words at research group meetings in Penryn and online.

I am grateful for the help that I received from Exeter University colleagues Dr Matthew Silk, Dr TJ McKinley, Dr MD Sharma and Professor Mario Recker, who—respectively—critiqued several of my early BUGS models, advised me on ways to reduce model run time, helped me to access the university's remote computing facilities, and provided me with an initial evaluation on what has now become Chapter 2 of this thesis. In addition, I am grateful to Professor Julian Drewe from the Royal Veterinary College who supplied me with the BUGS code underpinning Drewe *et al.*, 2010, a key paper that influenced the topic of this thesis.

I would like to acknowledge the role of Professor Jon Blount for accepting my initial proposals to study for a part-time doctorate in disease ecology at the university while also pursuing a career at Natural England; Dr Jenni McDonald

who kindly shared preliminary thinking and code on latent class approaches for estimating disease prevalence, which I used to generate my initial hypotheses; Dr Clare Benton at the Animal and Plant Health Agency who arranged my access to a portion of the Woodchester Park dataset used in Chapter 8 of this thesis; Dr Barbara Tschirren and Professor Erik Postma from the wider Biosciences department in Penryn who provided me with invaluable verbal and written feedback following my initial review viva in February 2019; and Exeter University library staff for speedily and efficiently sourcing and delivering to me non-catalogue reference texts upon request.

My inspiration and understanding has benefitted from attending the *Advanced Bayesian Modelling with BUGS* course at the MRC Biostatistics Unit, University of Cambridge, run by Dr Christopher Jackson, Dr Robert Goudie and Dr Anne Presanis; Professor Andrew Gelman's insightful blog on Statistical Modelling, Causal Inference, and Social Science; reading Professor Sir David Spiegelhalter's Guardian column on the statistics representing the Covid pandemic; and being on the mailing list of the BUGS archives, as well as a grateful user of StackExchange and StackOverflow.

I must also acknowledge my debt to my employer, Natural England, and particularly colleagues within the Chief Scientist's Directorate, for the flexibility and support provided to me at work.

Finally, it would not have been possible to write this thesis without the support and encouragement that I have had from family and friends, with especial thanks due to Samuel and Malcolm.

Table of Contents

List of Figures

List of Tables

List of Equations

Definitions

This section provides definitions for the key terminology employed within this thesis. Listed in alphabetical order, they provide a convenient reference resource for the reader, complimenting the more substantive terminology definitions and explanations provided in the core chapters one to eight.

Definitions are necessary, because throughout the relevant literature, there is widespread inconsistency in the usage of many of the terms in question. For instance, the terms “estimate”, “predict” and “infer” are often used synonymously, with precise definitions being difficult to come by; and the definition of “error”—a metric used to quantify “noise” in a system—is highly specific to the experimental design in question. None of these variations in usage are necessarily “wrong”, but the resulting discordance is hardly helpful to either practitioners or the broader development of the Bayesian Latent Class Models with which this thesis is concerned.

Three specialised dictionaries published by the Oxford University Press have been consulted when formulating this *lingua franca*: A Dictionary of Epidemiology 6th Edition (Porta, 2016), A Dictionary of Statistics 3rd Edition (Upton and Cook, 2014), and A Dictionary of Ecology 5th Edition (Allaby, 2015). Accordingly, the collection of 36 select definitions outlined below should form a welcome contribution to the discipline of disease ecology since it provides a summary of the words crucial for communicating to fellow academics—statisticians, ecologists, epidemiologists and beyond—on the subject of Bayesian Latent Class Models.

Accuracy: A metric summarising the distance between an inferred value and the truth (Cochran, 1977), with small values describing a relative lack of error

(Porta, 2016) associated with that inference. Accordingly, this thesis examines the accuracy of diagnostic test sensitivity (Se), diagnostic test specificity (Sp) and disease prevalence (P), where a high degree of accuracy is achieved when the posterior inference is close to the parameter's true value. Specifically, the accuracies of Se and Sp indicate how well a "test"—i.e., a diagnostic test for a wildlife disease—can produce a correct outcome, whereas the accuracy of P indicates how well a Bayesian Latent Class Model infers the proportion of infected individuals within a sampled population. In this thesis, accuracy is quantified by the metric inferential error or the metric inferential bias, and the term accuracy should not be confused with the term diagnostic accuracy.

Bayesian Latent Class Model (BLCM): A method for classifying observed data into unobservable groups using Bayesian inference (Li *et al.*, 2018). This method offers an approach to inferring Se, Sp and P using probability distributions given multiple imperfect tests, and given test data where the true disease statuses of individuals are unknown. In this thesis, BLCMs are declared using the JAGS language and may be referred to as "the model" or "models" for brevity.

Bias: see inferential bias.

Constraint: A condition placed on either the prior knowledge that a model uses, or the inference framework, in order to direct model outcomes (Berkvens *et al.*, 2006). Two types of constraint are defined for the purposes of this thesis: parameter constraints are used to control the space in which the truth can lie, and are applied to the true values in simulation studies; and prior constraints are used to direct the information provided to the BLCM and are applied by restricting prior distributions to justified ranges.

Degrees of freedom: The number of freely varying units of information associated with an inference (Lynn and Healey, 1992), which in this thesis is calculated as the number of possible diagnostic test outcomes minus one, in accordance with Siegel and Castellan, 1988. Note, a second type of degrees of freedom exists within the work presented in this thesis, associated with the regression models that are specified. Degrees of freedom in this case are calculated as a function of the sample size used, and are not explicitly reported given that each regression uses data from hundreds of simulations.

Deterministic method: An approach to calculating Se, Sp and P perfectly (Upton and Cook, 2014; Porta, 2016) under the assumption that all the required data is present, i.e. that a population has been censused perfectly. The work presented within this thesis tests this assumption, and so calculates Se, Sp and P using stochastic methods.

Diagnostic accuracy: A term that expresses the collective Se and Sp of an imperfect diagnostic test (Porta, 2016). The diagnostic accuracy of any diagnostic procedure or test describes how well it discriminates between health and disease, and improvements in diagnostic accuracy bring the diagnostic test closer to a gold standard diagnostic test.

Diagnostic Test: Any procedure or information—such as a medical observation or an expert opinion—that can be used to diagnose infections with an assigned diagnostic accuracy. Note, the term “battery of diagnostic tests” (McDonald and Hodgson, 2018) is used in this thesis to describe groups of two or more diagnostic tests.

Disease Prevalence (P): the proportion of infected individuals within a population (Porta, 2016).

Error: see inferential error.

Global statistic: Global statistics or “grand means” are arithmetic averages, calculated irrespective of group (Upton and Cook, 2014), that are used to provide signals that make meaningful statements at a given level (Vesely, Finos and Goeman, 2021). Accordingly, in this thesis global statistics describe the collective accuracy or precision of diagnoses by averaging across all inferred parameters—Se, Sp and P—for a given position in parameter space, where the subsets of values making up the average inferences of Se, Sp and P are of an equal sample size, and are considered irrespective of any biological implications of Se, Sp and P. The experiments described in this thesis therefore use global statistics to highlight volumes of parameter space that require further statistical investigation—i.e., to provide a preliminary determination of those regions where truth influences the accuracies and precisions of inferences of Se, Sp and P in different ways—but not to infer causation, in order to avoid making inferences that might be in conflict with Simpson’s Paradox (Simpson, 1951). Note, in this thesis, the inferential error of a global statistic is termed a global error, and the inferential bias of a global statistic is termed a global bias.

Gold standard: A diagnostic test where both Se and Sp have true values of one, meaning that diagnostic accuracy is perfect. Note, even this well-known phrase is subject to a diversity of usage, for example, it is used synonymously with the phrase “reference standard” (Bachmann *et al.*, 2005; Hahn, Schwarz and Frickmann, 2019), which is often used to mean a widely accepted but imperfect standard (Miller, 2012) that may theoretically be bettered. To address this diversity, diagnostic standards have been sub-classified as, for example, “silver” standards, when Sp is perfect and Se is imperfect, and “bronze” standards, when Se is perfect and Sp is imperfect (Wu *et al.*, 2016).

Importantly, the work presented in this thesis addresses ecologists' requirement to diagnose infections ante-mortem, in instances where gold standard tests are desired but not available. It follows that while "gold", "silver", and "bronze" standards are useful and of value, they cannot be applied to the problem of diagnosing ongoing infections in many wildlife disease systems—such as the badger-bovine tuberculosis system investigated in Chapter 8—where perfect diagnoses can only be established post-mortem.

Hyperparameter: A configuration variable given to a BLCM that defines how it should operate. The hyperparameters defined for the purposes of this thesis are described in

Table 10-3: The MCMC hyperparameters used to define the JAGS models written using the jagsUI package (Kellner, 2015), their values, and why those values were chosen. These hyperparameters are relevant to the simulation analyses conducted between Chapters 5 to 7.

.

Identifiability: A term to describe whether inferring Se, Sp and P is possible, given a model and the available data. The word possible is caveated by the fact that the BLCM may produce inferences, but these inferences sometimes may not be rational, or improve on the existing prior information.

Inference: Inferences describe characteristics of a posterior distribution using the best available evidence (Upton and Cook, 2014), therefore offer informed estimates of true parameter values, and in this thesis inferences are made using Bayes' theorem. Note, in contrast to an inference, a prediction is an evidence-based speculation, usually regarding the probability of certain outcomes (Upton and Cook, 2014), that for the purposes of this thesis is made

outside of the frameworks of a regression or a Bayesian inference. Moreover, in contrast to an inference or a prediction, an estimate is a realised value given data (Upton and Cook, 2014), which in this thesis describes the population-level findings of regression analyses.

Inferential bias or "bias": The systematic difference between an inference and the truth (Rothman, Greenland and Lash, 2014; Allaby, 2015). Biases represent the directionality of inferential error. Accordingly, bias is calculated in this thesis as the difference between inferred values of Se, Sp or P and the true values of Se, Sp or P. Bias indicates if a parameter has been underestimated or overestimated, where a negative bias corresponds to an underestimation and a positive bias corresponds to an overestimation, and a bias value close to zero is close to the truth. Bias measurements are particularly relevant when a single truth can be inferred by replicated simulations: the bias of the average inference shows whether a BLCM "tends" to overestimate or underestimate that truth.

Inferential error or "error": The degree to which a measurement is mistaken (Porta, 2016). Accordingly, in this thesis, error is calculated as the absolute value of the difference between the inferred values of Se, Sp and P, and the true values of Se, Sp or P. This method of calculating error can be used to represent a single simulation of a BLCM, or represent the mean difference between the truth and the sample mean, across all replicates of the model; the results of this thesis consider the latter. Error therefore represents a difference between probabilities, and errors are reported on in terms of their magnitudes.

Inferred parameter: an estimate of the true value of a parameter using Bayesian inference, with an associated accuracy and precision. In this thesis, when the term is used as a plural—inferred parameters—the inferred values of

Se, Sp and P are being referred to collectively, and this phraseology is not to be confused with a global statistic. Inferred parameters are often represented by the mean of the posterior inference, and as a rule, the terms “posterior inference”, “inference”, and “inferred parameter” should all be interpreted as describing the mean of the posterior inferences for a given parameter.

Latent Class Model (LCM): A statistical method to classify the unobservable heterogeneity within sampled data into subgroups (Andersen, Hagenaars and McCutcheon, 2003; Rothman, Greenland and Lash, 2014). Accordingly, in this thesis, individuals from infected wildlife populations are classified into the categories infected or uninfected based on information from multiple diagnostic tests.

Markov Chain Monte Carlo (MCMC): An algorithm used to explore likelihood functions—i.e., all the statistical evidence that the available data can provide—across parameter spaces while working to infer the posterior distribution. Monte-Carlo methods allow the estimation of the properties of a distribution by analysing random samples, and a Markov Chain is the enabling sequential process (van Ravenzwaaij, Cassey and Brown, 2018). To realise these methods in this thesis, the Bayesian modelling tool JAGS is used to sample probability distributions using verified methods, avoiding the need to write an MCMC sampler from scratch.

Mean-variance relationship: A statistical relationship describing how the variance of parameter values change as a function of the mean of parameter values, which in this thesis relates to how the accuracy or precision of inferred parameters vary across a parameter space.

Model validation: A process to evaluate whether an inferential method is sound (Porta, 2016). Model validation is used in this thesis to evaluate of the ability for a BLCM to infer values of Se, Sp and P in terms of the data available.

Parameter: Parameters, in this thesis, are latent population-level metrics that numerically describe a version of the truth. These parameters are described within the model using the JAGS language and can be inferred using given data to identify values of Se, Sp and P. These parameters are defined in Table 10-2.

Parameter space: A parameter space, in this thesis, is a reference to a defined point within a hyperdimensional space, an entire hyperdimensional space, or the space that an MCMC sampler is given to search for the posterior distribution within. A global parameter space is a fully unconstrained parameter space containing all feasible values of parameters Se, Sp and P.

Posterior distribution: Probability density functions that summarise the information that a Bayesian model can infer about a latent parameter (Upton and Cook, 2014).

Posterior inference: A metric that describes the posterior distribution of an inferred parameter. The preferred metric used in this thesis is the mean of the posterior distribution.

Precision: A metric summarising the ability to estimate consistently (Hellmann and Fowler, 1999) in terms of the quality of a single outcome, or the closeness of replicate outcomes to each other (Feinleib and Zar, 1975). Accordingly, in this thesis, precision is used to describe the replicability of inferred parameters, and is studied as an among-replicate metric representing the mean of the standard deviations of the posterior distributions. Note, as the standard

deviation of posterior distributions increase, the precision of the posterior distributions decreases.

Prior distribution: A probability density function that summarises what is already known about a parameter.

Prior information: This term is used to describe any existing beliefs about a parameter than can be supplied to the BLCM via a range of methods including the specification of prior distributions and their constraints, the provision of further diagnostic tests, the provision of more samples, and—for simulation analyses—the constraints of true values.

Sensitivity analysis: A method to evaluate the ability of models to infer values given new information, or changes to model assumptions (Porta, 2016). Accordingly, in this thesis the models are BLCMs, and the inferred values relate to the parameters Se, Sp, and P. A Global Sensitivity Analysis is a method that allows all uncertainties associated with an experiment to vary simultaneously across simulations (Saltelli *et al.*, 2020). This approach, applied in Chapter 7, tests the robustness of a BLCM across the full range of true values it could be presented with, i.e. the global parameter space.

Statistical artefacts: Observed errors in the statistical representation of data (Scott and Marshall, 2009). Accordingly, in this thesis, statistical artefacts are statistical trends that directly influence how “solvable” any region of parameter space is.

Stochastic method: A modelling approach where random processes are used (Porta, 2016). Stochastic approaches are used in this thesis to generate the true values of Se, Sp, and P for use in simulation analyses, accounting for the inability to trap an entire population of animals, and the resulting inability to

therefore test an entire population of animals for an infection (note, this thesis uses the words infection and disease synonymously).

Test Sensitivity (Se): The probability that a positive test outcome correctly describes infection (Rindskopf and Rindskopf, 1986).

Test Specificity (Sp): The probability that a negative test outcome correctly describes the absence of infection (Rindskopf and Rindskopf, 1986).

Truth (or true values): The known values of Se, Sp, and P in a simulation analysis. The true parameter is latent and so error free, but never known in the real world; but true parameters can be set in simulation analyses. Consequently, in this thesis the truth is a feature of a population.

Time decomposition: A statistical technique used to manipulate longitudinal data into categorical time-dependent components (Tuncer, Tanik and Allison, 2008), which is applied to the specification and capability of BLCMs in Chapter 8 of this thesis in order to unlock their ability to infer trends and change points in Se, Sp and P through time.

Abbreviations

A concise list of the abbreviations most used throughout this thesis.

BLCM – Bayesian Latent Class Model.

LCM – Latent Class Model.

LMM – Linear Mixed effects Model.

MCMC – Markov Chain Monte Carlo.

P – Disease Prevalence.

Se – Diagnostic test Sensitivity.

Sp – Diagnostic test Specificity.

Phat – The mean inferred value of disease prevalence.

Sehat – The mean inferred value of diagnostic test sensitivity.

Sphat – The mean inferred value of diagnostic test specificity.

n.tests trend – A trend showing that as the numbers of diagnostic tests available increase, the error of the inference decreases.

Chapter 1

1. General Introduction

The estimation problem that this thesis concerns is "how to accurately diagnose infection in wild animals". Ultimately, this is an ecological problem, with its resolution reliant upon statistical modelling. This chapter is an introduction to the ecological problems that motivated the statistical models advanced within this thesis. A specific introduction to these statistical models is provided in Chapter 4, which also serves as an overview of the 6 simulated datasets listed in Table 10-1.

Foreword

Any infection that can spill between animals and humans or *vice versa* is called a zoonosis, and zoonoses are responsible for most new diseases in humans.

On 11 March 2020 the World Health Organisation declared the outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (WHO, 2020)—the causative agent of COVID-19—to be a global pandemic. SARS-CoV-2, thought to originate from bats, overcame species boundaries to successfully maintain infections in humans, and is one of at least 250 known zoonotic viruses (Mollentze and Streicker, 2020) that have the potential to follow suit.

Mycobacterium bovis, the causative agent of bovine tuberculosis (bTB), is another example, this time bacterial, of a zoonosis with reported spillover infections into humans, that in England are largely controlled by the pasteurisation of milk. The bTB epidemic in England has been persisting in

reservoir i.e., primary, and secondary, wildlife hosts such as badgers for well over 50 years, and its control remains a “perfect storm” (Allen, Skuce and Byrne, 2018) despite long-term research focused on wild badger *Meles meles* reservoirs—for example at Woodchester Park, Gloucestershire, England.

Regardless of the pathogen—or how it may currently be controlled—these important zoonoses have brought to the fore the complexity of infection management, and testing. The pandemic has also highlighted that information about zoonoses still confined to their wildlife reservoirs is critical to both wildlife and public health disease management globally; and gathering this information is dependent upon being able to reliably test animals for infection.

We need to reliably test animals for infection.

In 2021, medical journal The Lancet reported that the number of COVID-19-related publications on the PubMed database surpassed that of any disease outbreak in the last hundred years: in March 2021 the count was greater than 110,000 (Winkler *et al.*, 2021); a statistic that in January 2023 was a count of greater than 330,000, illustrating just how important it is to understand wildlife reservoirs.

Most epidemics and pandemics in humans originate from human interactions with reservoirs of disease maintained across wild animal populations, and these diseases are often of concern for humans and livestock (Krebs *et al.*, 1998). For instance, the commonly recognised infections that cause measles, mumps and rubella are all thought to have originated from human interactions with domesticated animals and or wildlife (Wolfe, Dunavan and Diamond, 2007; Bennett *et al.*, 2020; Düx *et al.*, 2020). These interactions are complex and appear to be a consistent driver of zoonotic concern (Gibb *et al.*, 2020) common

to most circumstances via which zoonoses are thought to proliferate, inclusive of the spatial proximity of humans to livestock and or rodents; the taxonomic relatedness between humans and other primate species susceptible to the same pathogens; and the increasing anthropogenic disturbance of mammals such as bats, which are reservoirs of multiple viral pathogens.

While the SARS-CoV-2 virus is thought to have emerged from reservoirs of betacoronaviruses in bat populations (Andersen *et al.*, 2020), to have certainty in this, it is likely that a great many bat populations from across the world would need to be sampled and tested for betacoronaviruses. And here lies the estimation problem that this thesis addresses: how to accurately diagnose infection in wild animals.

The epidemiological challenge

Zoonoses emerge from complicated interactions between social and ecological systems, and are a threat compounded by our inability to accurately estimate disease parameters (DiRenzo *et al.*, 2018) in the absence of data to inform critical decisions about the species of the highest zoonotic concern. The risk of zoonoses, coupled with global declines in biodiversity, are therefore the key drivers of wildlife disease research, inclusive of this thesis.

The accurate diagnosis of infection in wild animal populations can be quantified by two parameters—the sensitivity (true positive rate) and specificity (true negative rate) of a diagnostic test—in addition to the parameter disease prevalence: the percentage of individuals in a population infected by a given pathogen (Jovani and Tella, 2006).

True diagnostic test sensitivity (Se) and specificity (Sp), and disease prevalence (P), are calculated as follows in Equation 1 to Equation 3, respectively.

Equation 1

$$Se = \frac{TP}{N_+} = \frac{TP}{TP + FN}$$

Where in Equation 1, TP is the number of true positive diagnoses, FN is the number of false negative diagnoses and N+ is the number of real infections.

Equation 2

$$Sp = \frac{TN}{N_-} = \frac{TN}{TN + FP}$$

Where in Equation 2, TN is the number of true negative diagnoses, FP is the number of false positive diagnoses and N- is the number of real negative infections.

Equation 3

$$P = \frac{N_+}{N_+ + N_-}$$

A second example of zoonoses is a group of bacteria called the *Mycobacterium tuberculosis* complex (MTC), which is inclusive of *Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB). Like SARS-CoV-2, *Mycobacterium tuberculosis* is maintained by humans, and can infect animals via spillover events, a transmission event where pathogens cross the human to animal boundary (Becker *et al.*, 2019) or *vice versa* (Ellwanger and Chies, 2021). TB is the leading cause of human deaths from infectious disease worldwide, surpassing COVID-19 in second place (WHO, 2022), and most cases of TB are hidden, exhibiting latency, i.e. a period where individuals are infected but lack

the ability to infect (Barreto, Teixeira and Carmo, 2006); with one third of the world's population likely to be infected by latent TB (Fogel, 2015). Yet latent TB does not have a gold standard diagnostic test, and therefore the disease cannot be diagnosed perfectly (Pourakbari *et al.*, 2018). A positive blood test result for *Mycobacterium tuberculosis* for instance does not provide discrimination between latent and active infection, for this, additional non-perfect diagnostics such as x-rays and or the clinical evaluation of symptoms may be used to increase the certainty of a diagnosis. While a fundamental aim of wildlife epidemiologists is to accurately diagnose infection, the need for better diagnostic tools strongly underpins both human and veterinary medicine.

Another bacteria of the MTC is *Mycobacterium bovis*, the causative agent of bTB, which in England are maintained by populations of the European badger *Meles meles*, the wildlife reservoir, supported by secondary hosts, such as deer (Collard, 2023). In England, bTB is costly to the agricultural sector due to the number of infected cattle that must be culled, in combination with the isolation and testing protocols that farmers must comply with; bTB is also costly to the taxpayer, who funds the Government's badger control efforts. Diagnostics for bTB in live badgers do not have a gold standard—which in this thesis is defined as a diagnostic test where both Se and Sp are 100%—and so possessing reliable measures of estimated diagnostic accuracy is critical to understanding the success of any badger control strategy.

The diagnostic accuracy of tests for *Mycobacterium bovis,* like for *Mycobacterium tuberculosis,* are complicated by, for example, an insensitivity towards disease latency, or mild infections with the absence of physical symptoms (Fitzgerald and Kaneene, 2013). In badgers, even post-mortem pathological examination is associated with a low sensitivity (Gavier-Widén *et*

al., 2009). Diagnostic accuracy is highly dependent upon the context in which the test is used; and this context includes the process in which disease spreads i.e., the pathogenesis, disease prevalence, and any badger control strategies being used.

Why is inferring disease prevalence a key challenge for ecologists?

This thesis does not claim that inferring disease prevalence, ante-mortem, where imperfect gold standards must be relied upon, is a new challenge. True disease prevalence is always a latent and population-specific parameter of diseased wildlife populations, and its inference is always likely to be inaccurate, yet its estimation is critical at the population level for fully understanding disease epidemiology; informing disease control strategies by providing point inferences of the number of infected individuals within a target population which can in turn explain disease dynamics (Helman *et al.*, 2020); and confirming disease control. Even small improvements to the accurate estimation of disease prevalence are therefore useful to researchers (Flor *et al.*, 2020), and the challenge lies in demonstrating to the scientific community that any reported estimates of disease prevalence have been obtained using transparent and robust methodologies.

A reliable inference of diagnostic accuracy is key to minimising false positive diagnoses.

Most diagnostic tests used in wildlife disease studies are not gold standards i.e., they are not error free (Dendukuri *et al.*, 2004), due to a combination of reasons inclusive of the costs of “better” tests, ethical considerations, procedural risks or invasiveness considerations, the need for specialist expertise, and laboratory

delays. Disease prevalence is usually estimated in contexts where the proportion of a population that is diseased is the minority—as is the case at the start of an epidemic—and where a majority of truly uninfected animals must be tested. In this scenario, there is an inherent risk that the number of falsely positive diagnoses will exceed the number of truly positive diagnoses simply due to the proportion of healthy individuals that require testing for disease elimination. This scenario is central to the problem that this thesis addresses, as the following example makes clear.

Consider a scenario where the true positive rate, Se, is calculated as per Equation 1, the true negative rate, Sp, is calculated as per Equation 2. In addition, the false positive rates of infection, and the false negative rates of infection are calculated using Equations 4 and 5 respectively.

Equation 4

$$False\ positive\ rate = 1 - \frac{TN}{N}$$

Equation 5

$$False\ negative\ rate = 1 - \frac{TP}{N}$$

Under such a scenario, a given population of wild animals is suspected of having a disease, infection X, which is thought to infect 2% of these animals. The best available diagnostic test for infection X has a reasonably high diagnostic accuracy, meaning that if animal A has infection X, it will produce a positive test result 95% of the time; and if animal A does not have infection X, it will produce a negative test result 95% of the time. 1000 animals could be tested.

Given this scenario, 68 tested positive (total positives), with 19 true positive and 49 false positive; and 932 tested negative (total negatives), with 931 true negative and 1 false negative. Although the best available test correctly identifies true positives or true negatives 95% of the time, 49 animals were classified as infected when they were not. This means that given a positive test result (of which there were 68), the actual probability of the individual having the disease given 19 true positive results is just 27.9%.

This scenario makes it clear that if diagnostic accuracy is not error free—like most diagnoses in real life—diagnostic tests should not be used naively to confirm the presence or absence of infection. Further, given that diagnostic accuracy is situation-dependent, information about the diagnostic situation should also be used to infer diagnostic accuracy. For example, diagnostic accuracy is influenced by variables such as testing strategy and sampling strategy as well as latent variables such as stage of infection at the individual level and pathogenesis at the population level.

Using the same “best available” test, if it is now suspected that 5% of animals are infected, a researcher could expect 48 false positives and 47 true positives. Meaning that a positive test result would only be ~50% likely to be true. And if the diagnostic accuracy of the test is now suspected to be 90%, a positive test result would only be ~32% likely to be true.

This simple example concerning the hypothetical infection X highlights that what is initially believed—from here on termed “prior information”—about disease prevalence and test performance can significantly influence the expected number of positive test results, and our interpretation of the truth. To emphasise, the above example reports frequencies; assumes that diagnostic

test sensitivity and diagnostic test specificity are known via the use of a reference test; and that disease prevalence is a point value i.e., measured independently of time at a specific point in time. In this thesis, these assumptions are removed, and the values of diagnostic test sensitivity, diagnostic test specificity and disease prevalence are inferred using probabilities; when their values are not known; when accounting for the uncertainty associated with testing capability and sampling strategy; and in Chapter 8, through time.

Understanding infection systems is directly applicable to the Bayesian philosophy of thinking, which can be used to create statistical models that incorporate prior information about the unknown parameters diagnostic test sensitivity, diagnostic test specificity and disease prevalence in the format of updateable and user-defined probability distributions.

A brief introduction to Bayesian philosophy

Information that is known about any complex system is rarely certain and often subject to additional information being provided. While this concept is just common-sense, it is also the cornerstone of the Bayesian philosophy, which allows a level of uncertainty about any assumptions used to create a Bayesian model, and therefore creates inferences that are essentially “a best guess”. Modern Bayesian analysis brings robustness to this framework by allowing the combination of prior information with data to yield powerful inferences using Monte-Carlo Markov Chain (MCMC) algorithms. In other words, given some awareness about how reasonable some data is, and if that awareness agrees with some newly available data, then it is possible to determine the probability of a hypothesis being correct using Bayes’ theorem (Bayes, 1763). This method

provides ecologists with a way of cultivating certainty in their knowledge about ecological processes by supplementing an often data-limited study with their own logic and expertise to extract a small number of reasonable solutions from a large number of probabilities.

Diagnosing infection without a gold standard test

Ecologists researching wildlife infections often work in an environment where an individual's true disease status is unknown, i.e. latent, and a gold standard diagnostic test is not available. So in this situation, statistical classification methods—belonging to a group of models termed Finite Mixture Models (McLachlan, Lee and Rathnayake, 2019)—must be used to infer diagnostic test sensitivity, diagnostic test specificity and disease prevalence given observed diagnostic test data and the subgroups infected or uninfected.

Latent Class Analyses are arguably the state-of-the-art (Toft *et al.*, 2007) means of estimating unknown diagnostic test sensitivity, diagnostic test specificity and disease prevalence (Hui and Walter, 1980). In essence, this is because:

1. The subgroups infected and uninfected are defined as probabilities, and memberships to each group are not fixed. Since class membership is not directly observed, classification could potentially differ between classifiers.

2. The parameters diagnostic test sensitivity, diagnostic test specificity and disease prevalence can be inferred.

3. The required diagnostic tests do not have to be perfect.

4. A reference test is not required, which is an advantage since it avoids the "Catch-22" situation of having to first evaluate the diagnostic accuracy of this test (Rydevik, Innocent and McKendrick, 2018).
5. Imperfect test data can be used that includes conflicting test results.
6. True diagnostic test sensitivity, diagnostic test specificity and disease prevalence can be inferred with associated accuracies and precisions.
7. The quality of individual diagnostic tests within a battery of diagnostic tests can be determined.

A Latent Class Analysis approach assumes that an individual's true infection status is latent, i.e. hidden, within an array of binary diagnostic test results described at the population level. Individuals are associated with a probability of being infected (+) or uninfected (-) given the results of *multiple*, inaccurate, and independent diagnostic tests (Helman *et al.*, 2020). And tests within a battery of diagnostic tests may be considered independent from each other if each test acts on a different biological component.

The dataset required for a Latent Class Analysis using three independent diagnostic tests would be categorised as the frequencies of subjects with the following sequences of diagnostic outcomes: +++; ++-; +--; ---; --+; -++. For wildlife disease researchers, Latent Class Analyses that are modelled within a Bayesian framework, i.e., within a Bayesian Latent Class Model (BLCM) are especially useful, since prior information such as expert opinion can be used to inform the likelihoods of values of diagnostic test sensitivity, diagnostic test specificity and disease prevalence.

There has been a small and recent increase in the application of BLCMs to wildlife disease research. For example, the approach has been used to infer the performance of tests for brucellosis (Pfukenyi *et al.*, 2020), feline foamy virus (Dannemiller *et al.*, 2020), as well as the diagnostic accuracy of anecdotal reports of foot-and-mouth disease (van Andel *et al.*, 2020). In general, inferences of diagnostic test sensitivity, diagnostic test specificity and disease prevalence are often difficult to obtain due to the lack of field data—which in the case of van Andel *op cit*, was addressed by informing the BLCM with proxy anecdotal tests.

Even with multiple independent diagnostic tests, and some other prior information, it appears that ecologists in practice require better tools and guidance to use the BLCM approach effectively. This is because the theory behind, and the application of BLCMs is complex, combined with an apparent lack of tools and guidance on Bayesian Latent Class Analyses specifically accessible to wildlife disease ecologists. The gulf between what BLCMs can theoretically deliver, and the complexities of actually deploying them—that this thesis attempts to bridge—has almost certainly contributed to why BLCMs have been “applied sparsely in wildlife systems” (Helman *et al.*, 2020).

The sources of bias when testing for infected wild animals using BLCMs

There are three key differences between the contexts of testing regimes applicable to human and wildlife studies, and these differences affect how diagnostic accuracy should be modelled. First, wildlife studies often require the trapping of animals whereas human studies are carried out using voluntary subjects and larger sample sizes can generally be attained. Second, the drivers

of pathogenesis in ecosystems differ to those in human systems, meaning that the environmental stochasticity that must be accounted for in human versus wildlife models is different. Specifically, the results of diagnostic testing regimes in ecological studies may be driven by latent variables characterising elements of wild host-pathogen systems, creating high levels of stochasticity in test results. Third, diagnostic testing regimes in human studies often adhere to widely accepted or gold standards, whereas the diagnostic testing regimes for wildlife diseases most often include imperfect tests, which have been used in comparatively fewer studies with highly variable contexts.

Understanding these biases, and understanding how to take account of them when adopting a BLCM approach, is central to facilitating the greater use of BLCMs within the ecologist community. Those goals are in essence the subject matter of this thesis, and so the three key biases in question are now discussed.

First, animals in wildlife studies must first be trapped prior to testing, and trapping efficiency is rarely 100%; trappability may vary according to, for example, physical and demographic traits; and the stage of infection within the trapped portion of individuals may not be representative of the population—for example, it may have been easier to trap diseased animals.

Second, in addition to the different testing contexts between human and wildlife studies, biological changes across ecosystems also influence pathogenesis via a network of diverse and mostly hidden mechanisms. This concept is usefully illustrated by Darwin's tangled bank theory (Darwin, 1859)—a metaphor for the complex heterogeneity of species and their interactions within the natural environment—which emphasises that the core of how ecosystems evolve and survive is based on co-dependencies (Plotkin, 2017) that may, for example, link

demography to immunity. And while the formal co-dependency between ecological networks and infectious disease was stated in the 1950s by ecologist Charles Elton (Richardson and Pyšek, 2007; Johnson, Ostfeld and Keesing, 2015) research on ecological networks in the context of trophic interactions predominates (Berlow *et al.*, 2009; Ings *et al.*, 2009; Kéfi *et al.*, 2012). Importantly, the systems we observe are only a small subset of those which could possibly exist or be modelled, and many of the co-dependencies that Darwin alludes to in his tangled bank theory will be relevant to understanding disease spread.

Third, this thesis employs BLCMs as a tool to extract “certainty” from imperfect diagnostic test data. A key part of using a BLCM is understanding when it is “identifiable”, i.e. whether deriving diagnostic test sensitivity, diagnostic test specificity and disease prevalence is possible given both the model and data available. Practical identifiability (Kao and Eisenberg, 2018) describes the fit between a BLCM, and the data used to inform it, a fit which includes how well environmental errors are represented (Roosa and Chowell, 2019). The environmental errors that must be accounted for by a BLCM are specific to the contexts within which human and wildlife diseases are studied, with key differences including the variation in possible testing conditions between predominantly clinical and field-based studies—with variation among field-based wildlife disease studies often attributable to latent ecological processes—and the accessibility of subjects to test. Consequently, when models have practical identifiability, the environmental realism that they infer can be applied to wildlife disease data with a greater confidence, with diagnostic test sensitivity, diagnostic test specificity and disease prevalence usually being more precise.

The study population

Long-term datasets detailing the infection history of diseased wildlife populations are rare and valuable (Barroso, Acevedo and Vicente, 2021). In Chapter 8, this thesis applies BLCMs to a longitudinal dataset of diagnostic test results obtained from a wild population of ~300 badgers (Drewe *et al.*, 2010) at Woodchester Park, Gloucestershire, England, in order to infer diagnostic test sensitivity, diagnostic test specificity and disease prevalence. The Woodchester Park mark-and-recapture study has been running since 1975 (Delahay, Cheeseman and Clifton-Hadley, 2001), and the population is naturally infected with bTB. Approximately 80% of the population is trapped each year (White and Harris, 1995), and badgers are trapped following seasonal patterns to avoid trapping lactating females that may have dependent cubs underground (personal communications, 28 March 2018). The long-term nature of the study means that a body of previous literature is available on the Woodchester Park badgers, including previous estimates of diagnostic test sensitivity, diagnostic test specificity and disease prevalence to compare findings to; as well as a library of published information about the study population itself, for example, see McDonald, Robertson and Silk, 2018.

Thesis outline

This thesis contributes to the small but growing body of work devoted to applying BLCMs to wildlife disease data; and the even smaller body of work focused on developing its proper application. The models and workflows presented are highly generalised, meaning that they can be quickly adapted to a diversity of real-world and hypothetical disease monitoring scenarios for both wildlife and human host-pathogen systems without the need for significant

changes to the contributed code available at https://github.com/annabush/PhD. Moreover, the models and workflows presented in this thesis build on two approaches to large-simulation analyses already well-described in ecological literature: Bayesian approaches using the BUGS language, and maximum-likelihood approaches using the R package `lme4` (DiRenzo, Hanks and Miller, 2023).

This thesis specifies a series of BLCMs run on simulated diagnostic test data, and then uses this architecture to demonstrate clear workflows allowing the robust inferences of diagnostic test sensitivity, diagnostic test specificity and disease prevalence. If model uncertainty is better understood, and more co-dependencies of the ecological networks in which disease spreads can be better accounted for within models, then the key epidemiological parameters of interest to this thesis—diagnostic test sensitivity, diagnostic test specificity and disease prevalence—can theoretically be better inferred.

Chapter 1 discusses the problems faced by ecologists when estimating diagnostic test sensitivity, diagnostic test specificity and disease prevalence in diseased wildlife populations, and outlines the requisite tools for addressing these problems. Chapter 2 then provides a literature review on the importance of modelling wildlife disease across ecological scales using a Bayesian framework. Chapter 3 describes and justifies the modelling architecture—i.e., the BLCMs and their enabling functions written in R code—that underpins the remainder of this thesis.

Chapters 4–8 then advance this modelling architecture throughout five empirical chapters, “stress testing” the architecture’s ability to infer diagnostic test

sensitivity, diagnostic test specificity and disease prevalence via the following four analytical approaches:

(a) **Model validation** (Chapter 4) i.e. assessing model fit in terms of the data available.

(b) **Uncertainty analyses**—the quantification of the confidence in BLCM inferences and a type of model validation—via the interrogation of two statistical artefacts, i.e. trends explainable by statistics rather than ecology. These trends are the relationships between the accuracies of diagnostic test sensitivity, diagnostic test specificity and disease prevalence (Chapter 5), and the observed effects on the robustness of inferences of diagnostic test sensitivity, diagnostic test specificity and disease prevalence at the “extreme” limits of their possible values (Chapter 6).

(c) **Sensitivity analyses** (Chapter 7)—an examination of whether BLCMs are sufficiently robust to new information, or changes in model assumptions.

(d) **Time decomposition** (Chapter 8)—a statistical procedure enabling the BLCM to infer the diagnostic test sensitivities, diagnostic test specificities and disease prevalence of a real-world dataset—from the Woodchester Park study—through time.

Finally, Chapter 9 draws together the contributions made, and outlines what they mean for ecologists wishing to use BLCMs for their own research.

Chapter 2

2. A perspective on the Bayesian modelling of wildlife disease across ecological systems

Introduction

Statistical methods have long been used to better understand disease data. In 1854, for instance, John Snow discovered the source of a cholera outbreak by cluster-mapping infections across Soho, London (Snow, 1856). Similarly, inferring disease prevalence is a critical tool for quantifying the number of infected individuals within a sample, group, or population. The challenge involved is clear: in ante-mortem animal studies, the value of disease prevalence can only be determined using statistical methods that can operate across ecological levels, and particularly at the levels of individuals and populations.

This chapter postulates the argument that the accurate estimation of disease prevalence involving wildlife host-pathogen systems calls for a better representation of ecological hierarchy, i.e. the multiple ecological levels involved; and a better understanding of the statistical hierarchy involved, i.e. the sources of bias at each level of the ecological hierarchy (Farnsworth *et al.*, 2005; McClintock *et al.*, 2010; Lachish and Murray, 2018). While this chapter's recommendations are unlikely to surprise statistical epidemiologists, this double hierarchy has not yet been clearly described for a single system. Research that reflects this double hierarchy is what is termed in this thesis as a "whole-system

approach", with the end goal being host-pathogen systems that can be described by "whole-system models".

If the latent parameters of host-pathogen systems are influenced by parameters belonging to other ecological levels, then statistical models used to predict population-level parameters may be more reliable if data is available at both the individual and population levels (Tompkins *et al.*, 2011). Given such data, multi-level regression models can be useful prediction tools, as they have the potential to uncover more sources of bias within a host-pathogen system than when a single-level regression model is used, since the applicable regression coefficients can vary by discrete groups, which in this example are ecological levels (Gelman, 2006).

Multi-level modelling is not a new technique to wildlife disease research (Cross *et al.*, 2010; Manzoli *et al.*, 2013; Raghavan *et al.*, 2016) but is undeniably a technique that has been less commonly applied to this field when used inside a Bayesian framework, and even less so when applied to more than one ecological level. The penultimate chapter of this thesis demonstrates the specific application of Bayesian multi-level modelling to wildlife disease research using real-world data on bTB infections in a badger reservoir population. Given this focus, and the fact that bTB host-pathogen systems are high-profile, and dominate the wildlife disease literature, the final section of this present chapter presents a case study on how Bayesian inference has already been applied to this specific body of work.

Eventually, it is perfectly possible that ecologists may wish to use Bayesian multi-level or "hierarchical" models to fully realise and explain (Feki-Sahnoun *et al.*, 2018) the latent relationships and interacting factors that make up disease

systems (Ting and Shaolin, 2008). To facilitate this, there is consequently a need to explain how these models may benefit ecologists, and this is the principal aim of this chapter.

That said, this chapter does not aim to reignite the debate on whether frequentist or Bayesian approaches are better in the round, or more appropriate in the context of disease modelling, or to indicate that alternative ways to infer data such as machine learning (Fountain-Jones *et al.*, 2019) aren't useful. Clearly, the Bayesian approach is intrinsically different from mechanistic methods of disease analysis that apply Ordinary Differential Equations—or frequentist methods such as variations of the Susceptible-Infected-Recovered models—because they offer a different type of flexibility to infer the noise of ecological processes (Zhuang *et al.*, 2013). And since the definition of likelihood functions for observed data is usually possible in disease analyses, the Bayesian approach can usefully be used to maximise the information known to an ecologist.

One immediate challenge encountered in this chapter was the difficulty in sorting studies by the specific type of model that they apply. This hurdle is reflective of the "terminological confusions" noted at several points in this thesis—including within the definitions section at the start of this thesis—regarding, for example, the terms model validation, or sensitivity analysis. For this chapter, it was found that many hierarchical models are not referred to as "hierarchical"; and it was suspected that many systems-level approaches will not use the word system. So, an important caveat is that literature selected from the various searches employed were only included for consideration in this chapter if the type of model used could be clearly identified. It was also assumed that the frequencies of published papers belonging to any well-

recognised type of model *should* still be revealed from the combinations of word searches described. For example, it was assumed that using the search term "state-space model" would reveal most publications that use a state-space model.

Finally, this chapter has benefited from comments received from two editors of the Wiley journal *Ecology Letters*, and a historic version is hosted on Authorea Preprints (DOI: 10.22541/au.164621773.37508959/v1). My PhD supervisor Professor Dave Hodgson also contributed to this preprint version by assisting with the development of ideas and providing edits.

In this chapter the term "scale" refers to the single level, or multiple levels, of a hierarchical system that a study may concern.

Why look at wildlife disease on a systems scale?

Host-pathogen systems are characterised by the complex networks of interactions (Sander, Wootton and Allesina, 2017) between an infectious agent and its host species (Forst, 2010). Examples of such host-pathogen systems include bovine tuberculosis (Böhm, Hutchings and White, 2009), avian influenza A (H5N1) or "bird flu" (Webster *et al.*, 2005), Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) (Li *et al.*, 2005) and the (yet-unidentified) wildlife host of SARS-CoV-2—the causative agent of COVID-19.

Given the severity and economic impact of these and similar diseases, improving the capability of statistical models to describe entire ecological systems is an important and desired advance in disease ecology, particularly since understanding wildlife health is critical to its management (Calenge *et al.*, 2021) and the risks posed to human health. Importantly, the ability to model

wildlife disease on a systems scale is likely to unlock more information about how human-wildlife interactions drive host-pathogen systems, supporting the One Health concept (Bordier *et al.*, 2020) of decompartmentalising human, animal, and ecosystem health. Emerging methods capable of uncovering missing links within entire host-pathogen systems have been termed “zoonotic risk technology” (Carlson *et al.*, 2021), a term also applicable to methods available to achieve a whole-system approach.

This chapter specifically proposes that a whole-system approach to studying wildlife disease is essential for a proper understanding of disease ecology, because most mechanisms of wildlife disease transmission co-vary with other ecological parameters, are not fully understood, and are impossible to measure directly. For example, substantial gaps in our knowledge of Chronic Wasting Disease (CWD) ecology have been identified, such as its unidentified reservoir species, and the biogeography of CWD transmission (Escobar *et al.*, 2020): a Bayesian whole-system approach could usefully fill such gaps by linking infection processes across ecological scales using prior information.

In comparison to models of human epidemiology, models of wildlife disease are usually created in a data-poor environment. As Chapter 1 emphasised, animals tend to be hard to track and or trap, and infection states can be hard to infer due to imperfect diagnostic tests. Accordingly, a frequent purpose of wildlife disease models is to infer latent parameters or associations between factors that make up a disease system, to better understand how disease spreads. Importantly, this chapter is restricted to statistical rather than algebraic models of disease processes because the parameters that may describe these processes vary stochastically, as well as in time and space; and are rarely, if ever, known (Zhuang *et al.*, 2013).

Studying disease at a systems scale involves recognising the hierarchy of interacting levels over which disease dynamics persist—termed in this chapter as the ecological hierarchy (Figure 2-1)—and requires the adoption of hierarchical models in the broadest sense, i.e. models that are capable of investigating multiple levels of organisation. The classical definition of hierarchical models is focused on in this chapter, and this starts from the premise that a hierarchy of scale exists across ecological systems (King, 1997; Wu and David, 2002; Allen and Starr, 2017) that can be used to explore the nested relationships between differently scaled variables through sub-models, which then link together to form a full model.

Wildlife disease studies using hierarchical models have made exciting discoveries. For example, major progress in eliminating the *Sarcoptes scabiei* mite from bare-nosed wombat populations was facilitated by considering the disease statuses of wombat burrows at the metapopulation level as well as of the individual wombats; and consequently, both the burrows *and* the wombats were modelled as hosts (Martin *et al.*, 2019). Hierarchical modelling has also enabled a database of bat hibernation roost surveys to be analysed through time, space, and across five species, to determine the latent disease severity of *Pseudogymnoascus destructans* infections—the causative agent of white-nose syndrome—at species and regional scales in North America (Cheng *et al.*, 2021).

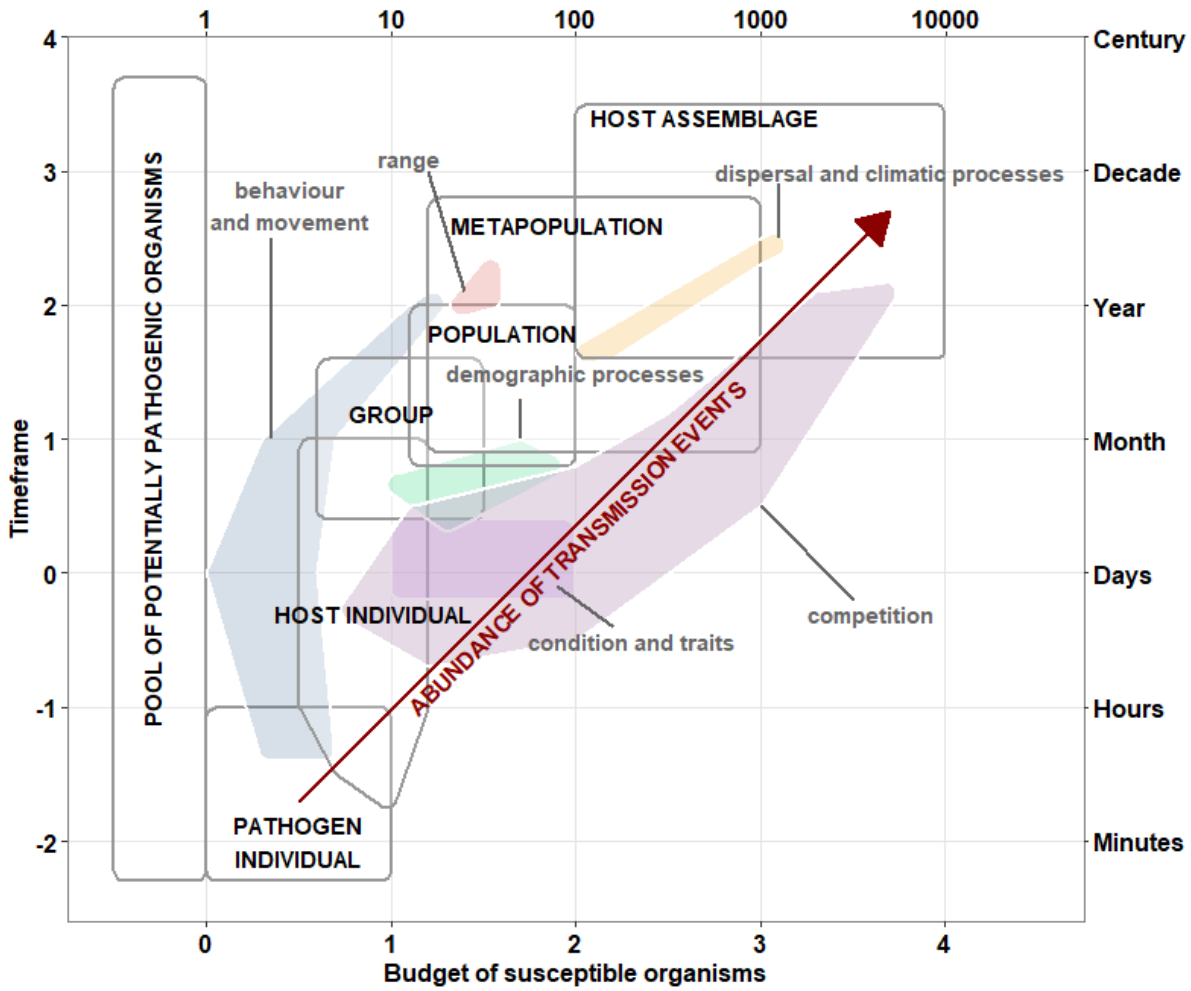

Figure 2-1: A stommel diagram illustrating the concept of an “ecological hierarchy” on a logarithm base 10 grid.

Within Figure 2-1, hierarchical ecological information, defined by the grey rectangular outlines—spanning from a single pathogen to a host assemblage—may contribute to a whole-system model that describes disease flow within a host-pathogen system. As the time since an initial infection increases on the y-axis, transmission events may also increase, as well as the budget of susceptible organisms, represented on the x-axis, that could be exposed to infection. The diagram represents both the structures and processes acting throughout an ecosystem, on a scale that could be parameterised by dynamics—such as host or population behaviour(s) or group demographic processes—which describe expected fluctuations around ecological equilibria

(Cushman, 2010). Coloured polygons therefore propose example locations for these latent dynamics in the context of a whole-system model.

The statistical modelling of ecological hierarchies.

In hierarchical models, the individual level is defined as the smallest measurable unit of that system (Kéry and Schaub, 2011): for example, genes, such as those coding for disease susceptibility, could be viewed as the individual unit; alternatively, detailed models might consider individual pathogens; coarser models might start with the individual host and model within-host infection processes in the abstract. When Bayesian methods are used to infer latent and unmeasurable states, truly binary or categorical states such as dead or alive; infected or uninfected; can be inferred as probabilities, which better reflect their lack of direct measurement (Buzdugan *et al.*, 2017).

Parameters and processes of interest to ecologists act at the individual level up to higher levels of the ecological hierarchy. Within this hierarchy pathogens are clustered into biological and environmental reservoirs; hosts are structured socially and into (meta)populations; host species are members of assemblages; and system dynamics play out through time and space. At each level, different sets of predictors influence outcomes: host condition might relate to pathogen load; social context might influence transmission; host age might influence mortality hazard and susceptibility to infection; weather might influence population level epidemiology; wildlife management might affect the host assemblage; climatic and anthropogenic change might influence the prevalence of disease and the risk of epidemics or host-shifts. In short, a system—with linkages to and dependencies on pathogens—is at work.

Expanding on Figure 2-1, infection status or “being infected” is an example of a host-specific latent variable, which may inform processes at all levels of an ecological model. Likewise, ageing is an example of a demographic process that could reasonably influence multiple levels of an ecological hierarchy (Jones and Vaupel, 2017), and age itself could be considered a feature of a pathogen, a pathogen reservoir, an individual host, an individual infection, a social group of hosts or even whole host populations.

This chapter also draws attention to a second type of hierarchy typical of data associated with wildlife: the statistical hierarchy (Table 2-1). This hierarchy maps onto the ecological hierarchy, recognising that different model parameters, and different predictor variables, are relevant to each stratum. For example, host condition is likely to relate to the individual host’s susceptibility to infection, its propensity to suffer disease, and its role in transmitting infection to other hosts. Other predictors vary at higher levels of the ecological hierarchy, for example weather conditions varying weekly, seasonal conditions varying annually, and climate varying over longer timescales; for example, density-dependent transmission varying at the scale of social groups, sub-populations or whole populations.

The statistical hierarchy must avoid problems of pseudoreplication—i.e., the incorrect assumption that all replicates are independent (Lazic *et al.*, 2020)—by recognising independent survey units on each stratum of the hierarchy or by accounting for stratum-specific spatial, temporal, genetic or social nonindependence. Another important feature of the statistical hierarchy is any mismatch between the parameters that researchers wish to infer, and the data that they are able to collect. Often the hidden network of latent variables that researchers wish to infer, such as being infected or being dead, can only be

measured by proxy, such as through live trapping, or by analysing the results of imperfect diagnostic tests. The statistical hierarchy therefore includes a state-space representation of many parameters (Table 2-1).

Within Table 2-1, a broad statistical method used to infer each respective latent parameter is also suggested. As the strata increase with respect to the number of organisms present, the statistical methods that may be used to infer common latent parameters become broader. Consequently, a large variety of spatio-temporal methods may be used to infer latent parameters above the host level using proxy data. As the strata represent more complex latent data, the need for Bayesian state-space models increases, and the number of commonly reported statistical methodologies to infer common latent parameters decrease. A whole-system model would be a spatio-temporal modelling technique able to infer latent parameters within any stratum of an epidemiological system, which may themselves be dependent on latent parameters in other strata.

Table 2-1: Examples of latent parameters that wildlife disease researchers may wish to infer and their proxy measures that may be chosen given each stratum of a typical wildlife disease system. Statistical methods that may be used to infer the latent parameters are suggested. The need to infer latent parameters using state-space approaches increases as the number of organisms belonging to the ecological layer at which the latent parameter is being inferred increases.

Layer of ecological network	**Latent parameter to infer**	**Proxy measure in host**	**Statistical method to infer latent parameter**
Pathogen	Virulence	Observation of symptoms (physical/behavioural) in host	Logistic regression of pathogen versus host survival
	Basic Reproduction Number	Serosurveys or behavioural surveys in host	Logistic regression of proxy measure versus time
	Presence/absence	Laboratory culture of host serosurvey data	Latent Class Model to account for imperfect testing
Host	Location	Telemetry	Home range analyses using kernel density estimation
	Infection status	Diagnostic test outcomes	Multi-event analyses
	Alive/dead	Capture-mark-recapture	Dynamic occupancy modelling
Group	Membership	Social co-dynamics	Spatio-temporal analyses

	Social Contact Network	Social co-dynamics	Spatio-temporal analyses
Population	Temporal/spatial abundance	Camera trapping (presence/absence)	Spatio-temporal analyses
	Population size	Resource abundance	Spatio-temporal analyses
	Disease prevalence	Diagnostic test outcomes	Latent Class Models to account for imperfect detections
Metapopulation	Connectivity	Topography	Spatio-temporal analyses
	Colonisation	Telemetry	Spatio-temporal analyses
	Gene flow	Capture-mark-recapture	Spatio-temporal analyses
Assemblage	Species interactions	Prey kills in wild	Spatio-temporal analyses
	Species distribution	Camera trapping (presence/absence)	Spatio-temporal analyses
	Species richness	Stable isotopes	Spatio-temporal analyses

For ecological systems in which the definition of likelihood functions for observed data is possible, the Bayesian approach is both rigorous (O'Hare *et al.*, 2014), and also capable of parameterising the "double-hierarchy" of whole-system disease models, i.e. the ecological levels that must be represented by different statistical levels. Where complexities prevent the definition of likelihood functions, analysts might look to Approximate Bayesian Computation (Benavides *et al.*, 2017) or machine learning techniques (Pandit and A. Han, 2020) to guide understanding of the system.

Throughout this chapter, attention is drawn to two variables "disease status" and "mortality status" because these variables are of primary interest to wildlife epidemiologists. Despite this concentrated focus, a whole-system model is referred to as one that can describe as many aspects of an ecological network as possible, across a hierarchy of ecological and statistical scales (Figure 2-1, Table 2-1).

The remainder of this chapter is structured as follows: first, a consideration of why Bayesian inference should be used to model wildlife disease; second, a review of the application of Bayesian methods to wildlife epidemiology; third, a consideration of the importance of considering latent variables and individual heterogeneities for a whole-system model of wildlife disease; and finally, a demonstration of how Bayesian modelling has informed research into wildlife reservoirs of bovine tuberculosis (bTB). The literature surveyed suggests that Bayesian approaches to the modelling of wildlife disease are (a) relatively scarce, and (b) tend to infer only limited subsets of a whole-system model.

Why use Bayesian inference to model wildlife disease?

A Bayesian model can be loosely defined as any model deriving its inference from a posterior probability distribution, acquired from a prior probability distribution and its associated likelihoods, using Bayes' theorem and *any* available data or prior knowledge (Pearl, 1988). Powerfully, a disease ecologist can consequently combine all known ecological information relating to a host-pathogen system, drawn from disparate sources, into a single, integrated model (Dunson, 2001). Bayesian models have already influenced our understanding of disease risks from invasive species (Lohr *et al.*, 2017), the potential for disease transmission (Lau *et al.*, 2017), and vulnerabilities within livestock systems to foot and mouth disease (Manyweathers *et al.*, 2020).

The primary goal of statistical epidemiology is to understand parameters most relevant to the understanding and management of epidemics, particularly infection prevalence, severity and spread. A current focus of disease ecologists is to understand and differentiate among interactions and relationships within a complex host-pathogen system (Milns, Beale and Anne Smith, 2010), despite the multiple complications this entails. For instance, when modelling disease systems, network complexity is known to add to "network fragility" (Milns, Beale and Anne Smith, 2010)—a somewhat vague graph theory term that in essence means "less stable"—largely due to increasingly unpredictable ecological responses to perturbations (Montoya, Pimm and Solé, 2006). Examples include social perturbation, i.e., individual dispersal in response to management interference, as observed during badger culling (Woodroffe *et al.*, 2006; Carter *et al.*, 2007) and wider anthropogenic perturbation from the threats to wildlife from human activity. A further complication is that the causative pathogens

themselves may cause potentially bi-directional—i.e., host to pathogen and pathogen to host—behavioural alterations affecting transmission (Weber *et al.*, 2013; Ezenwa *et al.*, 2016; McDonald, Robertson and Silk, 2018). Constructing a realistic host-pathogen network, including these fine-scale interactions such as individual behaviours, remains a key challenge to the development of a whole-system model.

The many benefits of Bayesian approaches to inference, and the Monte-Carlo Markov Chain algorithms usually used to implement them, are well documented (Kéry and Schaub, 2011; Hooten, Hobbs and Ellison, 2015). However, in the context of wildlife-disease modelling, the benefits of adopting a Bayesian approach include, but are not limited to:

1. The ability to use prior information when available.
2. The flexibility to describe a hierarchy of states, processes and their noise in a single model.
3. A clear approach to inferring latent variables and parameters.
4. The ability to combine across multiple sources of data and multiple statistical processes.
5. And, the flexibility to work with a wider-than-usual range of likelihood functions (van de Schoot *et al.*, 2021) such as computationally expensive likelihood functions that are slow to evaluate.

In contrast, the costs of adopting Bayesian methodologies include:

1. The learning of new statistical concepts and software.
2. The dropping of ingrained allegiances to tests of significance or information criteria (Halsey, 2019).

3. The computational expense of running long, iterative chains of likelihood calculations.
4. The lack of consensus on how to judge the importance of rival models (Harrison *et al.*, 2018).

Recent advances in computation, methodology, education, and software are already helping to minimise these apparent costs.

Bayesian inference is particularly useful to disease ecologists because field data from real-world, diseased, or healthy wildlife populations is sparse but can often be supplemented by expert prior knowledge. Therefore, a Bayesian modeller has the flexibility to combine both quantitative and qualitative data (Wijesiri *et al.*, 2018). Further, Bayesian hierarchical techniques can capture the intricacies of level, scale and hierarchy within ecosystems by simultaneously accounting for their uncertainties and handling a hierarchy of predictors as fixed or random effects (Wikle, 2003). Consequently, the uncertainty in latent epidemiological variables (Drewe *et al.*, 2010) such as an individual's infection status, can be both accounted for, and inferred.

Bayesian Inference for Wildlife Disease: Examples

To date, Bayesian hierarchical methods have been applied only sparingly to wildlife disease problems. For example, only eight examples (Table 2-2) of Bayesian hierarchical methods can be found using a Web of Science search—dated March 2022—given combinations of the terms: "Bayesian"; "hierarchical"; "model"; "wildlife"; "animal"; "disease"; "infection"; "system". Naturally, such a paucity of citations will not adequately embrace every paper that uses Bayesian hierarchical methods to model disease, but certainly serves to characterise its limited application. What is more, Bayesian hierarchical models also seem to

have been applied to the field of wildlife disease epidemics in something of a “scattergun” manner across disease systems, levels of the ecological hierarchy, or in terms of the process that is being inferred. They have rarely been used to explore individual, group, and population hierarchies within the same model.

Table 2-2: Examples of wild host-pathogen systems that have been investigated using Bayesian hierarchical modelling where S = spatial, and T = temporal.

Host-pathogen system	**Scale(s) of study**		**Key parameters investigated**	**Key ecological finding**	**Reference**
	T	**S**			
bTB in European badgers	✓		Survival; recruitment	Life history and recruitment characteristics of badgers ensure that the bTB reservoir is maintained	(McDonald *et al.*, 2016)
Devil Facial Tumour Disease (DFTD) in Tasmanian devils (*Sarcophilus harrisii*)	✓		Survival; fecundity	DFTD affects the most reproductively valuable devils	(Wells *et al.*, 2017)
CWD in white-tailed deer (*Odocoileus virginianus*)		✓	Likelihood of infection	How CWD may be spatially distributed	(Evans *et al.*, 2016)

CWD in North American elk (*Cervus elaphus nelsoni*)	✓	✓	Prevalence; allele frequency	The relationship between CWD prevalence, and the PRNP 12L allele, which may extend the latency of CWD in North American elk	(Monello *et al.*, 2017)
Influenza A in captive mallard (*Anas platyrhynchos*) and lesser snow geese (*Chen caerulescens*); *Yersinia pestis* in coyotes	✓		Time since infection; force-of-infection	A method to estimate force-of-infection from individual antibody data	(Pepin *et al.*, 2017)
Brucellosis (*Brucella abortus*) in wild elk and livestock herds	✓	✓	Probability a region has brucellosis infections in its livestock	The spillover of brucellosuis from elk to livestock may happen more in regions where unfed elk are contracting the disease from fed elk	(Brennan *et al.*, 2017)

CWD in mule deer (*Odocoileus hemionus*)	✓	Prevalence	The impacts of CWD on population growth rate, and covariates which moderate disease dynamics	(Geremia *et al.*, 2015)
Five pathogens (porcine reproductive and respiratory syndrome virus, pseudorabies virus, Influenza A virus, Hepatitis E virus, and *Brucella spp.*) infecting wild pig (*Sus scrofa*)	✓	Seroprevalence	Demographics were not good at predicting seroprevalence. It is important to account for detection error when estimating the sensitivity and specificity of a diagnostic test.	(Tabak, Pedersen and Miller, 2019)

The information in Table 2-2 provides corroboration that Bayesian hierarchical models are useful for generalising large within-population and or landscape-scale processes, and that they are broadly applicable across disease systems. Despite this, studies across broad ecological levels and scales are rare. The information in Table 2-2 also suggests that studies purely investigating disease spatially, or spatially and temporally, are less common than those that have a temporal investigation alone. Equally, as is demonstrated by the bTB case study in the final section of this chapter, Bayesian hierarchical analyses of evolving longitudinal datasets are also rare but are likely integral to the discovery of fine-scale ecological interactions pertinent to understanding disease processes.

Modelling latent variables is essential to the whole-system approach.

Bayesian state-space models are a form of Bayesian hierarchical model that allow Bayesian networks to easily distinguish dynamic biological processes such as changes through time (Beyer *et al.*, 2013; Auger-Méthé *et al.*, 2016) from unavoidable errors due to the imperfect detection of disease, host survival or transmission events. This chapter's review of examples of Bayesian inference in wildlife disease research reveals a suite of latent variables that can be inferred from the capture and diagnosis data that are typically collected (Table 2-3). Specifically, state-space models achieve this inference by accounting for whether a parameter is unobserved or observed, as well as any associated sampling error (Royle and Young, 2008). This means that Bayesian state-space models are especially good at, for example, teasing apart

demographic stochasticity and sampling error (Newman, 1998; Patterson *et al.*, 2008).

Table 2-3: Examples of inferred wildlife disease parameters investigated using Bayesian state-space models.

Host-pathogen system	**The Bayesian state-space model**			**Inferred disease parameters**	**General ecological finding**	**Reference**
	Observed time series		**Unobserved State**			
	Serology	**Other data**				
Swine Influenza in domesticated Chinese swine	✓	Virological	Probability of exposure	Force-of-infection; risk of exposure	Early life exposure to Influenza in swine populations is increasing	(Strelioff *et al.*, 2013)*
CDV in lions (*Panthera leo*) and domesticated dogs (*Canis lupus familiaris*)	✓	regional vaccination coverage	Probability of infection	Seroprevalence; impact of vaccination	CDV infection in lions is becoming more frequent	(Viana *et al.*, 2015)†

Morogoro virus in multimammate mice (*Mastomys natalensis*)	✓	weight; infection patterns	Natural infection patterns	Time of infection	No evidence suggesting that natural and laboratory infection patterns are not similar	(Mariën *et al.*, 2017)‡
Hantavirus in striped field mice (*Apodemus agrarius*)	✓	NA	Seasonal transmission rates	Risk of hemorrhagic fever with renal syndrome (caused by Hantavirus) in human populations	Hantavirus spillover is driven by seasonality and dynamics of Hantavirus in rodent reservoir populations	(Tian *et al.*, 2017)‡
CDV in grizzly bears (*Ursus arctos*) and wolves (*Canis lupus*)	✓	NA	Timing of infection	CDV exposure in wolves and bears	How CDV dynamics vary temporally in wolves and bears	(Cross *et al.*, 2018)†
Fox (*Vulpes vulpes*) rabies	✓	NA	Demographic data, spatial data,	Transmission heterogeneity; probablility of	Information about the local transmissive	(Baker *et al.*, 2020)†

			vaccination rate	infected fox moving area; observation rate; environmental noise	processes of rabies in foxes	
Avian malaria (*Plasmodium relictum*) in Hawaiian honeycreeper species	✓	NA	Age-prevalence model, demographic data	Prevalence; intensity of infection	Patterns of prevalence, transmission, and mortality rates	(Samuel *et al.*, 2015)†

* Web of Science one-term and one-topic search; † Web of Science five-term search; ‡ ad hoc search

The inferred latent variables and ecological findings of the Bayesian state-space models presented in Table 2-3 demonstrate that disease parameters are mainly studied at the group level or population level (Osada *et al.* 2015) even though most of them used individual serology data to inform models. Table 2-3 highlights that the application of Bayesian state-space models within wildlife disease epidemiology is limited, but the search was hampered by vague or inconsistent model terminologies. Underpinning this observation are three search methods referenced within Table 2-3. The first is a Web of Science search using combinations of the terms: “Bayesian”; “state-space”; “disease”; “wildlife”, which only yielded four relevant studies. A further relevant study was found using the search topic “state-space model” when filtering by the Web of Science category “ecology”. Two additional relevant examples were found in the absence of either “state” or “space” as a keyword.

Three key observations can be drawn from the examples contained in Table 2-3. First, that observed serological data is common to all studies, presumably because most disease states in wildlife remain latent following visual surveillance. Second, it is encouraging that ecological stochasticity is modelled in the dimensions of space and time, often within the same study. And thirdly—and most importantly—the observations drawn by all the examples in Table 2-3 only regard population- or species- levels. Based on these observations, it is found that state-space models often span two levels of a hierarchy but rarely multiple latent variables. An example of this is demonstrated within previous work on the badger-bTB system (McDonald *et al.*, 2016), which used state-space models to infer the latent variable “alive” but ignored uncertainty in diagnostic test outcomes.

Table 2-3 convincingly demonstrates that serological data collected over space and time can, when combined with Bayesian state-space methods, yield powerful conclusions about high-level disease parameters. But state-space models are also an obvious tool for filling in any unknown relationships between individual disease states. For example, multi-state modelling using maximum likelihood methods revealed that epidemiological and demographic parameters vary between disease states in badgers (Graham *et al.*, 2013), yet a significant number of unknown complexities still exist within this relationship which cannot be quantified without Bayesian methods. There is a fundamental need to parameterise processes within disease models more rigorously by applying Bayesian state-space theory. Although the need for good epidemiological parameter inference has been apparent for over a decade (Simmons *et al.*, 2006; Craft *et al.*, 2008), the potential of state-space models has not yet been realised: they can help define the mutable nature of disease across *any* level of the whole-system model, inclusive of space as well as time.

Including individual heterogeneities is essential, but difficult.

Studying the spread of infection or disease among individual hosts can be challenging because single transmission events are not just impossible to observe in the wild, but also associated with a wide variety of host characteristics that are difficult to measure and monitor, such as behaviour, immunity, age, movement and crucially the interactions between infected and susceptible individuals.

Wildlife diseases are often studied using data on antibodies or general pathological observations (Mariën *et al.*, 2017). Consequently, many epidemiological state-space models are based on serology records (Gilbert *et*

al., 2013; Benavides *et al.*, 2017). Although the seropositive statuses of most wild host species are unknown (Benavides *et al.*, 2017), it is becoming increasingly important to look at the information that serology records provide, to reveal individual heterogeneities. For example, in *Eidolon helvum* fruit bats, seropositive thresholds were used to distinguish between the genetic and acquired immunities to Lagos bat virus and African henipavirus, using Bayesian mixture-models (Peel *et al.*, 2018). Here, Bayesian inference determined that immunity relied on patterns in disease transmission (Peel *et al.*, 2018), suggesting that serological data is a useful way to measure individual heterogeneities. In turn, this suggests that estimating seroprevalence is a good proxy for inferring the probability of infection in the absence of reliable testing. Yet even though the state-space models described in Table 2-3 are based on individual-level data, the inferences are usually population-level parameters, with highly generalised disease processes (Viana *et al.*, 2015), illustrating the difficulty in disaggregating the individualistic characteristics of disease processes.

A further difficultly in representing the individual state within state-space models is the complexity of the data involved. For example, ageing is a latent individual process that is difficult to understand, particularly in terms of its relationship with disease. Serological data has been directly associated with age to infer infection rate, the probability of antibody loss, and recovery rates in brucellosis-infected Elk (Benavides *et al.*, 2017). Yet to infer these parameters, the authors adopted Approximate Bayesian Computation methods due to the difficulty in writing closed form likelihood functions for the study parameters, and the associated difficulty of then implementing them within a standard MCMC algorithm (Benavides *et al.*, 2017). This is an example of where the usefulness of

Bayesian state-space modelling is currently limited in terms of its accessibility to disease ecologists. The latent process of ageing is intrinsically linked to disease via a host of known and unknown latent variables which should be accounted for in a whole-system model. Bayesian methods are a practical tool of choice for modelling complex systems, but realistically, the modelling of whole systems using the Bayesian hierarchical approaches described within this chapter will rely on stronger collaborations between statisticians, epidemiologists, and ecologists.

CASE STUDY: Use of Bayesian inference to research wildlife reservoirs of bTB

Bovine tuberculosis (bTB) infections—caused by zoonotic bacteria *Mycobacterium bovis*—are globally relevant, difficult to control, and scrutinised by disease ecologists across many host species. Research on mammals maintaining bTB reservoirs over wildlife-livestock boundaries dominate the literature, and the disease is high-profile and economically important. Yet researchers continue to find new wildlife reservoirs of *Mycobacterium bovis* (Varela-Castro *et al.*, 2021), any of which could influence the transmission and spread of disease among livestock. Bayesian approaches could help bridge the data gaps between rarely studied and well-studied bTB hosts by enabling information on host ecology from non-disease studies to inform future epidemiological models.

In badger-bTB research, a better understanding of pathogen transmission within and among badger reservoirs, as well as between badgers and cattle, or other non-reservoir host species, is required. Like all disease systems, the understanding of the badger-bTB system is constantly shifting with new pieces

of information, which can act to better inform priors with expert knowledge and improve our beliefs. For example, the rapid serological Dual-Path Platform VetTB test has recently been validated for bTB testing in badgers (Arnold *et al.*, 2021) and a badger behaviour called super-ranging has been detected, which is potentially responsible for long-distance bTB transmissions (Gaughran *et al.*, 2018). Consequently, these specific pieces of information *could* help provide updated estimates of disease transmission and disease progression within a badger-bTB system (McDonald *et al.*, 2016).

Although the number of "how-to" papers describing the power of Bayesian inference in the context of wildlife epidemiology is increasing (Enright and O'Hare, 2017; Conn *et al.*, 2018), research incorporating Bayesian modelling strategies specifically focused on the ante-mortem badger-bTB system is limited, with the result that its benefits to wildlife disease research are not widely appreciated. Six known studies of primary research (Drewe *et al.*, 2010; McDonald *et al.*, 2014, 2016; McDonald, Robertson and Silk, 2018; Crispell *et al.*, 2019; Hudson *et al.*, 2019) that used Bayesian methods to explore badger-bTB transmissions on a "landscape-scale" were considered. All six studies defined landscape-scale as the geographical extent of Woodchester Park, Gloucestershire, UK, a 7km^2 region where the capture-mark-recapture data common to all six studies was collected.

In South Island, New Zealand—where brushtail possum (*Trichosurus vulpecula*) were speculated to be the keystone reservoir species of bTB for *circa* three decades (*Trichosurus vulpecula*) (Morris and Pfeiffer, 1995)—recent Bayesian research (Crispell *et al.*, 2017) has provided confirmation that its possum population is responsible for South Island's bTB maintenance; rather than its cattle population. In the UK, although it has been confirmed via Bayesian

Integrated Population Models *why* Woodchester Park badgers are an efficient bTB reservoir (McDonald *et al.*, 2016), the directionality of bTB transmissions between badgers and cattle remains debated, and it is suspected that badgers are responsible for roughly half of bTB infections in cattle within high cattle-bTB incidence areas (Donnelly and Nouvellet, 2013). Another analysis concluded that badger to cattle transmissions were ~10.4 times more frequent than *vice versa* (Crispell *et al.*, 2019).

A whole-system model of bTB systems, capable of linking information throughout an ecological hierarchy, is required, and an example of what this model may look like is presented in Figure 2-1. A particular limitation in the development of such a model is the ability to incorporate individual badger heterogeneities: individual traits are often neglected in disease models since detailed longitudinal datasets of individuals within diseased populations—such as the Woodchester Park dataset—are rare.

Within the badger-bTB system, heterogeneities among badgers (McDonald, Robertson and Silk, 2018)—such as gender, inbreeding, disease, social group and age (Benton *et al.*, 2018)—act as proxies for infectiousness or “risk” (VanderWaal and Ezenwa, 2016), and are thought to drive fine-scale bTB dynamics. Fundamentally, an understanding of fine-scale disease processes in combination with Bayesian methodologies arms ecologists with the ability to parameterise previously unobservable processes, such as actuarial senescence (Hudson *et al.*, 2019), gender-differences in susceptibility to bTB (McDonald *et al.*, 2014) and on the diagnostic accuracies of badger-bTB tests (Drewe *et al.*, 2010): information which improves our capability to model badger heterogeneities in the future.

The idea of achieving a better understanding of the bTB system in wildlife hosts using a whole-system approach is not a new one. Analogous to the whole-system model posited within this chapter, Silk *et al.*, 2017 proposed the need for a novel modelling framework, and McDonald, Robertson and Silk, 2018 recommended a comprehensive epidemiological model. In addition, White, Forester and Craft, 2017 suggested that combining contact networks with Bayesian inference is the future direction for understanding wildlife epidemics. The inclusion of a hierarchy of scale within whole-system ecological models in general has been recommended by several authors (Tonnang *et al.*, 2017; Fountain-Jones *et al.*, 2018).

Conclusion

Modelling host-pathogen systems can be considered a “wicked problem” (Rittel and Webber, 1973): its success is dependent on multidisciplinary thinking (Benjamin-Fink and Reilly, 2017) between statisticians, epidemiologists, and ecologists; and there is a balance between accepting over-simplified solutions and being overwhelmed by overly complex ones (Defries and Nagendra, 2017). Moreover, any solution involves balancing conflicting and fluid temporal and spatial ecological scales (Waltner-Toews, 2017). In addition to space and time, the environmental processes that describe host movement—such as climate or seasons—are often disregarded, yet essential, dimensions required to model disease systems (Merkle *et al.*, 2018).

A deeper forensic approach (Benton *et al.*, 2018) is required to better understand and parameterise complex host-pathogen systems, and the Bayesian toolkit provides a good starting place for this. Overall, future studies of host-pathogen systems require a better representation of scale, which needs to

be examined in terms of applying suitable Bayesian methods (statistical hierarchy) and by paying attention to the complexity of the system that is being analysed (ecological hierarchy). The connection between these different scales has rarely been studied within ecological systems, and has never been completed for a single system, yet is essential to providing a whole-system model.

This survey of the wildlife disease literature demonstrates that the current application of Bayesian networks to solving wildlife disease problems is limited: in particular, there is a paucity of hierarchical analyses that infer truly latent parameters or individual heterogeneities across ecological scales. While Bayesian methods are now being used in several wildlife disease systems, they are usually only used to tackle standard hypotheses at a single level of the ecological hierarchy or, at most, span two levels of the ecological or statistical hierarchies.

By developing Bayesian hierarchical modelling methods and integrating them with real-world empirical data that is not exclusively serological, the potential exists for ecologists to create whole-system models that can provide unique insights into the epidemiology of wildlife disease networks. The first step towards the whole-system model is to develop a Bayesian hierarchical model that spans the state-space nature of each level of the host-pathogen ecological hierarchy.

With the complexities of the Bayesian modelling of wildlife disease across ecological systems considered, this thesis turns to presenting the general modelling architecture underpinning all empirical chapters.

Chapter 3

3. Generalised methodologies for generating diagnostic test data, and parameterising and calibrating Bayesian Latent Class Models.

From this point, the true parameters diagnostic test sensitivity, diagnostic test specificity, and disease prevalence are abbreviated in-text to Se, Sp and P respectively due to their frequency of use; with the abbreviations Sehat, Sphat and Phat indicating where an inferred value is being referred to.

Introduction

Conventionally, the management of disease—including newly emerged diseases and zoonotics that have crossed geographic or species boundaries—rely on the accurate estimation of epidemiological parameters at the ante-mortem stage (DiRenzo *et al.*, 2018). Since gold standard reference tests are a rarity for wildlife diseases, metrics describing disease in wildlife systems are largely reliant on statistical alternatives to such tests, particularly Latent Class Models. These alternatives, however, pose significant statistical challenges in respect of their proper application, and meeting these challenges is a prerequisite to accurate and precise inferences of the epidemiological parameters of interest.

With this in mind, Chapter 3 outlines the generalised structure of the BLCMs, and the associated modelling architecture employed in the remainder of this thesis. As such, this chapter does not aim to be a background text on BLCMs;

for this, already-published reviews—such as Wang, Lin and Nelson, 2020 and Li *et al.*, 2018—are more appropriate.

With the general modelling architecture duly outlined, each subsequent technical chapter—Chapters 4 to 8—simply describes the specific modifications to this generalised modelling structure, made in order to allow specific investigations into the particular research questions that they address.

An introduction to model power

In general terms, a model's power is its ability to find a signal when a signal exists, and this is usually conditional on the sample size available, since the standard error of the parameter being estimated is dictated by sample size (Gelman and Hill, 2006; Gelman and Carlin, 2014). Power analyses are traditionally associated with frequentist studies intending to determine the statistical significance of a signal given a null hypothesis (Gelman, Meng and Stern, 1996).

For Bayesian studies, support exists (Cumming, 2014; Gelman and Carlin, 2014; Kruschke and Liddell, 2018), for a "*shift of emphasis away from null hypothesis significance*", and instead the emphasis moves towards analyses that consider the magnitudes and uncertainties of error structures, and therefore the credibility of inferences (Kruschke and Liddell, 2018). This shift of emphasis has been termed "The New Statistics" (Cumming, 2014). As part of this thinking, a Bayesian New Statistic termed Bayesian Generalised Power has been set out as an alternative measure of model power for Bayesians (Kruschke and Liddell, 2018). For simulation analyses, Bayesian Generalised Power is "*the proportion of times that a goal is achieved*", where the "*goal*" is simply an *a priori* assumption based on real or hypothetical data (Kruschke and Liddell, 2018).

A central goal of this thesis is to evaluate how accurately and precisely BCLMs can infer Se, Sp and P. And with the notion of the New Statistics in mind, this thesis redefines model power as the relative accuracies and precisions of inferences of Se, Sp and P when compared to other BLCMs. The power of a BLCM in this thesis therefore indicates the *quality* of its diagnostic abilities, i.e. the ability to discriminate between infected and uninfected individuals. Model power is therefore a measure of the usefulness of a BLCM in terms of its performance, and the trust that can be assigned to its inferences.

In essence, for this thesis, model power provides a qualitative metric which can be used to compare the performance of BLCMs under different modelling conditions. Model power is informed by the accuracies and precisions of inferred parameters across parameter spaces, and these statistics can be visualised on heatmaps to enable qualitative analyses (see Chapter 5).

Parameter imperfection and model usefulness

The performance of a BLCM is dependent on the complex interactions among inferences of the latent parameters Se, Sp and P, which for batteries of diagnostic tests, are not fully understood.

Importantly, within this thesis, a “parameter” is a latent population-level metric, which numerically describes the “truth”, and an inferred parameter is the output of a BLCM which describes a version of that truth. A parameter is both a component of a model, and a latent feature of a population which we wish to infer. Fundamental to this concept is the understanding that inferences of the parameters Se, Sp and P cannot ever achieve “perfection”. Diagnostic perfection is not a logical research ambition—and certainly not the goal of this thesis—since statistical diagnoses only exist due to the absence of gold-

standard diagnoses. Instead, a core aim of this thesis is to better understand *when* and *how* parameters are *not* perfect, as it is impossible to truly know whether a parameter is perfect—even with unlimited modelling.

Accordingly, this thesis focuses on improving model power—inclusive of the levels and sources of error that can contribute to a parameter inference—rather than model parameterisation, which is already well-established for Latent Class Models emulating diagnostic tests (Hui and Walter, 1980; Joseph, Gyorkos and Coupal, 1995; Enøe, Georgiadis and Johnson, 2000). This focus, and the Bayesian context in which it is applied, conforms to the concept of "model-dependent realism"—which is "*the idea that a physical theory…is a model…and a set of rules that connect the elements of the model to observations*" (Hawking and Mlodinow, 2010)—since the studies presented in this thesis place importance on the usefulness of models, rather than their deterministic perfection. In short, model power is the metric that this thesis uses to qualify the usefulness of a BLCM.

A note on the levels and sources of the uncertainty of posterior distributions.

The following list describes the key sources of what this thesis terms the error (Equation 16), bias (Equation 17), and precision (Equation 18)—here, collectively termed the uncertainty (Porta, 2016)—of posterior distributions.

1. Given a selected model and prior, the uncertainty of a single simulation could be due to the choice of initial value when setting MCMC algorithm, a lack of identifiability, or any mistakes made by the MCMC algorithm. The error of a single simulation is the difference between the truth and

the inference. The precision of a single inference can be described as a credible interval, i.e. the width of a posterior distribution.

2. Given a selected model and prior, the uncertainty of multiple replicates is reflected in the confidence that can be attributed to the posterior mean. This thesis reports uncertainty in terms of among-replicate accuracy or precision, for a given volume of parameter space, and given modelling conditions. Among-replicate measures of accuracy describe how far from the truth the posterior mean sits from the prior mean. Among-replicate measures of precision can be determined by taking the mean of the standard deviations of each simulated posterior distribution, and these metrics describe the average variation associated with inferred values of Se, Sp or P.
3. The power of a BLCM—as previously defined, a qualitative metric to compare the performance of BLCMs under different modelling conditions informed by the accuracies and precisions of inferences of Se, Sp or P across parameter space on heatmaps.
4. The power of a BLCM as a function of the prior information provided, i.e., the method or methods used to supply a BLCM with existing beliefs about a parameter.
5. Selecting the truth, which could be easier or harder to infer dependent on the precision of its prior distribution.
6. A biased sample of diagnostic test data, which is not necessarily representative of the study population.

An introduction to parameter space

The parameters Se, Sp and P can obviously take any one of an infinite number of values, which are conventionally bounded above by one and below by zero as the probability scale is most useful for discussing values of Sehat, Sphat and Phat. It can therefore be useful to consider the set of possible values that a given parameter might possess as comprising the "parameter space" for that parameter. When multiple parameters are being considered at the same time, parameter space becomes multidimensional, given that the values of individual inferred parameters Se, Sp and P might not constrain each other. The number of such individual parameters gives the "dimensionality" of the parameter space, and within this multidimensional space are all the possible parameter values that characterise a particular solution (Vaseghi, 2008).

Within this thesis the term "parameter space" therefore represents a conceptual space in which the truth must lie, and is used interchangeably, as a noun, in one of three senses:

1. A one-dimensional space encompassing the range of possible values for a single parameter, Se, Sp or P.
2. A multidimensional space in which the true combination of parameters—or the truth—must lie.
3. The space explored by an MCMC algorithm (Kosmala *et al.*, 2016; Hu, Gonzales and Gubbins, 2017; Vehtari *et al.*, 2020; Ragonnet-Cronin *et al.*, 2021) that is defined by prior distributions. MCMC algorithms investigate parameter spaces while working to infer the posterior distribution of credible inference. The truth, which is usually fixed for simulation studies in order to provide controlled study

environments—or alternatively that is unknown when observed data is used—exists in MCMC parameter space.

When we as ecologists “simulate” diagnoses, we inevitably set a truth belonging to the parameters concerned. Consequently, this thesis investigates the hypothesis that model power depends on the position of the truth within parameter space. The remainder of this chapter outlines, in general terms, the overarching framework and rationale used to test this hypothesis.

The statistical challenge

Host detection is conventionally regarded as imperfect through a reliance on imperfect capture-mark-recapture studies, and the unpredictability and randomness to which they are subject. There is also a deeper layer of uncertainty to consider: the imperfect detection of pathogens within a sample of captured, live hosts (Kellner and Swihart, 2014).

As already alluded to, the data underpinning BLCMs is impacted by two stochastic processes that influence the uncertainty of diagnoses: first, the inability to trap an entire population of animals, and second, the inability to therefore test an entire population of animals for an infection. In practice, when undertaking theoretical studies—that may have the purpose of supplementing or validating real-world studies—most ecologists and researchers side-step these considerations and instead employ deterministically-calculated test data (Clark, 2005) that obey mathematical equations (Pool, 1989). Deterministically-calculated test data is generally easier to understand, particularly since the test outcomes are exactly predictable.

The limitations of deterministic approaches include the difficulty of using multiple data sets; the assumption that the process behind the parameters is

known; the inability to fully model existing *a priori* understanding; and the difficulty in integrating many layers of complex interactions (Clark and Gelfand, 2006). Since the outcomes of deterministic approaches are entirely predictable, or “idealised” (Sharkey, 2008), they can be thought of as having an ambiguous “conceptual status” (Gelman *et al.*, 2010) simply because models are never perfect. Ultimately, when modelling deterministically, if any information is incomplete, then predictions made from the governing equations will be imperfect (Hastings *et al.*, 1993), and the resulting uncertainty will be difficult to retrospectively compensate for (Omurtag and Fenton, 2012; Uusitalo *et al.*, 2015) or determine via model checking, since there is no sampling distribution to compare the data with (Gelman *et al.*, 2010).

Consequently, this thesis argues that the data inputted into theoretical BLCMs should ideally account for two important stochastic processes: the inability to trap entire wildlife populations, and inability to consequently test entire wildlife populations. The distinction between stochastically- and deterministically-derived test data is important and can have a significant impact on the power of a BLCM.

This is easily illustrated with a simple thought experiment. If P is 20% and 100 individuals are captured, deterministically calculated test data would not reflect the real-world studies that rarely capture the expected 20 infected individuals. Furthermore, if 20 known infected individuals are tested, and the Se and Sp of the test is 80%, a deterministic study would report exactly 16 positive test results, with which a real-world study would be unlikely to agree. Actual wildlife test data is difficult to decipher because it includes the random and often imperfect processes that characterise trapping and testing in the field.

Consequently, it is possible that the power of BLCMs when using deterministically calculated data would be overestimated.

Perfect trapping efficiencies in wildlife populations are rarely attainable, and so this is a further source of uncertainty that should be considered. Moreover, it is already known that imperfect trapping efforts, when combined with imperfect diagnostics, contribute to biased estimates of Phat, which—regardless of the method for its prediction—is an often-underestimated parameter of interest (Lachish *et al.*, 2012; Miller *et al.*, 2012). Consequently, inferences of Se, Sp and P if based on an unknown population size, will have a high likelihood of being biased by the impossibility of trapping an entire population (Smith and Vanderweele, 2019).

For example, in badgers, “trappability” is known to vary among badger individuals (Byrne *et al.*, 2012), and trapping conditions vary among trapping events (Noonan, 2015) thereby impacting population-level trapping efficiencies (Tuyttens *et al.*, 1999). Nor is this impact relatively insignificant: estimates of trapping efficiencies in badgers range between 34% (Byrne *et al.*, 2012) and a figure greater than 80% (Smith and Cheeseman, 2007).

To complicate matters further, P in the field, as opposed to historic estimates of apparent P from sampled data (Lewis and Torgerson, 2012), is thought to associate with covariates such as host trappability, weather (Martin *et al.*, 2017), as well as the performance of already imperfect diagnostic tests in largely unknown ways. These findings only reinforce the need to account for imperfect trapping and testing in simulation studies using BLCMs, since these limitations present unavoidable sources of uncertainty.

Consequently, understanding how sampling error may affect estimates of Se, Sp and P—on a population level, where population size is unknown—must be regarded as a significant gap in ecologists' ability to quantify disease in live populations.

Parameterising this challenge

Conventionally, a specific output y is expected to occur when models are run with parameters denoted θ. Consequently, in deterministic modelling, when given a parameter θ, the same outcome y is expected no matter how many times the model is run. In contrast, to model stochastically, ecologists must associate many observed outcomes of y with θ via a probability distribution $Pr(y|\theta)$, where Pr is the likelihood function better reflecting the realities of testing environments in the field. Since the likelihoods of Se and Sp are not derived using P, a stable expression of test performance can be expected.

In this thesis, stochastic methods are used to generate arrays of expected binary diagnostic test results by using two random binomial processes to account for the dependencies of Se, Sp and P on theoretical diagnostic test outcomes and sampling efforts. The parameters of interest—Se, Sp or P—represented in their unconstrained state are all bounded above by one and below by zero to remain on an interpretable probability scale, and are associated with an accuracy and precision specific to their location within parameter space.

Most statistical alternatives to gold standard diagnostics infer Se, Sp and P using the latent class probabilistic models first derived during the 1980s, such as the Hui-Walter model (Hui and Walter, 1980), which—usually via maximum-likelihood methods—work to describe the link between observed test results

and latent true infection statuses. To model this link, it is a common perspective (Dendukuri, Bélisle and Joseph, 2010; Jones *et al.*, 2010) that at least three diagnostic tests should be employed when gold-standard tests are not available, though this is not always essential (Goodman, 1974), and so in this thesis, experiments are designed to explore flexible parameter spaces to verify when parameters can be inferred.

The Hui-Walter Latent Class Model is an extendable statistical tool used to overcome the impossibility of assessing diagnostic test accuracy and disease prevalence in infected wildlife and human populations (Hui and Walter, 1980). By advancing maximum-likelihood approaches to BLCM approaches, previously difficult-to-quantify disease parameters may be inferred, since prior information can explicitly resolve their otherwise missing values.

A further challenge in estimating the latent parameters Se, Sp and P is that even their best inferences may vary widely between published studies of the same host-pathogen system, and even the same study population—for example as explained by Greiner and Gardner, 2000—often due to commonly-cited reasons such as biological differences between sampled populations, methodological differences in sampling strategies or efforts, and or changes in the specifications of a diagnostic test such as its cut-off point, i.e. the agreed threshold at which a diagnostic test result can be perceived as positive or negative.

This thesis examines a less frequently cited but also important reason for variations in the inferences of Se, Sp or P which is the specification of the BLCM itself, including the impact of stochastic test data and the prior distributions used on a BLCM's explanatory power. For ecologists, the need to do this arises because of a lack of standard specifications for BLCMs, and due

to the absence of available procedures for validating a BLCMs' algorithms, data, and assumptions. The implications of this are explored further within Chapter 4.

Broadly speaking, however, the more accurately and precisely posterior inferences can describe the likelihood of a parameter, given some data, then the closer the model represents reality, and the more powerful it becomes.

In summary, for the power of a BLCM to be known across the required parameter space, the ecologist's toolkit needs to include general methods for stochastic data generation, and a verified means of generalising latent class probabilistic models across tests and populations.

Consequently, the modelling framework and associated infrastructure described in this chapter is integral to the data generation and pre-inference data processing employed throughout the studies presented in Chapters 4 to 8. This framework constitutes a stochastic modelling framework written in R (R Core Team, 2023) version 4.2.2 to generate theoretical diagnostic test data that is paired to a BLCM using the jagsUI package (Kellner, 2015). The BLCM is programmed using JAGS (Just Another Gibbs Sampler) version 4.3.1—a C++ language with similarities to the software BUGS (Bayesian inference Using Gibbs Sampling) (Lunn *et al.*, 2000)—and is used to infer results via a relevant MCMC sampler. Specifically, using JAGS as the Bayesian modelling tool provides an easy way to conduct Gibbs sampling without the need to derive the full conditional distributions, or write an MCMC sampler.

While the specifications of each simulation scenario presented will of course vary along with the specific hypothesis being tested, all models, in each chapter, conform to this general specification. It is presented here to avoid chapter-by-chapter repetition.

How is the diagnostic test data generated?

First, the stochastic methodology is applied in order to simulate arrays of diagnostic test results. Then, a BLCM uses this test array to instruct the relevant MCMC algorithm to infer the values for parameters Se, Sp and P. All the simulated data is determined by seed—a feature that can ensure the repeatability of experiments using randomly generated data—in order to facilitate the comparison of experiments as well as an understanding of how the observed data impacts on a model's ability to infer the truth.

When true values for Se, Sp and P are given, the infection statuses and resultant test outcomes of individuals can be sampled from binomial distributions to introduce stochasticity, i.e. random noise. The inclusion of this random noise sets this thesis apart from studies such as Johnson *et al.*, 2009, and previous models by Branscum, Gardner and Johnson, 2005, where deterministically-generated data is used. The studies present in this thesis use stochastically-generated test data, which accounts for the random noise associated with the imperfect trapping and testing of animals. To the best of this author's knowledge, Helman *et al.*, 2020 are the only authors to have used stochastically-generated test data to study wildlife disease.

The function, in pseudocode, that is used to generate diagnostic test results from the true values provided works as follows. This function is termed `get.values` and its various specifications can be found within the online repository at https://github.com/annabush/PhD.

Inputs: P, Se, Sp, D, M

Output: simulated diagnostic test outcomes

```
FOR each of the M individuals
    SET the individual's infection status, s = Bernoulli(P)
    FOR each of the D diagnostic tests
        SET the probability of a positive result,
                        q = s · Se + (1 − s)(1 − Sp)
        SET the test outcome d = Bernoulli(q)
    STORE each of the individuals test outcomes in binary
    format
TALLY the number of each combination of test outcomes
```

The true values for Se, Sp and P are then inputted, along with the number of diagnostic tests D, and number of individuals, M that control the simulation outside of the model. The infection status, s, of each individual is drawn using a single trial from a Bernoulli distribution, B, such that $s = B(P)$.

For each of the given diagnostic tests, the probability, q, that an individual returns a positive result can be calculated according to Equation 6.

Equation 6

$$q = s \cdot Se + (1 - s)(1 - Sp)$$

And the subsequent test outcome, d, can also be drawn from a Bernoulli distribution, such that $d = B(q)$. All diagnostic test results can then be tallied to quantify the number of observed positive (1) and negative results (0).

The generation of diagnostic test data is a crucial part of the general workflow required to produce the datasets described in Table 10-1, this workflow is outlined in pseudocode below.

SET parameters (Table 10-2) and hyperparameters (

Table 10-3: The MCMC hyperparameters used to define the JAGS models written using the jagsUI package (Kellner, 2015), their values, and why those values were chosen. These hyperparameters are relevant to the simulation analyses conducted between Chapters 5 to 7.

```
)
DEFINE functions (Table 10-4)
     EXECUTE
          INITIALISE an array to collect simulation outputs
          GET diagnostic test data (see get.outcome.matrix,
          Table 10-4)
          COMPILE all required data into a list for
          simulation
RUN simulations over multiple cores
STORE simulation outputs into results array and save
```

Why is a generalisation of the Hui-Walter model necessary?

A generalised Hui-Walter model can simultaneously test hypotheses concerning any number of independent diagnostic tests and or populations without the need to re-parameterise; with a core benefit being flexibility in the amount of information available for making inferences. This section identifies five specific reasons why a generalised Hui-Walter model is required.

Batteries of tests are used in many diagnostic settings in human and animal health, for example the use of molecular and antigen tests for detecting infection with SARS-CoV-2. Nevertheless, in the field of infectious wildlife

disease—where there is often a paucity of real-time and longitudinal disease data—it is critical that wildlife epidemiologists can utilise all the information that they have on a specific disease to fill this data gap. Bayesian ecologists sometimes account for this data gap by providing BLCMs with informative priors, yet successful modelling calls for a careful balance between the quality of prior information and "enough" diagnostic tests (or populations)—which could be inputted by proxy—in order to meaningfully direct the MCMC sampler.

The degrees of freedom of a statistical problem is a value describing the number of independent pieces of data that are free to vary when solving it (Rodríguez *et al.*, 2019). Regarding the studies presented within this thesis, the statistical problem is the identifiability of the BLCM, and the data, in this context, are the parameters that must be inferred by the BLCM. Following this logic, the degrees of freedom of any diagnostic testing scenario using multiple tests is calculated using the rule $N - 1$ (Siegel and Castellan, 1988), where N is the number of possible test outcomes available. Using this rule, Table 3-1 describes the degrees of freedom available in Latent Class Modelling situations of one to five tests.

Table 3-1: The degrees of freedom available to an estimation problem given batteries of binary diagnostic tests.

Number of tests	**Number of parameters** $(\boldsymbol{2D+1})$	**Number of test outcomes** $(\boldsymbol{2^D})$	**Degrees of freedom** $(\boldsymbol{2^{D-1}})$	**Are the degrees of freedom ≥ number of parameters?**
1	3	2	1	N
2	5	4	3	N

3	7	8	7	Y
4	9	16	15	Y
5	11	32	31	Y

A standard maximum-likelihood construct of the Hui-Walter model assumes that two imperfect tests and two populations are available, providing six degrees of freedom (Enøe, Georgiadis and Johnson, 2000). But in many cases, researchers do not have access to two study populations, and modelling one population as two subpopulations based on a selected splitting characteristic (Enøe, Georgiadis and Johnson, 2000) is not without risk, as it can be difficult to ensure that this characteristic is truly independent of diagnostic accuracy and P. Therefore, to satisfy the degrees of freedom rule, logic dictates that the simplest Latent Class Model must have three independent diagnostic tests and one population, and this is termed the Walter and Irwig 1988 model (Walter and Irwig, 1988).

Therefore, a “Three-Test, One-Population” BLCM—the Walter-Irwig model—is functionally equivalent to the “Two-Test, Two-Population” archetype termed the Hui-Walter paradigm common to wildlife disease literature (for example, Johnson, Gastwirth and Pearson, 2001) since it provides the minimum model identifiability required in terms of degrees of freedom versus the number of parameters to be inferred.

Considering this, the models developed for this thesis build on the concept of the Walter and Irwig “Three-Test, One-Population” model described in Drewe *et al.*, 2010 and McDonald and Hodgson, 2018, though they are specified

differently. A generalised “Any-Test, Any Population”—a generalised Hui-Walter model—in the form described is required for five main reasons:

1. The accuracy of a diagnostic test—i.e., its ability to produce correct results (Gardner *et al.*, 2000)—is dictated by the values of Se and Sp. In turn, any inferences of Se, Sp and P will have their own accuracies and precisions, defining the power of the BLCM. The sensitivity of BLCMs must be validated across the entire parameter space to ensure that their power and assumptions do not break down within parameter space (see Chapter 7). Accordingly, to test this, the BLCM specification must be flexible across different numbers of tests and populations, though the latter does not apply within this thesis as splitting characteristics are not explored (the infrastructure is however supplied for population-based studies). A non-generalised Hui-Walter model does not meet this specification.
2. The quantity of both tests and populations are limiting factors for real-world studies, and so ecologists may wish to include proxy information—such as the expert analysis of clinical information, which can be considered a diagnostic test in its own right if associated with an Se and Sp (Albert and Dodd, 2008)—in order to ensure that the number of tests and populations required satisfies the degrees of freedom rule. Proxy information can be used to substitute for a biological diagnostic, and further examples of proxy tests might include expert elicitation (van de Schoot *et al.*, 2021) based upon veterinary opinion, proximity-logged information to known infected individuals, or expert opinion from animal behaviouralists. For example, Mazeri *et al.*, 2016 used cattle inspection data as proxy for classical diagnostics in a study on liver fluke *Fasciola*

hepatica in cattle, finding expected diagnostic accuracies for five fluke-specific tests. The Bayesian framework allows such proxy tests to be included, often in the form of beta distributions, enabled by software such as the R version of Wes Johnson's and Chun-Lung Su's Betabuster tool (Stevenson *et al.*, 2020). A non-generalised Hui-Walter model is not flexible enough to integrate new proxy tests without re-parameterising the model.

3. A generalised Hui-Walter model is particularly useful for ecologists because gold standard field tests are rare, imperfect tests are expensive, and there are rarely more than two of them, but the broad ecological knowledge of infected populations is usually large. And it is this ecological knowledge—such as area-specific population densities that in turn can inform predictions on probable trapping efficiencies—that can be combined with existing imperfect diagnostic tests as a proxy to improve inferences. Mainly, this information is indirectly related to the epidemiology of the infected population. A single non-generalised Hui-Walter model cannot include this information readily without the use of multiple models.
4. While authors such as Berkvens *et al.*, 2006, Ochola *et al.*, 2006, and Pereira *et al.*, 2012 do reference generalisations of the Hui-Walter paradigm, algorithms for an Any-Test, Any-Population model have yet to be made accessible or available for ecologists to use in the context of sensitivity analyses. Once again, the desirability of a generalised Hui-Walter model is apparent.
5. The first BLCM was published in 1995 by Joseph, Gyorkos and Coupal, and the first estimable Bayesian Hui-Walter model—i.e., a model in

which all nonseparable parameters Se, Sp and Pi were identifiable (Ponciano *et al.*, 2012)—was published six years later in 2001 (Johnson, Gastwirth and Pearson, 2001). Yet while it is clear that a third test allows a simple one-population study to satisfy the degrees of freedom rule—and the first BLCMs were now published nearly 30 years ago—only one analysis has been found on of the importance of the third diagnostic test in understanding batteries of non-gold diagnostics (Dendukuri, Bélisle and Joseph, 2010). The importance of the third test is clearly in doubt: indeed, the Three-Test, One-Population scenario has been disparagingly referred to as "*not exactly estimating the parameters, merely rewriting data*" (Toft, Jørgensen and Højsgaard, 2005). Yet again, a non-generalised Hui-Walter model is unsuitable for studying BLCMs using differing numbers of tests simultaneously.

Although latent class methods were popularised by Hui and Walter, the sheer volume of studies published over the past few years indicate that Latent Class Analysis is a rapidly evolving field of study.

The Bayesian specification of the extended Hui-Walter paradigm

"*Bayesian inference is the re-allocation of credibility across the overall parameter space*" (Kruschke and Liddell, 2018).

The results presented in this thesis are generated from an MCMC sampler implemented using JAGS (Plummer, 2003), which constructs Markov chains over parameter spaces that converge and provide posterior distributions of interest. The JAGS "black box" is relied upon to decide the exact sampler required, as well as the MCMC parameter space in which posterior credibility lies.

When considering a One-Test, One-Population model, the diagnostic test can—for any given individual—return one of two possible outcomes: positive (1), or negative (0). And, once each individual has been tested, all the diagnostic test results can be tallied into a single vector: $N = [N_0 \quad N_1]$, where N_0 is the number of negative test results, and N_1 is the number of positive test results. However, when considering a Three-Test, One-Population model, the combined results from each of the three diagnostic tests can be categorised into one of eight (2^3) possible outcomes—a negative result from each test (000), through to a positive result from each test (111)—but can still be tallied into a single vector, N, as shown in Equation 7.

Equation 7

$$N = [\, N_{000} \quad N_{001} \quad N_{010} \quad N_{011} \quad N_{100} \quad N_{101} \quad N_{110} \quad N_{111}].$$

A diagnostic test will be negative if it correctly determines that a healthy individual is not infected, or if it incorrectly infers that an infected individual is healthy. Therefore, when d is the diagnostic test outcome, and s is the infection status of the tested individual the probability, q_0, that a single diagnostic test returns a negative result can be expressed as shown in Equation 8.

Equation 8

$$q_0 = Pr(d = 0 \mid s = 0) + Pr(d = 0 \mid s = 1)$$

On the other hand, a diagnostic test will be positive if it correctly determines that an infected individual is in fact infected, or if it incorrectly infers that a healthy individual is infected. Therefore, the probability, q_1, that a single diagnostic test returns a positive result can be expressed as shown in Equation 9.

Equation 9

$$q_1 = Pr(d = 1 \mid s = 1) + Pr(d = 1 \mid s = 0)$$

If an individual is selected at random from a population, the probability of that individual being infected is determined by disease prevalence, P. If the individual in question is in fact infected, then the probability that the diagnostic test returns a correct result is determined by Se. If, on the other hand, the individual is healthy, then the probability of a correct diagnostic test result is given by Sp.

Consequently, the probability that a diagnostic test is negative, q, can be calculated as:

Equation 10

$$q_0 = (1 - P) \cdot Sp + P(1 - Se)$$

And the probability that a diagnostic test is positive, q_1, can be calculated as:

Equation 11

$$q_1 = P \cdot Se + (1 - P)(1 - Sp)$$

Extending this construct to a Three-Test One-Population scenario increases the number of possible diagnostic test outcomes to eight, and so when a testing a single, randomly-selected individual, the probability of each outcome being reported can be expressed as:

Equation 12

$$q_{000} = (1 - P) \cdot Sp_1 \cdot Sp_2 \cdot Sp_3 + P(1 - Se_1)(1 - Se_2)(1 - Se_3)$$

$$q_{001} = (1 - P) \cdot Sp_1 \cdot Sp_2(1 - Sp_3) + P(1 - Se_1)(1 - Se_2) \cdot Se_3$$

$$q_{010} = (1 - P) \cdot Sp_1(1 - Sp_2) \cdot Sp_3 + P(1 - Se_1) \cdot Se_2(1 - Se_3)$$

$$q_{011} = (1-P)\cdot Sp_1(1-Sp_2)(1-Sp_3) + P(1-Se_1)\cdot Se_2\cdot Se_3$$

$$q_{100} = (1-P)(1-Sp_1)\cdot Sp_2\cdot Sp_3 + P\cdot Se_1(1-Se_2)(1-Se_3)$$

$$q_{101} = (1-P)(1-Sp_1)\cdot Sp_2(1-Sp_3) + P\cdot Se_1(1-Se_2)\cdot Se_3$$

$$q_{110} = (1-P)(1-Sp_1)(1-Sp_2)\cdot Sp_3 + P\cdot Se_1\cdot Se_2(1-Se_3)$$

$$q_{111} = (1-P)(1-Sp_1)(1-Sp_2)(1-Sp_3) + P\cdot Se_1\cdot Se_2\cdot Se_3$$

And the probabilities derived from Equation 12 can be incorporated into a single vector, Q, as shown in Equation 13.

Equation 13

$$Q = [\, q_{000} \quad q_{001} \quad q_{010} \quad q_{011} \quad q_{100} \quad q_{101} \quad q_{110} \quad q_{111}]$$

If varying numbers of tests are required within a study, including information that can serve as a proxy for a diagnostic test such as veterinary opinion, the Hui-Walter model can be generalised to any number of diagnostic tests, D, as follows, where $\odot$ denotes component-wise multiplication, and Ω denotes a $2^D \times D$ matrix of all possible diagnostic test outcome combinations (see `get.outcome.matrix`, Table 10-4):

Equation 14

$$Q = (1-P)\prod_{i=1}^{2^D}(SP^T \odot (1-\Omega) + (1-SP^T)\odot \Omega) + P\prod_{i=1}^{2^D}(SE^T \odot \Omega + (1-SE^T)\odot \Omega)$$

Importantly, the counts of the different observations, N, are assumed to have independent multinomial sampling distributions (since there are usually more than two outcomes) as shown in Equation 15.

Equation 15

$$N \sim multinomial\left(Q, \sum_{i=1}^{n} N_i\right)$$

The calculation of Q as shown in Equation 14 also relies on the following five assumptions, noting that Hui and Walters' assumption of independent disease prevalence does not apply in the Walter-Irwig construct since only one population is studied.

1. Each test is equal in its diagnostic capability across different stages of disease progression.
2. Diagnostic accuracy is independent of population.
3. All tests are conditionally independent of each other in terms of how they measure infection, i.e. the results of a second test do not rely on the results of the first, and the testing mechanisms are sufficiently different so that they are not, for example, both blood tests. Violations of this assumption are studied elsewhere (Branscum, Gardner and Johnson, 2005; Toft, Jørgensen and Højsgaard, 2005).
4. The combinations of test outcomes, for any number of tests, follows a multinomial distribution since more than two outcomes are possible.
5. Diagnostic test accuracy is independent to individual-level heterogeneities in the ability to diagnose.

The following JAGS code within the `set.model` function provides a generic example of the Bayesian specification of the Any-Test, Any-Population model. This function below writes the BLCM definition to file.

```
set.model <- function(filepath="model.txt") {

  writeLines(
"model{
  for (i in 1:n.tests) {
```

```
    se[i] <- mu.se[i]
    sp[i] <- mu.sp[i]
  }
  pi <- mu.pi
  for (i in 1:n.outcomes) {
    for (j in 1:n.tests) {
      # A = se if badger is positive, 1 - se otherwise
      # B = 1 - sp if badger is positive, sp otherwise
      A[i, j] <- outcomes[i, j] * se[j] + (1 - outcomes[i, j]) * (1 -
se[j])
      B[i, j] <- outcomes[i, j] * (1 - sp[j]) + (1 - outcomes[i, j]) *
sp[j]
    }
    p[i] <- pi * prod(A[i, 1:n.tests]) + (1 - pi) * prod(B[i,
1:n.tests])
  }
  y[1:n.outcomes] ~ dmulti(p[1:n.outcomes], n)
  for (i in 1:n.tests) {
     mu.se[i] ~ dnorm(prior.se[i], precision) T(se.limit[1],
se.limit[2])
     mu.sp[i] ~ dnorm(prior.sp[i], precision) T(se.limit[1],
sp.limit[2])
  }
  mu.pi ~ dunif(pi.limit[1], pi.limit[2])
}",
    con=filepath,
  )
}
```

Calibrating three important model performance indicators of BLCMs

Performance indicator 1: prior distributions

Constraining prior distributions according to *a priori* assumptions, is a widely used method for limiting the size of parameter space. A model given constrained priors can be thought of as a nested version of a full model—i.e., a model capable of searching the entirety of parameter space—as it represents a proportion of space within it. There are two general types of constraint applied in this thesis that should not be confused. One is the application of constraints to prior distributions activated within the BLCM construct in order to limit the parameter space searched by the MCMC (prior constraints) (Gelman and Carpenter, 2020). The second is the constraining of the true values generated at the same time as the stochastic test data, in order to ensure that the truth does not lie outside of the given prior distributions (constraints to true values).

The latter method does not preclude the former, but experiments where the truth lies outside of the given prior distributions are redundant, and both methods may be classified as supplying forms of prior information (see Figure 7-1).

For prior constraints, limits are applied to prior distributions using reasonable assumptions. In this case, a constrained parameter space is generally that where values of P are less than 0.5 and values of Sp are greater than 0.5 (see Appendix 2: Key parameters, hyperparameters and functions for justifications).

Constraints are mechanisms to provide posterior inferences without overburdening a computer with a prohibitive number of model runs (Berkvens *et al.*, 2006; Gonçalves *et al.*, 2012) and thus extending the computational runtime. This is an important consideration given that the number of possible combinations of parameter values increases exponentially with the dimension of parameter space being considered. For example, Gelman and Carpenter, 2020 used constraint to set the prior scale of Se to 0.5 or less to rule out the possibility of a very high values of P corresponding to an unrealistic Se values. If priors do not cover the range of expected true values, then the model is over-constrained.

Accordingly, constraint is specifically employed within this thesis to improve the accuracy and sometimes identifiability (Wu *et al.*, 2021)—in situations where parameters are able to be inferred though the inputted data contains limited information about the parameter of interest (Ponciano *et al.*, 2012)—of any solutions to Equation 14. Despite valid solutions existing across hyperdimensional parameter space, the answer may be incorrect unless constraint is applied, and so constraint is required in order to direct models

towards sensible solutions. Constraint has been found to be particularly useful when the number of true values is small (Wu *et al.*, 2021).

The purpose of this form of constraint can be usefully thought of as follows. Considering an infected population of X infected individuals and N total individuals, the fact that the population is infected means that the statement $X > 0$ must be true, while the statement $X > N$ cannot be true, so to assist a model in providing meaningful answers these conditions could be inputted as constraints.

As outlined by Hobbs and Hooten, 2015, model parameters can be constrained via their prior distributions in order to find more certain models, and this includes any prior distribution that is uniform. Importantly, valid inferred values—and therefore outcomes available to a Bayesian model—are largely based on the information contained in priors, which must be elicited appropriately. Within the workflow in this thesis, parameter space exists according to the constraints assumed within the process for selecting true values.

Within ecology, debate remains on how to specify prior information (Banner, Irvine and Rodhouse, 2020), and in studies employing BLCMs, poor prior specification often causes inconsistent model conclusions (Hobbs and Hooten, 2015) but the adverse effects of this are often nullified by the fact that prior information can be provided to improve model success (Gonçalves *et al.*, 2012).

It is considered that a BLCM's performance should ideally be analysed given a normal prior, no matter how informative (Gelman, Simpson and Betancourt, 2017). However, in this thesis, the performance of BLCMs given normal priors is compared to the performance of BLCMs given uniform priors to understand whether uniform priors can be useful to BLCM analyses, particularly given that

ecologists often choose uniform priors as default (Banner, Irvine and Rodhouse, 2020).

This section now describes the general methodology for ensuring that BLCMs do not have an excessive reliance on priors. Broad definitions for the common types of priors available are described in Banner, Irvine and Rodhouse, 2020. In this thesis informative priors are referred to as “precise” priors, weakly informative priors are referred to as “imprecise” priors, and non-informative priors are referred to as “uniform” priors.

To ensure that no intrinsic correlations between true and inferred values exist among parameters, the known true values of each stochastic method are inputted as either randomly generated fixed values, or values randomly selected from the uniform distribution, U, where $U = unif(0,1)$. To avoid bias, it was ensured that uniform distributions on the interval $[0,1]$ were used for the selection of true values, as outlined in Toft, Jørgensen and Højsgaard, 2005.

For tests which specify prior information via normal priors, the true values are initially used as the mean for a truncated normal distribution specified as $N(U, sigma)T(0,1)$ on the probability scale, where sigma is the given standard deviation signifying how closely the prior information matches the true value for each draw. The mean of the prior distribution is then drawn from the truncated normal distribution in order to ensure that the prior distribution is not introducing bias by being centred exactly on the true value (Figure 3-1). Defining the draw standard deviation is particularly important considering that the density of the prior information, may look like the distribution shown in Figure 3-1 for any given parameter with the same truth.

Normal priors were specified rather than the alternative option of using beta priors—where two shape parameters must be specified—due to their flexibility to inform models, i.e. symmetrical, unimodal, and zero-mean priors could be specified with known variances. The prior distributions used in this thesis were first examined visually (Figure 3-2) to verify their expected shape and behaviour around the truth.

The following pseudocode illustrates how arrays of true values and arrays of prior means are generated.

```
Inputs: P limits, Se limits, Sp limits, n.sim, n.tests. draw.sd
Outputs: array of true values, array of prior means
FOR each simulation
     FOR each test
          SELECT true Se from U(Se, limits)
          SELECT true Sp from U(Sp, limits)
          SELECT true P from U(0, 1)
          STORE true values in an array
          SELECT prior Se from N(true Se, draw.sd) T(0, 1)
          SELECT prior Sp from N(true Sp, draw.sd) T(0, 1)
          STORE values of prior means in an array
```

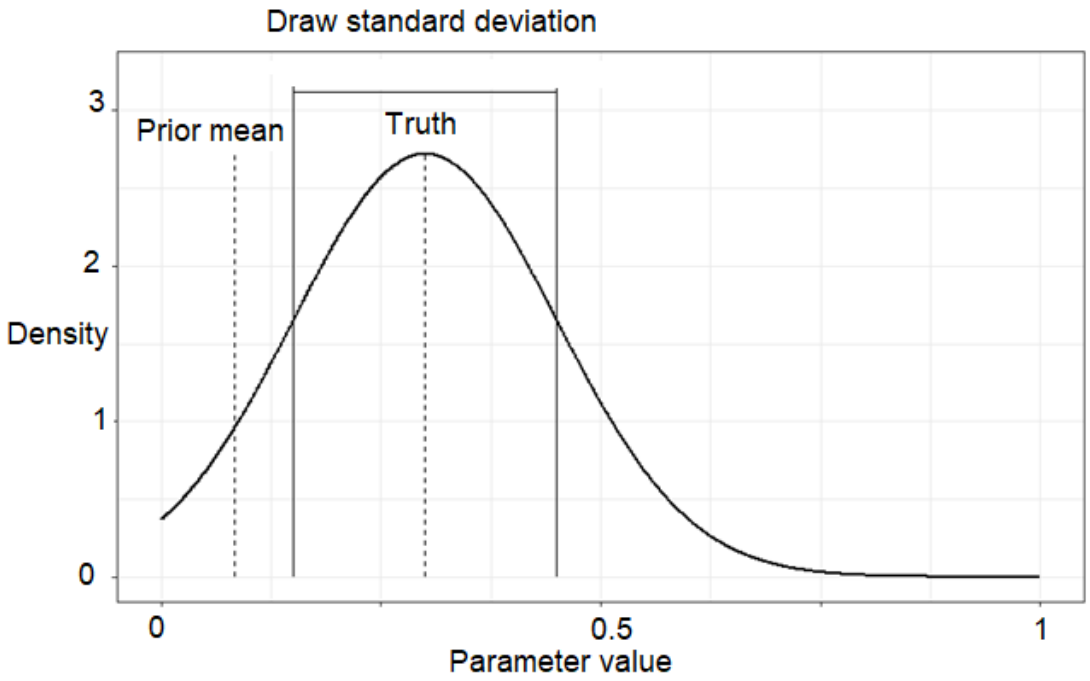

Figure 3-1: A schematic illustrating that in simulation studies the means of informative prior distributions are selected from a probability density function with the truth as its mean, and a standard deviation that avoids the generation of over-informative priors. The “prior mean” is just one realisation of the draw from the distribution of means.

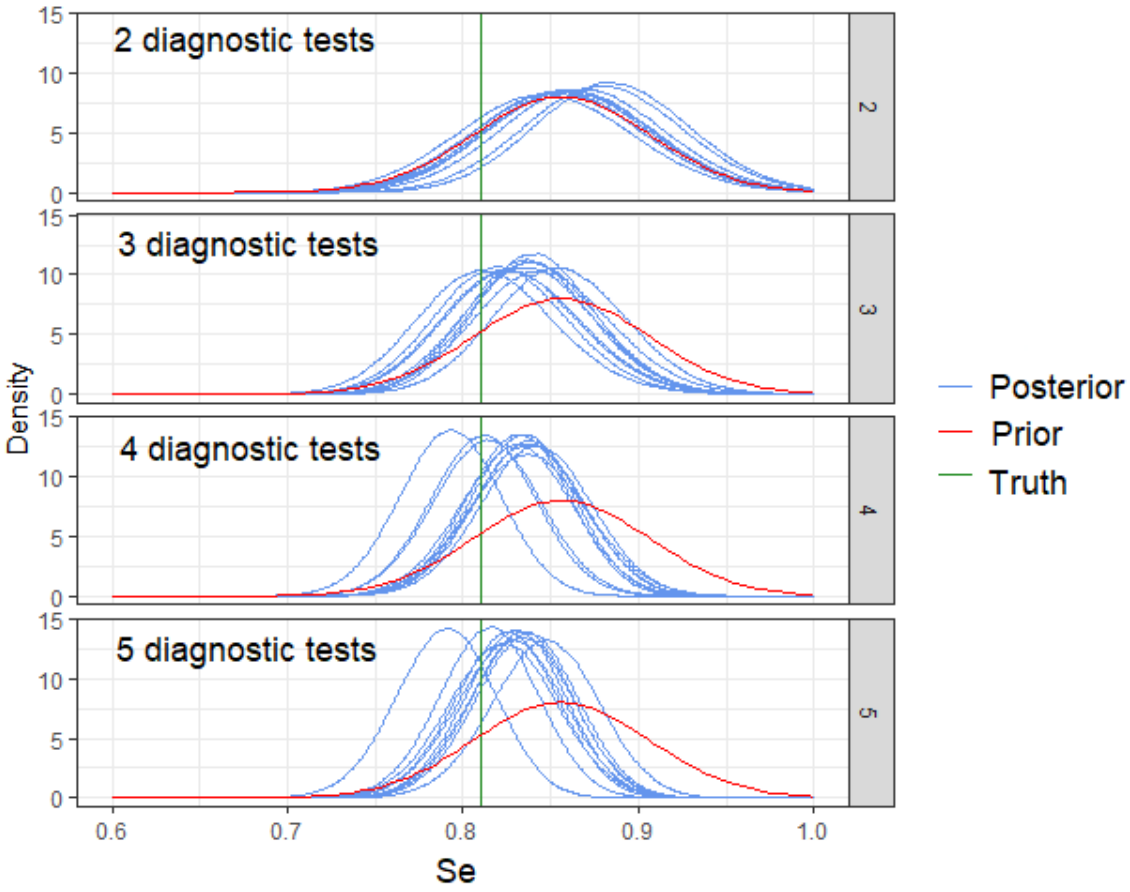

Figure 3-2: An example of a visual prior-posterior check. The probability densities of the posterior inferences of Se are in blue—where each function relates to a single simulation—and can be compared to the probability density of the informative prior of Se in red, given a set truth shown in green. The visual shows that as the number of diagnostic tests available increase, the posterior density moves closer towards the common truth.

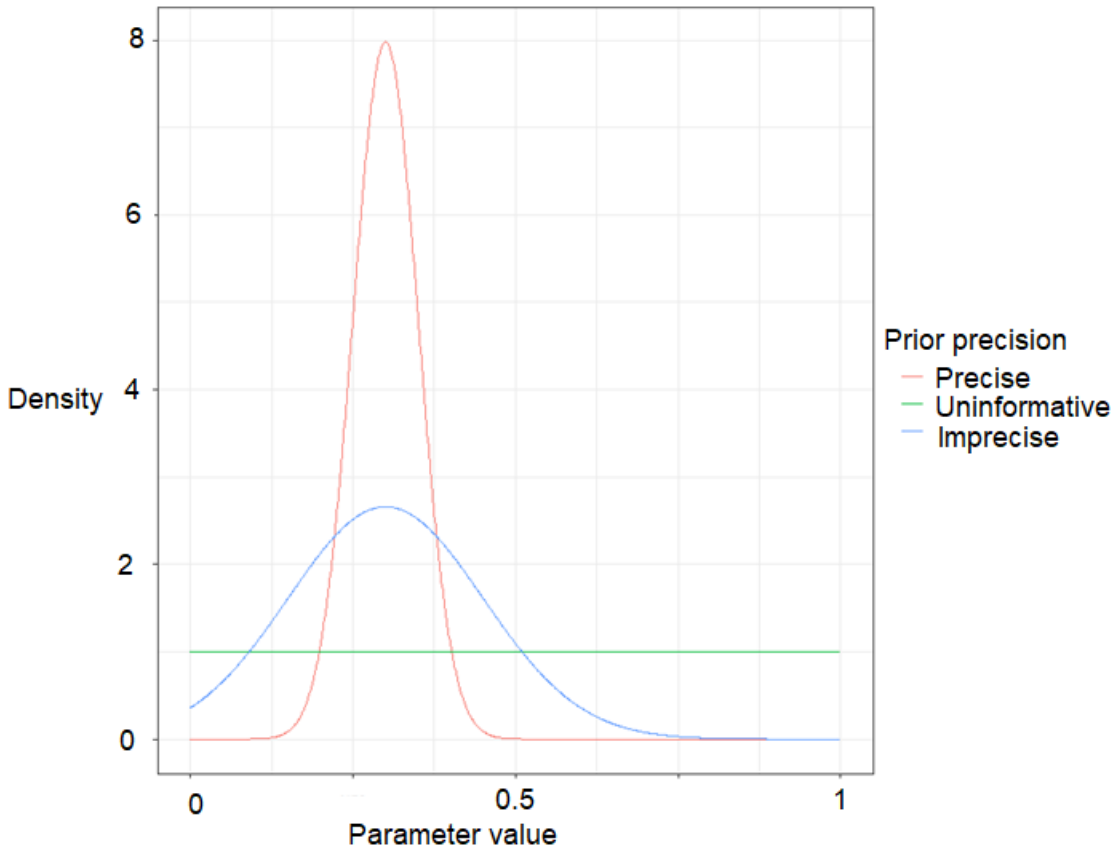

Figure 3-3: Hypothetical probability densities of precise, imprecise, and uninformative prior precisions of a parameter where the given truth is 0.3.

Performance indicator 2: addressing non-convergence.

In ecology, checks for convergence in general are often only undertaken visually, such as those undertaken by Arango-Sabogal *et al.*, 2019. Any subsequent degree to which convergence issues can be corrected—usually via re-parameterisation—is largely dependent on further computation and therefore the amount of RAM available, since MCMC calculations are dependent on in-memory computation.

For this thesis, the amount of RAM available was maximised via parallelisation across the available cores on the University of Exeter's remote Linux servers, situated in its High-Performance Computing facility. Each instance of Linux was accessed via the remote networking software MobaXterm, with processing

conducted by two RStudio Pro servers, enabling additional parallelisation through the use of two concurrent R sessions. The “tuning” parameters that must be set when using an MCMC model are described and justified in Table 10-3. Importantly, while access to High-Performance Computing offered a considerable time saving, access to large amounts of RAM is not prerequisite to running the supplementary code to this thesis available at https://github.com/annabush/PhD.

The `traceplot` function of the `coda` package (Plummer *et al.*, 2006) was used to check for convergence, and it was ascertained that convergence *could* have been achieved throughout all the models in this thesis through a visual “convergence check”. Ecologists however must be mindful that visual convergence checks can be less helpful for complex models searching in up to 11-dimensional parameter space—for example, the sampler may become “stuck” in specific volumes of parameter space (Gelman and Rubin, 1992) disguising whether models have in fact mixed properly. For this reason, it was not considered that visual modelling diagnostics could provide certainty that convergence was achieved, and a more quantitative approach to diagnosis was explored.

The quantitative convergence checks.

Ensuring the absence of non-convergence enables MCMC chains to be irreducible, i.e. the chains can reach all places of the target distribution, and aperiodic, i.e. the chains do not get stuck in cycles (Roberts, 1995). It is assumed that since the parameters Se, Sp and P are intrinsically linked (Equation 14), there will always be some inter-correlation between these parameters (see Figure 3-4). Computational difficulties in achieving convergence—and specifically eliminating autocorrelation—were therefore

expected, and while one approach to the problem is to re-write the likelihood function (Liu *et al.*, 2022), the purpose of this section is to describe how the authenticity of convergence was verified quantitatively.

While some autocorrelation is inevitable, autocorrelation in general affects the amount of information available to a Markov chain and is a measure of how dependent any current value of a chain compares to previous values due to the iterative steps taken by the MCMC algorithm not being entirely random. Since convergence cannot be assured for any model apart from the simplest textbook examples (Lunn *et al.*, 2000), it could be argued that—using inductive reasoning (Saint-Mont, 2022)—any proffered claims for achieving, or not achieving, convergence could become circular logic: for example, remedying autocorrelation by thinning removes even more information available to the Markov chain, information that could be regarded as useful for reducing autocorrelation (Link and Eaton, 2012). In short, guidance for ecologists describing how to address autocorrelation and associated non-convergence issues within complex MCMC algorithms is limited.

Consequently, the approach to autocorrelation taken in the research described in this thesis begins with the observation and agreement with Link and Eaton, 2012 that addressing complex autocorrelation via thinning—the apparent standard for addressing autocorrelation—is not robust. Next, it was noted that autocorrelation was smaller in longer MCMC chains compared to when thinning was used but thinning comes at the cost of a loss of precision in the subsequent model outputs due to the reduction of data available to construct posteriors. Autocorrelation was consequently addressed by maximising the number of MCMC iterations to provide “effectively independent” samples (Link and Eaton, 2012).

The Effective Sample Size (ESS) is the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to i.e., the sample size adjusted for autocorrelation. Markov chains with a high autocorrelation have a low ESS per unit of computational time and this causes another convergence issue called the “slow mixing problem” (Duan, Johndrow and Dunson, 2018), which is an additional drain on RAM. Addressing autocorrelation becomes less important when the ESS is above the minimum ESS—which should not be below 10,000 in complex models (Kruschke, 2014)—required to produce good quantile estimates at 95% confidence.

The JAGS hyperparameters (

Table 10-3: The MCMC hyperparameters used to define the JAGS models written using the jagsUI package (Kellner, 2015), their values, and why those values were chosen. These hyperparameters are relevant to the simulation analyses conducted between Chapters 5 to 7.

) were chosen after an analysis of whether the number of iterations had achieved the minimum ESS, using the `minESS` function in the `mcmcse` package in R (Flegal *et al.*, 2017) where the dimensions of the estimation problem are calculated as $2D + 1$ where D is the number of diagnostic tests. The minimum ESS was exceeded in every model, for every parameter; and true ESS was found to be of a magnitude large enough to fully account for autocorrelation within the posterior inference. Specifically, multivariate ESSs—for one MCMC chain using the `multiESS` function of the `mcmcse` package *op cit.*—were calculated since it could not be assumed that the MCMC algorithm would carry out fully independent sampling (Vats, Flegal and Jones, 2019).

Using this heuristic, and to provide illustrative figures only, the ESS of a randomly selected Three-Test, One-Population model within this thesis was calculated to be 441,608 with a 0.8% tolerance level (Flegal *et al.*, 2017)—indicating that the MCMC model can in general be expected to be within 0.8% of the posterior inference 95% of the time.

Addressing the label switching problem

When unconstrained parameter space is studied in theoretical settings, and particularly with uniform priors, it is possible that convergence may suffer from the 'label switching problem' (Celeux, 1998). This can be illustrated using Equation 12, where switching Se with $1 - Sp$, Sp with $1 - Se$ and P with $1 - P$ on the right-hand side, would yield the same result on the left-hand side, meaning that while posterior values could be resolved, an incorrect or bimodal distribution could result.

Toft, Jørgensen and Højsgaard, 2005 propose that one solution to the label switching problem is to require Se and Sp to sum to above one, which was the method employed within this thesis. In addition, a second rule to avoid the label switching problem was investigated, requiring P to be less than 0.5 and Sp to be above 0.5. A third solution to the label switching problem required the provision of at least some prior information and to then use multiple chains (Collins and Huynh, 2014); accordingly, three chains were always used (see Table 10-3: The MCMC hyperparameters used to define the JAGS models written using the jagsUI package (Kellner, 2015), their values, and why those

values were chosen. These hyperparameters are relevant to the simulation analyses conducted between Chapters 5 to 7.

). Despite this, some hypotheses investigated within this thesis required the provision of uniform priors—it is important to still test models with uniform priors to, for example, ensure that the posteriors are informed by data rather than the prior—and unconstrained parameter space (see Chapter 7 and Chapter 8).

Checking for correlation between true and inferred values

Correlation density plots demonstrated that the values of Sehat and Sphat are not intrinsically correlated and are normally distributed (Figure 3-4). This check was completed using the `correlationPlot` function from the `BayesianTools` package (Hartig *et al.*, 2017), and suggests that correlations between parameters are not biasing the MCMC sampler. The Pearson coefficients did not exceed the widely accepted threshold for a significant correlation, which is 0.7 (Ratner, 2009). Note, some level of correlation is expected between the true and inferred values given randomly selected truths; this bias was accounted for by running each chain for many iterations.

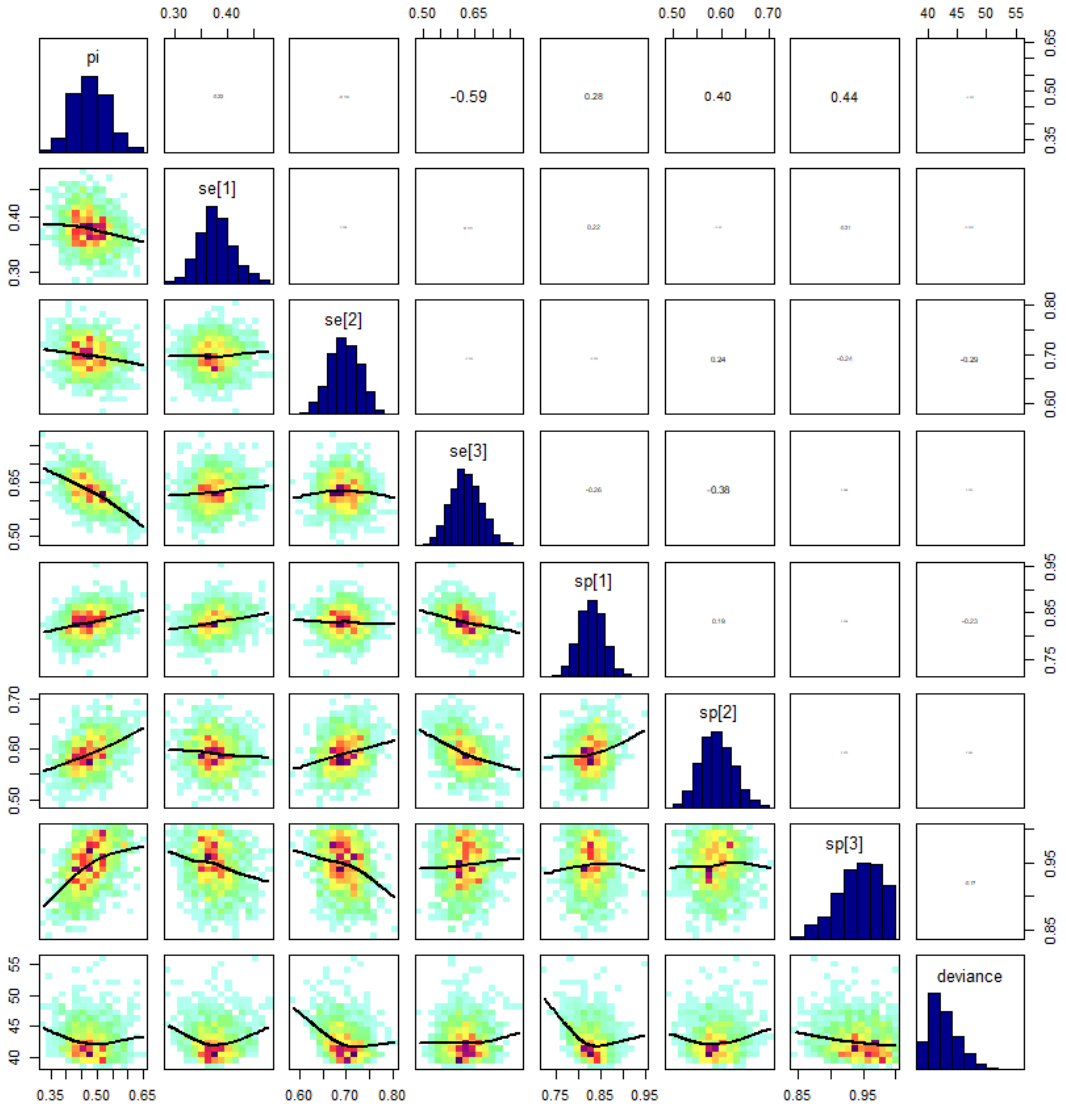

Figure 3-4: A correlation density plot showing the densities and Pearson correlation coefficients of the true and inferred values for each parameter within a randomly selected three-test model, where [1] denotes diagnostic test 1 and so on.

Performance indicator 3: The accuracies and precisions of BLCM inferences.

In short, the ability of a model to identify the truth with certainty is determined by measuring the accuracy and precision of replicate inferences, and is dependent on how easily the parameters can be inferred by the model (Ponciano *et al.*,

2012). These measures are specifically defined in Equation 16, Equation 17, and Equation 18 below.

In general, the accuracy of Se and Sp describe how well a diagnostic test can produce a correct outcome, and the accuracy of P describes how well the model inferred the proportion of infected individuals in a population, based on a sample from that population. In contrast, the precision of a parameter is indicative of how much credibility we can place on its inference, inclusive of how consistent that inference should be: precision is often an ignored indicator of assurance (Toft, Jørgensen and Højsgaard, 2005) when BLCMs are reported.

The accuracy measure termed “error”, describes the absolute inferred difference between the truth and the mean posterior inference for a given parameter (Equation 16),

Equation 16

$$Error = |\hat{y} - y|$$

where $\hat{y}$ is the predicted value, and y is the truth. Error is generally calculated as a mean error inclusive of all replicate inferences for either Se, Sp or P.

In contrast, the accuracy measure termed “bias”, describes the raw inferred difference between the truth and the mean posterior inference for a given parameter (Equation 17).

Equation 17

$$Bias = \hat{y} - y$$

The precision of a parameter is defined in accordance with Equation 18, where σ is the mean of the standard deviations of the simulated posterior distributions. This definition is used in preference to other possible definitions—for example,

the definition within Lunn *et al.*, 2000—since σ is measured in the same units as the parameter.

Equation 18

$$Precision = \frac{1}{\sigma^2}$$

Conclusion

This chapter has outlined the general underlying methodology that forms the basis of the stochastic framework and BCLMs employed within this thesis. It is recognised that the concepts are not all new, and that the correction of false positive and false negative diagnoses is a well-known probability problem—especially when P is low, as is the case in many persistent wildlife infections—but that general agreement is still lacking among researchers regarding the "right" combination of tools for interpreting stochastically-generated diagnostic test data across multidimensional parameter space.

The generalisation of the Hui-Walter paradigm, capable of handling the Walter-Irwig construct, is described: its functions are generalised to handle any number of diagnostic tests—and populations—within the same simulation. And, a novel stochastic modelling architecture is described, advancing previously deterministic versions within the wildlife disease literature (Branscum, Gardner and Johnson, 2005; Drewe *et al.*, 2010; McDonald and Hodgson, 2018). The stochastic framework accounts for noise when applying latent class methodology to real-world scenarios in terms of detection error, and the use of imperfect testing information.

To summarise, a generalised Hui-Walter model is required for the following five reasons:

1. Models should be flexible across numbers of tests (and populations). This is because any confidence in inferred values is dependent on the power of the BLCM itself.
2. Models should be flexible across the type of tests employed. This is because expert prior elicitation in Bayesian frameworks can account for proxy tests; and a prior is only as good as its justification. Proxy tests may increase the ability for many disease studies to meet the minimum number of degrees of freedom for parameter identifiability.
3. Models must be flexible about how existing imperfect tests are specified. This is because the general knowledge base on infected populations is often larger than the knowledge of the latent parameters of interest, but the general knowledge base is often intrinsically linked—via other latent parameters—to the latent parameters of interest.
4. No generalisations of the Hui-Walter model have been specified in the ecological literature or are accessible to ecologists particularly to use for sensitivity analyses.
5. No known research examines the important of the third opinion needed to satisfy the minimum degrees of freedom required to identify the Three-Test, One-Population scenario investigated in this thesis.

Based on the general methodologies described—that are modified as appropriate for the purposes of each chapter—subsequent Chapters 4 to 8 use generalisations of the Hui-Walter model to interpret stochastically-generated diagnostic test data across multidimensional parameter spaces.

Chapter 4

4. Considerations for the validation of Bayesian Latent Class Models using simulated data.

Introduction

The World Organisation for Animal Health (OIE) endorsed the use of Latent Class Models for the estimation of epidemiological parameters as recently as 2013 (Gardner *et al.*, 2021), despite the fact that the first Latent Class Model—termed ‘latent structure model’—was published in 1968 (Lazarsfeld and Henry, 1968), and that BLCMs have been widely adopted since 1995 (for example, Joseph, Gyorkos and Coupal, 1995; Johnson, Gastwirth and Pearson, 2001; Branscum, Gardner and Johnson, 2005). Only since 2017 have researchers (such as Krolewiecki *et al.*, 2018; Rahman *et al.*, 2019; Islam *et al.*, 2020) been able to follow the 30-point checklist (Kostoulas *et al.*, 2017) making up the Standards for the Reporting of Diagnostic Accuracy studies that use Bayesian Latent Class Models (STARD-BLCM).

Currently, even with OIE’s global advocacy for the use of LCMs for diagnosing animal disease, and the existence of a standard protocol for presenting research using BLCMs, ecologists still lack a standard protocol for describing how to validate their custom-built BCLM algorithms—a procedure that should ideally occur before any model selection (Hobbs and Hooten, 2015) takes place, before any diagnostic test performances are validated, and certainly before any research is presented.

Accordingly, this chapter presents two model validation examples, which can serve as a foundational template for model validation exercises by ecologists wishing to evaluate their own BLCMs. These are based around seven general “stylised facts” (Kaldor, 1961) to be identified across the parameter spaces explored within this chapter, which demonstrate the specific type of information that can be gathered from validating BLCMs. These findings, or stylised facts, relate to the accuracy of BLCM inferences, specifically in terms of the magnitude of error (Equation 16), and are:

1. Seemingly successful inferences in two-test scenarios may simply be due to the posterior replicating the prior, and so should be treated with caution.

2. Unidentifiable areas of parameter space may occur where error does not decrease when the number of diagnostic tests available for inference increase.

3. Increasing the number of diagnostic tests has the greatest effect on decreasing the error of Phat.

4. Prior constraints are particularly important for reducing errors associated with Sphat over and above the reduction in errors associated with increasing the number of diagnostic tests.

5. Prior precision is particularly important for reducing errors associated with Phat and Sphat in addition to the reduction in error from increasing the number of diagnostic tests.

6. The errors associated with Sehat are inversely proportional to the errors associated with Sphat.

7. Phat is particularly difficult to infer when Sp is low.

The findings within this chapter have implications for researchers who consult both the medical and wildlife literature for what to consider when validating a BLCM, or even when considering the intended use of a specific diagnostic test. This work provides a stepping-stone towards further research that can map specific volumes of parameter space that may lack practical identifiability, a term describing how confidently a BLCM can infer parameters given noisy data.

To summarise, the use and interpretation of BLCMs demands care (Schofield *et al.*, 2021), specifically in terms of understanding how accurately BLCMs infer Se, Sp and P in the required parameter space. To do this, it is important to validate the power of a BLCM before it is used, yet there are no standardised ways to do this, and it is likely that publication of models in ecology that have not been validated are "*more common than appreciated*" (DiRenzo, Hanks and Miller, 2023). This chapter addresses this problem in the following three ways.

1. Model validation is defined, and the critical connection between model validation and robust BLCMs is made.
2. Two examples of how a BLCM could be validated are provided.
3. The specific type of information that can be gathered from validating BLCMs is demonstrated.

What is model validation, and why do it?

Model validation is a process for verifying "*that a model is acceptable for its intended use because it meets specified performance requirements*" (Rykiel, 1996). Within this thesis, the term is used to describe an evaluation of the ability of a BLCM to infer Se, Sp and P given simulated data. While model validation is not a prerequisite to inference, it *is* a prerequisite to drawing valid conclusions from inferences (Tredennick *et al.*, 2021).

Model validation according to its “classical” definition as stated above is only possible if independent data is available (Yates *et al.*, 2018), which in the field of wildlife disease ecology generally means simulated data. Model validation on simulated data can provide useful information to real-world studies with no known truths, including a better understanding of the requisite BLCMs and their assumptions, as well as the reliability of a BLCM’s inferences (DiRenzo, Hanks and Miller, 2023). For these reasons ecologists should not consider model validation as a secondary or subsidiary type of analysis but should instead consider validation as a key part of model development.

Standard protocols on how to *present* ecological models are increasingly common, with the STARD-BLCM—that describes how to present BLCMs—already available. Despite this, within ecology, standard protocols describing how to *validate* the algorithms ecologists write remain a rarity (DiRenzo, Hanks and Miller, 2023). This disparity, while recognised in the ecological literature (Rykiel, 1996; Augusiak, Van den Brink and Grimm, 2014; Mouquet *et al.*, 2015), is at odds with common practice in other fields of research such as physics, mathematics, data science and beyond.

Before applying inferences from a BLCM to real world scenarios, it must first be established that the BLCM can produce credible inferences, i.e. inferences that are expected, and to which researchers are willing to assign confidence (Cordes, 1980). This is particularly important when models are custom-built, complex, when there are uncertainties in how a model has been specified (Wu and Li, 2006), or when models appear to be forcing inferences in accordance with the given prior information (Chivers, Leung and Yan, 2014). Published standards for the validation of ecological models in general are rare, and a literature search for such standards completed in March 2022 only uncovered

only two papers on the topic (Augusiak, Van den Brink and Grimm, 2014; Prowse *et al.*, 2016). For BLCMs, model validation should be a standard step in the inference-checking process, and to do this, the availability, accessibility, and use of standardised validation approaches is important. This is particularly crucial, given that differing implementations of a BLCM can impact how an MCMC sampler may interpret the same test array (Albert and Dodd, 2004).

Key to model validation is understanding whether the accuracy and precision associated with a BLCM's inferences is dependent on the true values of Se, Sp and P, and whether any trends in the reliability of inferences across parameter space exist and can be usefully generalised. The ability to evaluate positions of truths in parameter space is important for determining whether model identifiability changes relative to nearby truths, or whether certain volumes of parameter space lack identifiability altogether, and this information may direct how BLCMs should be applied. Identifiability issues can be described in terms of "practical identifiability", a measure of how confidently a BLCM can infer parameters given noisy data (Dendukuri, Bélisle and Joseph, 2010); as well as in terms of "structural identifiability", a prerequisite to practical identifiability describing whether a BLCM is able to infer the parameter values given error free data (Yates *et al.*, 2018). Reporting on any practical and structural identifiability issues are forms of model validation.

The two validation examples presented in this chapter explore error under two scenarios.

Validation Example A explores the error (Equation 16) associated with a disease testing scenario based around a defined single position in parameter space—described by the true values of Se, Sp and P—while **Validation**

Example B explores the error (Equation 16) across a random slice of parameter space when multiple sets of truths are randomly selected.

Relevant to both validation examples is the understanding that a simulation is a proxy for a defined position in parameter space—i.e., the true value of Se, Sp and P. The number of simulations describes how many times a BLCM has been replicated, and each replication may define a fixed or random set of truths, depending on how the BLCM has been specified. Note, the validation exercises presented do not analyse whether the precision of inferences of Sehat, Sphat and Phat is dependent on their respective true values, however the procedures described may also be applied to this dependent variable.

Methods

Three key building blocks underpin the Validation Examples presented. First, the hypothetical modelling scenario and its assumptions are described. Second, the purposes, models, and predictions for Validation Examples A and B are outlined. And third, the specifications of the Linear Mixed Effects Models (LMM's) that are used to analyse the response variable error are presented.

The hypothetical modelling scenario

The hypothetical modelling scenario considers theoretical ante-mortem diseased populations of 500 wild animals. True parameter values are provided for Se, Sp and P, and up to five different diagnostic tests are then used to infer their values via a BLCM. Within this framework, the effect of position in parameter space on the errors of Sehat, Sphat and Phat are explored across the two examples. Post-hoc analyses are then conducted using LMM's and likelihood ratio test procedures (see Table 10-5 to Table 10-12 for the results of these analyses), which are used to understand how error (calculated in

accordance with Equation 16) changes given position in parameter space, as well as the modelling conditions described.

In addition, the following assumptions were made when considering the hypothetical modelling scenario described:

1. Constrained prior distributions are a valid source of prior information.
2. Test one and test two parameters—i.e., Se1, Sp1, Se2, Sp2, the parameters common to all models simulating two to five tests—are assumed to behave similarly, though independently. This assumption was made to avoid additional noise within the random effects resulting from the potential for the accuracies of test one and test two parameters to be highly dependent on their respective positions, *and or* if one of the parameters within either test is consistently difficult to identify. Here on, and to avoid confusion, results are generally reported on using the general acronyms Se and Sp only; rather than Se1, Se2 and so on.
3. It is important that all truths are not assigned fixed values, since it has been proven that Se, Sp and P change as a function of many external biological factors (Begg, 1987; Greiner and Gardner, 2000).

Validation Example A: a basic validation simulation, based on a fixed point in parameter space, with replication.

PURPOSE: To explore the sources of error associated with a single position in parameter space. That is, the error associated with (i) the variation among independent replicate inferences; (ii) the posterior inference when each position in parameter space is manipulated by +/- 0.1; and (iii) the number of diagnostic tests.

MODEL: A fixed set of true parameters was inferred across a battery of five diagnostic tests. The initial set of fixed truths were randomly selected and are as follows: `P=0.4, Se1=0.81, Se2=0.71, Se3=0.66, Se4=0.52, Se5=0.59, Sp1=0.51, Sp2=0.56, Sp3=0.91, Sp4=0.94, Sp5=0.72.` Parameters Se, Sp and P are then, in turn, increased and decreased by a value of 0.1, creating six new sets of truths. Each scenario was inferred across 10 simulations. No prior distributions were constrained in Validation Example A, and the prior distributions used are as specified in Table 10-2.

PREDICTIONS:

(a) The error of posterior inferences is sensitive to small changes in parameters.

(b) If parameter space is fixed to a single point, this baseline can be used to detect changes in the errors of posterior inferences when this baseline is increased or decreased (by a value of 0.1) across a small volume of parameter space.

Validation Example B: a basic validation simulation, based on randomly selected points across parameter space, without replication.

PURPOSE: To understand how the accuracies of posterior inferences vary when multiple sets of truths are randomly selected across parameter space.

MODEL: 25 randomly selected truths for a battery of five diagnostic tests were drawn from parameter spaces bounded by zero and one, unless constraints were applied. For each truth, one set of diagnostic test outcomes were simulated, and used to infer Se, Sp and P. All random processes within the stochastic framework were seeded for repeatability. Note, given that the 25 truths are drawn across an 11-dimentional space, it was not considered that these samples could fully represent the variability in the accuracy of inferences

across parameter space, including at the edges of parameter space. Rather, the effect of prior precision and constrained true values on the magnitude of error at random positions in parameter space is investigated.

Several differences exist between the methodologies for Validation Examples A and B. In Example B, true values are randomly selected from parameter space using the uniform distribution to ensure that there are no intrinsic correlations among Se, Sp and P; and within the BLCM, the prior distributions are defined using two levels of standard deviations, creating precise (σ=0.05) and imprecise (σ=0.15) priors (see Table 10-2 for a full justification of these values). The variation in truth, and the extent of the prior information provided are also considered to be additional sources of noise when compared to Example A.

Note, constrained models are defined as models where the prior distributions informing inferences of P and or Sp are restricted in accordance with the justifications described in Table 10-2.

PREDICTIONS:

(a) A random choice of truths will make Se, Sp and P more difficult to infer—i.e., associated with greater errors—since volumes of high-dimensional parameter space may have complex regions of posterior density.

(b) Informative prior information will improve inference by directing the MCMC algorithm into more identifiable regions of posterior density.

(c) The errors of Se, Sp and P will be affected by the different levels of constraints applied in this chapter, which can be summarised as: unconstrained; constrained Sp; constrained P; and constrained Sp and P.

Analysing error using Linear Mixed Effects Models

LMM's have been identified within the work of this thesis as a useful means of interrogating the error attached to BLCM inferences, revealing the structure of the random effects that influence accuracy, and going some way towards explaining parameter-specific errors.

LMM's are a type of linear regression which can model fixed and random effects. Fixed effects are explanatory variables which describe an explicitly chosen treatment to investigate in comparison to other explanatory variables (Bennington and Thayne, 1994; Upton and Cook, 2014). Random effects are explanatory variables where the levels of effects under investigation have not been explicitly chosen (Bennington and Thayne, 1994). Random effects are not used to test the differences of values belonging to a hierarchy. In this chapter, random effects are used to understand the variance among values describing inferential error given the true values of Se, Sp, and P. To avoid confusion between an LMM and a BLCM, the word "model" is used in reference to BLCMs, and is not used in reference to LMM's.

The LMM's were written in R, using packages `lme4` and `lmerTest` (Bates *et al.*, 2015; Kuznetsova, Brockhoff and Christensen, 2017). Specifically, the LMM's were used to determine how the errors of Sehat, Sphat and Phat are affected by the number of diagnostic tests and prior information, by partitioning the noise associated with position in parameter space from the variation among simulations. Note, all scripts used to specify the LMMs used within this thesis can be found on https://github.com/annabush/PhD.

For Example A, noise due to position in parameter space is included as a random effect within LMM's using the relative truths of Se, Sp and P. In this

case, a relative truth is the difference between the manipulated truth and the original set of true values, and it can take the values 0.1, -0.1, or 0. For Example B, the true values are entered into the LMM as random effects.

Any residual random effects within all LMM's can be attributed to the variation of test arrays between simulations, and this can be affirmed by the fact that parameter identity is explicit within each LMM. Importantly, identifying between the noise attributable to position in parameter space, and the noise among the random processes of testing, allows us to define the sources of error from the LMM results.

Coding the Linear Mixed Effects Models in Chapter 4

For Example A, LMM's were specified in pseudocode as follows:

```
value ~ n.tests + (1 | P.rel) + (1 | Se.rel) + (1 | Sp.rel),
```

where the dependent variable "`value`" is the relevant absolute error for an explicit parameter, the variable "`n.tests`" is the number of diagnostic tests available, and the random intercepts denoted as `parameter.rel` indicate where relative truths have been inputted.

For Example B, LMM's were specified in the pseudocode as follows:

```
value ~ n.tests + prior.information + (1 | pi.truth) + (1 |
se.truth) + (1 | sp.truth),
```

where in addition to the parameters defined above, the variable "`prior.information`" describes the relevant prior precision or constraint, and the random intercepts denoted as "`parameter.truth`" indicate where true values have been inputted.

A note on the specification of random effects.

This note applies to Chapters 4, 5 and 6 of this thesis.

It is hypothesised that inference—including the "error structures" of an inference—may in some instances be strongly dictated by position in parameter space. For these reasons the identity of parameters needs to be accounted for within LMM's, so that the impact of truth on error, given the groups Se, Sp and P, could be evaluated without (a) needing to evaluate the specific true value, given that truths are randomly selected (Barr *et al.*, 2013) and known; and (b) conflating the variation among replicates with the differences among parameters.

The truths of Se, Sp and P were applied as crossed random effects since the groupings do not represent levels in a hierarchy. The continuous nature of the values that make up the truths of Se, Sp and P were ignored, since the biological meaning of each value was not being explicitly investigated. Instead, it is assumed that Se, Sp and P govern the variance structure of the variable `value`, that this structure can vary over three different intercepts, and that when interpreted together, Se, Sp and P can help diagnose infection.

Results

Table 10-5 to Table 10-12 show the full set of results for cross-referencing purposes. Table 10-5 to Table 10-7 present the results relevant to Validation Example A, and Table 10-8 to Table 10-12 present the results relevant to Validation Example B.

Validation Example A

(i) How do the number of diagnostic tests affect the errors of Sehat, Sphat and Phat?

Error generally decreases as the number of diagnostic tests increase (Figure 4-1 and Figure 4-2), with reductions in the errors of Phat (LMM1) and Sehat (LMM2) being particularly responsive to the number of diagnostic tests available. Likelihood-ratio tests confirm that the number of diagnostic tests available is not an important predictor of the magnitude of Sphat errors within the small region of parameter space explored; however Sphat errors are surprisingly small in two-test models (Figure 4-2).

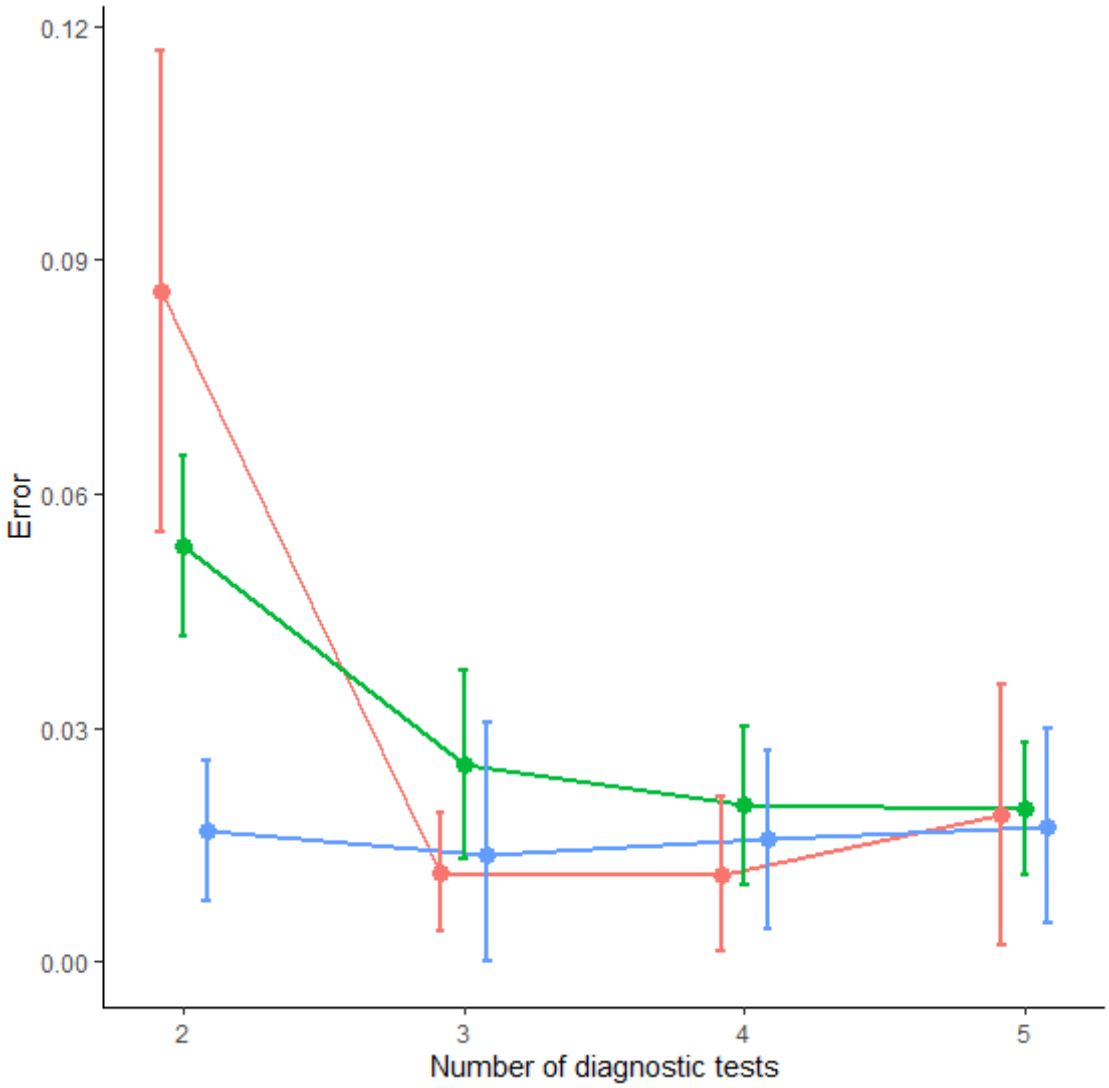

Figure 4-1: On average, the errors of Phat (red), Sehat (green), and Sphat (blue) generally decrease as the number of diagnostic tests available increase from two to five. This general trend is termed the “n.tests trend”. The error bars show the standard deviations of the mean posterior inferences. When five diagnostic tests are available, the errors of Phat, Sehat and Sphat are of similar magnitudes.

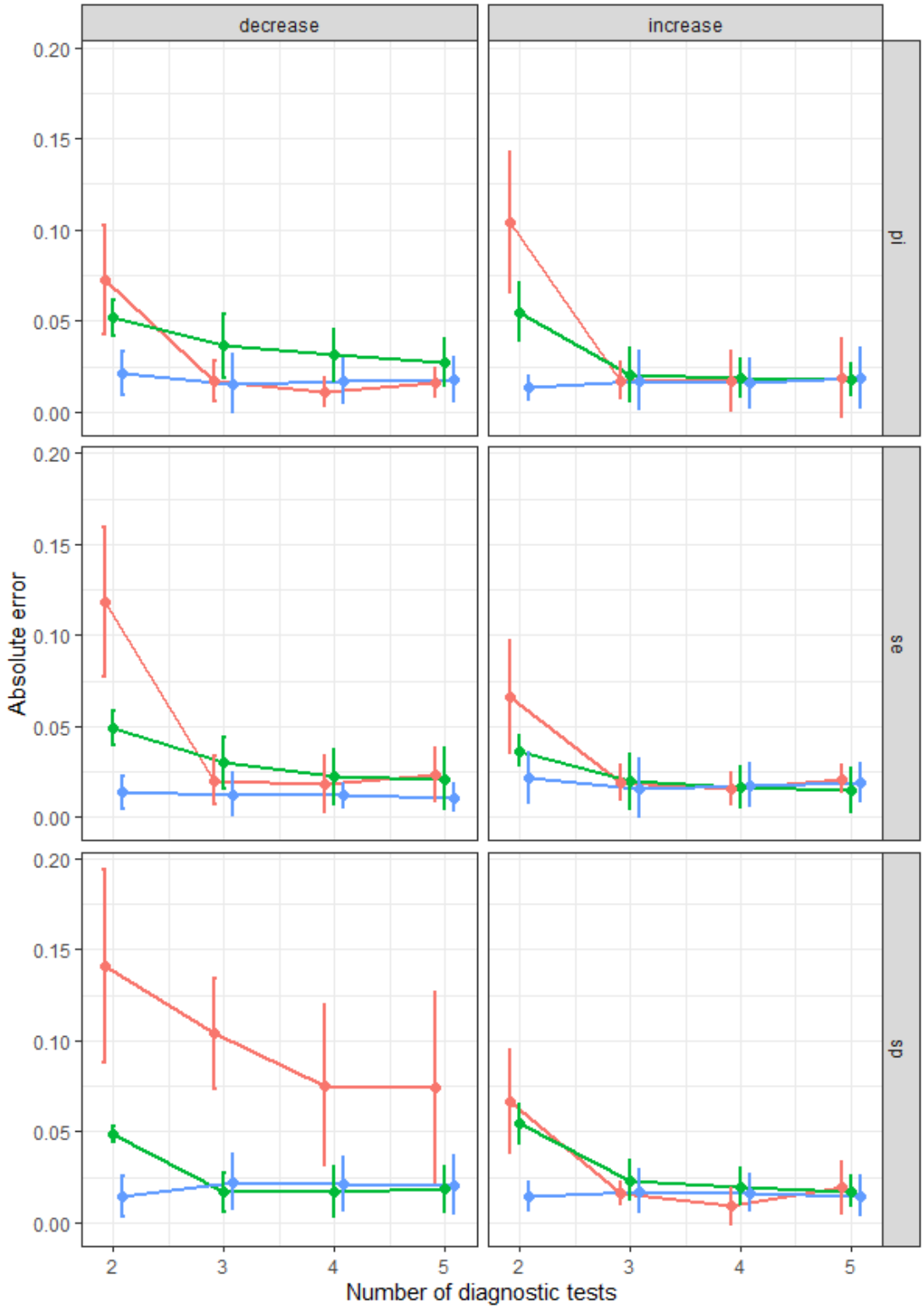

Figure 4-2: How the errors of Phat (red), Sehat (green) and Sphat (blue) change as the number of diagnostic tests available increase, when Se, Sp and P are increased or decreased by the value of 0.1 in comparison to the “original” true values. This baseline set of true values are as follows: `P=0.4, Se1=0.81,`

Se2=0.71, Se3=0.66, Se4=0.52, Se5=0.59, Sp1=0.51, Sp2=0.56, Sp3=0.91, Sp4=0.94, Sp5=0.72.

(ii) What are the dependencies influencing the errors of Sehat, Sphat and Phat?

The variance in the errors of Sehat, Sphat and Phat not explained by the number of diagnostic tests available is likely due to the variance among simulations. The variance not explained by the number of diagnostic tests available accounts for 83.8% and 66.1% of variance in Se1 and Se2 errors respectively, and 95.3% and 93.3% of variance in Sp1 and Sp2 errors, respectively (LMM4 and LMM5) providing confidence in the decision to restrict the reporting of analyses to Se1 and Sp1 only across Chapters 5 to 7 of this thesis. The average error and variation of error associated with Phat increases when the true value of Sp was decreased (Figure 4-2).

Validation Example B

(i) Checking how BLCMs respond to the information provided matters

Three key findings from the plotting exercises were:

1. The accuracies of Sehat, Sphat and Phat generally increase as the number of diagnostic tests increase, and this was found to be true across all the BLCMs specified for Example B (Figure 4-3 and Figure 4-4)
2. The accuracies of Sehat, Sphat and Phat are dependent upon model constraint (Figure 4-2)
3. Models provided with more precise prior information provide comparatively better inferences of Se, Sp and P than those that are supplied with imprecise prior information (Figure 4-4).

In addition, the regression analyses (see LMM 6 to 11 described in Table 10-8 to Table 10-9) suggest that:

1. The accuracies of Sphat are least affected by changes in the number of diagnostic tests.
2. The accuracies of Phat are least reactive to changes in model constraint.
3. The accuracies of Sehat are least reactive to changes in prior precision.

(ii) The effect of constraining true values on the errors of Sehat, Sphat and Phat.

The error of Phat is generally increased by the application of constraint; the errors of Sphat are only decreased when the values of both Sp and P are constrained; and the errors associated with Sehat are higher when Sp is constrained, and *vice versa* (Figure 4-3).

Perhaps surprisingly, the errors of Sphat increase with constraint (Table 10-8), yet the errors of Sehat and Phat generally decrease with constraint when either Sp, Pi or Sp and Pi are constrained (Table 10-8). To complicate matters, the errors of both Sehat and Sphat have an apparently significant relationship with constraint when both Sp and Pi are constrained (Table 10-8). Despite this, likelihood ratio tests show that only the errors of Sehat and Sphat are significantly affected by constraint (Table 10-9). Constraint appears to be particularly important for reducing the errors of Sehat and Sphat further to the reduction in error that can be achieved by increasing the number of diagnostic tests alone (Table 10-9).

(iii) The effect of the number of diagnostic tests used on the errors of Sehat, Sphat and Phat

As expected, the errors of Sehat, Sphat and Phat are all reduced as the number of diagnostic tests increases (Table 10-8, LMM's 6–8). In addition to this, it was found that prior information in the format of constrained or precise priors, consistently decreases the errors of Sehat, Sphat and Phat compared to when information is added via constrained or precise priors alone (Table 10-10). Despite this, the reverse situation—that information from diagnostic tests, *and* constraint or precise priors is always better than information from diagnostic tests alone—is not always true (Table 10-9).

For example, likelihood ratio tests show that Sehat is more accurate when prior precision and constraint are not provided, i.e. in situations where models are only informed by information from increasing numbers of diagnostic tests (Table 10-9). The number of diagnostic tests available to a model, *in addition* to any improvements in prior precision or constraint that is available, strongly

associates with a model's ability to accurately infer P, over and above its ability to accurately infer Se or Sp (Table 10-10).

(iv) The effect of changing the precision of normally distributed priors on the errors of Sehat, Sphat and Phat

The errors of Sehat, Sphat and Phat generally decrease when precise priors are used (Table 10-9), and becomes less parameter-specific, i.e. different between inferences of Se, Sp and P. Precise priors seem critical for the inference of P and Sp, over and above the reductions in error that can be attributed to increasing the number of diagnostic tests alone (Table 10-9).

(v) The effect of the information supplied, and the position in parameter space, on the errors of Sehat, Sphat and Phat.

When BLCMs are informed by precise priors, the errors of Sehat, Sphat and Phat are least affected by the position of P (Table 10-11). This is not the case when constraint is used to inform the BLCM: in this case, the errors of Phat are least affected by the position of Se; the errors of Sehat are least impacted by the position of P; and the errors of Sphat are least affected by the position of Sp (Table 10-11).

Further when precise priors are used, the random variance within the LMM's used is dominated by variation across simulations, accounting for 74.3% of the random variance of Phat errors; 56.6% of the random variance of Sehat errors; and 42.7% of the random variance of Sphat errors (Table 10-12). It appears that random variance is also dominated by variance across simulations when models are informed by constrained priors (Table 10-12).

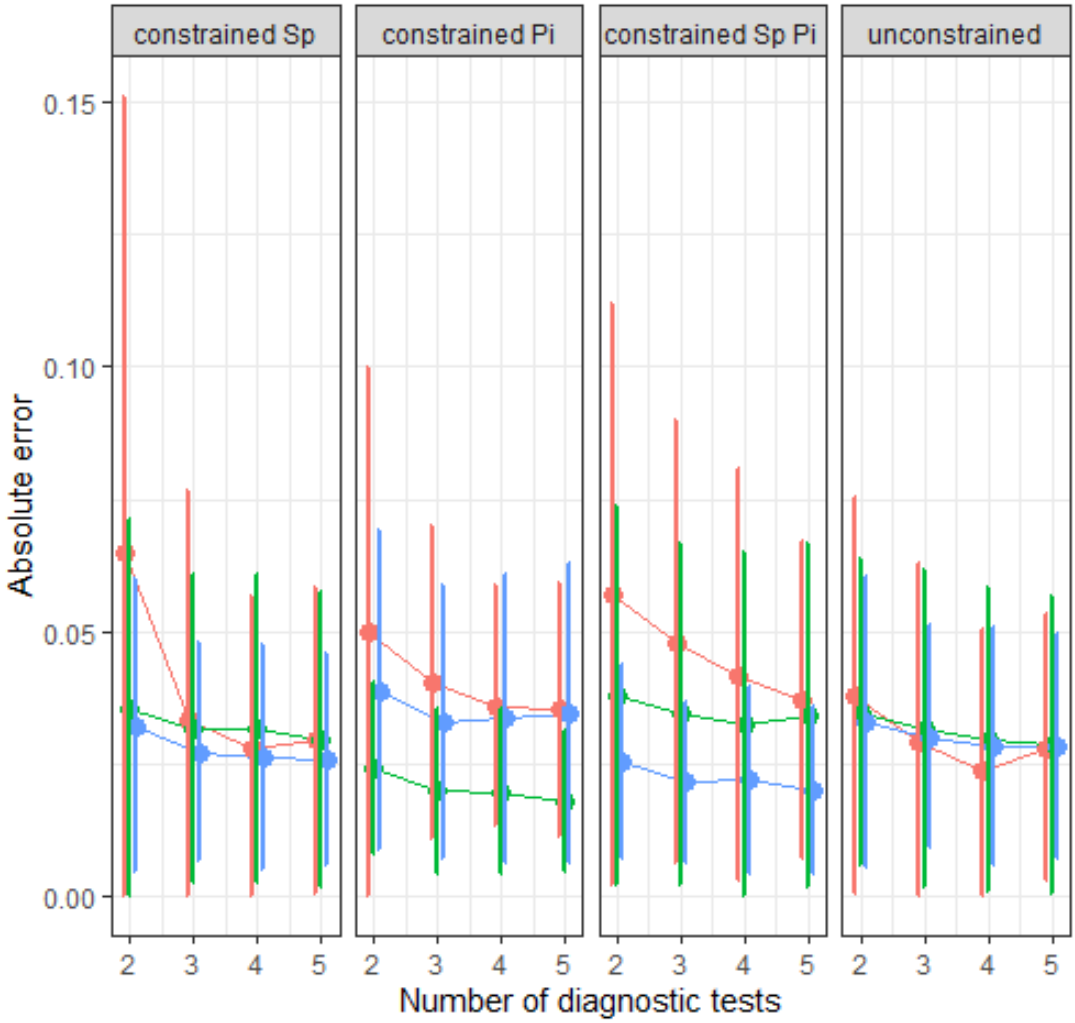

Figure 4-3: How the absolute error of parameters Phat (red), Sehat (green) and Sphat (blue) change over number of diagnostic tests in scenarios where given truths are either unconstrained or constrained.

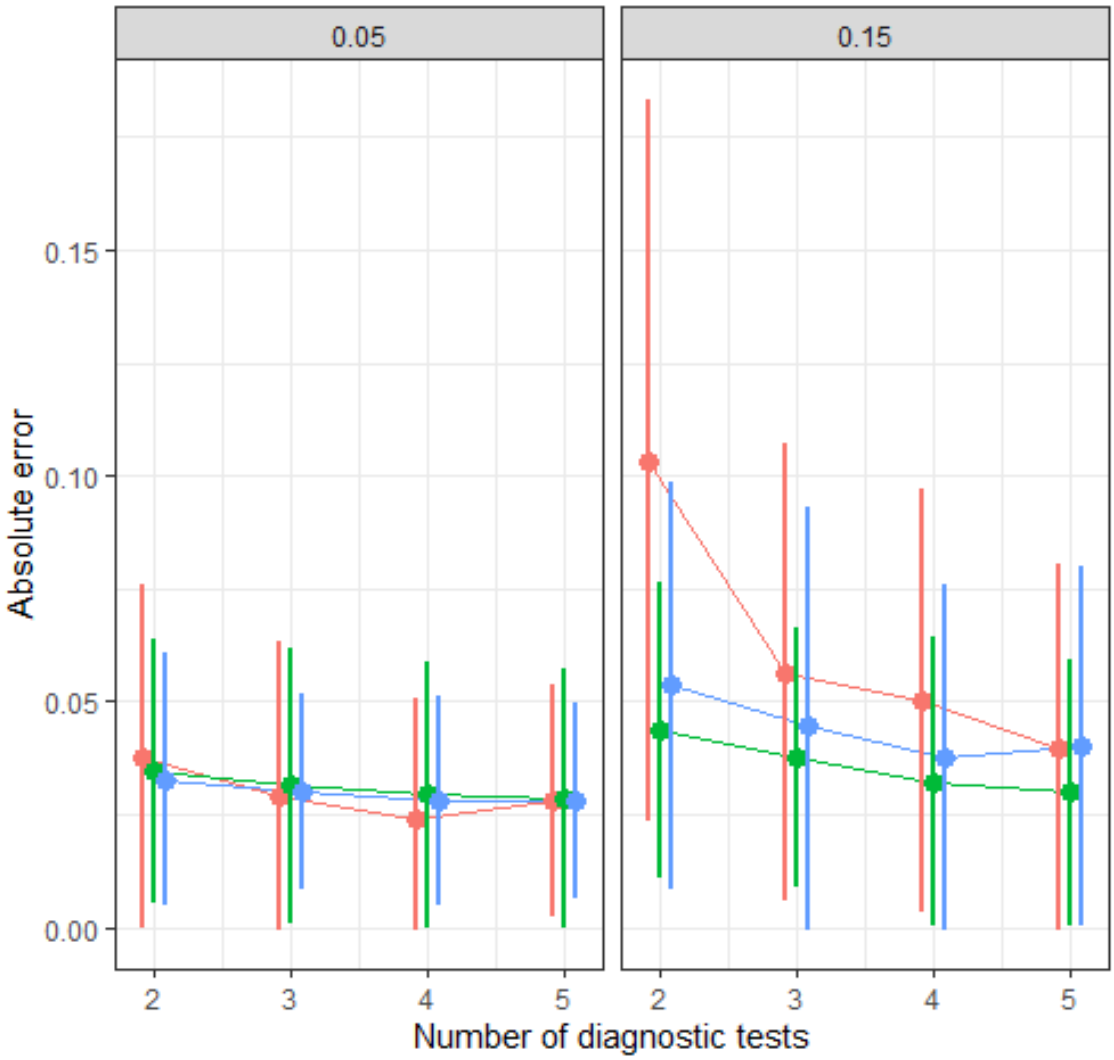

Figure 4-4: How the errors of Phat (red), Sehat (green) and Sphat (blue) change over the number of diagnostic tests available when prior precision is precise compared to when it is imprecise.

Discussion

Although it is widely accepted that data from multiple diagnostics can improve existing reference standards via Latent Class Analyses (Rydevik, Innocent and McKendrick, 2018; de Bronsvoort *et al.*, 2019; McAloon *et al.*, 2019; O'Hagan *et al.*, 2019), the Latent Class Models that demonstrate this are rarely validated across a wide range of diagnostic outcomes and true parameter values (Hobbs and Hooten, 2015; DiRenzo, Hanks and Miller, 2023). Before inputting real

data, a careful validation of BLCMs using simulated data is critical to the ability to confidently infer parameters (Bermingham *et al.*, 2015) and to establish useful priors (McAloon *et al.*, 2019). Accordingly, this section outlines seven key findings that emerged from the two model validation examples demonstrated in this chapter, with each finding being presented in the form of a stylised fact.

It is recognised that the results presented within this chapter are specific to the small number of truths that were tested, and that consequently, the noise among simulations was the greatest source of noise among the random effects. Further, it is recognised that some regions of parameter space may be over-sampled during random simulations while others may be under-sampled. These two aspects of study design are addressed in the following chapters. Despite this, Validation Examples A and B demonstrate that error does depend on position in parameter space.

This discussion was written after reflecting on a remark by Agresti, who writes that "*A danger with latent variable models… is the temptation to interpret latent variables too literally…One should realize the tentative nature of the latent variable. Be careful not to make the error of reification – treating an abstract construction as if it has actual existence*" (Agresti, 2003). Model validation on simulated data helps us to understand *how literally* latent variables may be interpreted, and how much weight a theoretical BLCM should have in the real world.

While latent variables such as P, Se and Sp have a greater likelihood of representing the truth when prior information can be verified (Schofield *et al.*, 2021)—i.e., if the precision of the prior information can be evidenced—this verification exercise is complex, and particularly in the field of wildlife disease, is also open to debate, and so the problem of "better representing the truth" must

often be answered post-inference. By validating theoretical BLCMs under a variety of testing situations, the practical limitations of a BLCM can be better understood.

In short, it is possible to offer ecologists an insight into the accuracies of Sehat, Sphat and Phat summarised as seven stylised facts—broad conclusions, that in line with Agresti's considerations, generalise over the simulations concerned.

STYLISED FACT 1: Seemingly successful inferences in two-test scenarios may simply be due to the posterior replicating the prior, and so should be treated with caution.

Although the posterior distribution does not seem to precisely replicate the prior distribution in a two-test model, there was much less variation among simulations in the two-test scenario, and practical identifiability cannot be assumed. This finding was expected, since according to Hui and Walter (*op. cit.*), it is not possible to estimate 5 parameters, and 4 potential test outcomes, with only 3 degrees of freedom (Table 3-1).

This observation highlights an important point: beware of successful inference in two-test BLCM scenarios since the posterior may just be replicating the prior, and the associated inference may be false. This is the most probable reason why Sp is successfully inferred in two-test models, and highlights that for real-world applications an understanding of exactly how prior information creates a "successful" inference, particularly for a two-test model, is critical.

Importantly, non-identifiability in two-test models was not observed in Validation Example B, where some prior information was provided to every model, suggesting that appropriate prior information can aid model identifiability in two-test models.

STYLISED FACT 2: Unidentifiable areas of parameter space may occur where error does not decrease when the number of diagnostic tests available for inference increase.

Validation Example A shows that the errors of Sehat, Sphat and Phat decrease over increasing numbers of diagnostic tests. And this trend persists despite the model being provided by randomly selected truths in example B.

This trend is both under-researched and important to note for two reasons. First it is hypothesised that when this trend breaks down, parameter space may be unidentifiable. Second, reductions in the errors of inferences attributable to increases in the number of available diagnostic tests are separate to, and can be in addition to, reductions in the errors of inferences from the provision of prior information.

STYLISED FACT 3: Increasing the number of diagnostic tests has the greatest effect on decreasing the error of Phat.

The finding that increasing the numbers of diagnostic tests markedly reduces the error of Phat indicates that proxy tests may be a simple way for ecologists to optimise BLCMs. While the concept of a “third opinion” as a proxy test to aid model identifiability has already been discussed (Dendukuri, Bélisle and Joseph, 2010), as well as the use of anecdotal proxy tests in animal disease research (Leeflang *et al.*, 2013), ecologists have ready access to a wealth of information—such as opinion on infection presence or absence from experts, documentations of historic infections, or research on geographically separate reservoirs of a pathogen of interest—that could be used to develop a proxy test.

Validation Example A demonstrated that relative to the errors of Sehat and Phat, the errors of Sphat are the least reactive to changes in the number of

diagnostic tests across the chosen true values of Sp that range from 0.51 to 0.94. This finding—that the numbers of diagnostic tests available is potentially not a critical dependency of the errors of Sphat—is also replicated in Example B when truths were randomly selected. This is suspected to be a statistical issue—it is likely that constraining Sp values ensures the MCMC algorithm remains in useful parameter space (see stylised fact 4), and that in comparison to inferring P, solutions to the values of Se and Sp (to satisfy Equation 14) prove more difficult for the MCMC algorithm to find.

For most wildlife diseases, tests with suitably high values of Sp—i.e., values close to 1—are rarely available, but due to the P of wild disease in animals usually being low, the need for most regimes to identify true negative cases is high. This means that optimising the inference of Sphat is arguably more important than optimising the inference of Sehat. The theory that minimising the errors of Sphat is particularly important when sample size and P are both low is both logical, and in agreement with recent research by Helman *op cit*.

STYLISED FACT 4: Prior constraints are particularly important for reducing errors associated with Sphat over and above the reduction in errors associated with increasing the number of diagnostic tests.

In wild animals, many diseases persist with low values of P—such as Bovine viral diarrhoea virus (Casaubon *et al.*, 2012) and Brucellosis (Godfroid *et al.*, 2005)—and it is reasonable to assume that most endemic wild diseases infect less than half of the population at any one time. This means that most sampled individuals are true negatives, and implies that the ability to reduce the errors of Sphat in preference to reducing the errors of Sehat is sensible.

While Validation Example B demonstrates that constraint is a comparatively important source of prior information for accurately estimating Sp compared to Se or P, a qualitative analysis of Figure 4-3 suggests that this trend may not be uniform across parameter space: the error of Sphat seems to only decrease by constraining P too.

Overall, the influence of constraint on error seems less significant than the effects on error given more diagnostic tests, indicating that increasing the number of diagnostics is a more powerful way to improve the accuracies of inferences. Nevertheless, prior constraints remain a valuable source of information, since they can reduce the parameter space (Hobbs and Hooten, 2015) in which the MCMC algorithm must search. In fact, Berkvens *et al.*, 2006 suggests that the only way to accurately estimate P is by introducing external knowledge through constraint. When researchers can speculate on the parameter space in which the truth likely lies by offering the BLCM broad constraints—for example, give or take 25%—Sp estimates can be improved, and in turn, those of P.

STYLISED FACT 5: Prior precision is particularly important for reducing errors associated with Phat and Sphat in addition to the reduction in error from increasing the number of diagnostic tests.

Prior precision is a valuable source of prior information for accurately inferring P, and unlike constraint, it has the same magnitude of effect on the error of Phat as increasing the numbers of diagnostic tests available. However, increasing prior precision decreases the error of Sphat more than increasing the number of diagnostic tests.

Despite this, the errors of Sehat seem the least responsive to the provision of precise prior information, suggesting that for Validation Example B, Sehat is both difficult to provide useful prior information for, and highly dependent on prior information. This finding is substantiated by Liu *et al.*, 2014, who find that the prior of Se is more important than that of Sp: and this in turn could be because diseased animals are, in general, less likely to be recorded in test arrays than healthy animals. Despite this, ecologists should be mindful that providing precise priors is not as important as maximising the number of diagnostic tests if the model does not have the minimum degrees of freedom that it requires.

STYLISED FACT 6: The errors associated with Sehat are inversely proportional to the errors associated with Sphat.

The Se-Sp trade-off, or reciprocal relationship, is well cited, and its properties—that describe the ability of a given test to determine “noise” from “signal plus noise” (Green and Swets, 1966)—have particularly important implications for the classification threshold of a positive or negative test, as well as the power of a BLCM. Since the 1970’s it has been known that the properties of Se and Sp are not stable (Ransohoff and Feinstein, 1978), and further insights into the trade-off remain of inherent value to disease researchers.

Insights into a second trade-off that exists between the accuracies of Sehat and Sphat, belonging to a given test, have not seemingly been published. Accordingly, this chapter reports on five underlying trends concerning the errors of Sehat and Sphat:

1. Constraining P reduces the error of Sehat but increases the error of Sphat.

2. Constraining P and Sp increases the error of Sehat and reduces the error of Sphat.
3. Prior precision has more influence on the errors of Sphat than constraint.
4. Constraint has more influence on the errors of Sehat than prior precision.
5. Increasing the number of diagnostic tests reduces the errors of both Sehat and Sphat.

While the above five tendencies may be unique to Validation Example B, they provide an insight into how to prioritise the prior information that a BLCM should be given to maximise its model power, as well as information useful for the calibration of cut-off thresholds.

STYLISED FACT 7: Phat is particularly difficult to infer when Sp is low.

Understanding how the errors of Phat can be biased is fundamental to wildlife disease ecology, particularly since diagnostic accuracy is dependent on P (Brenner and Gefeller, 1997; Gardner, Johnson and Norris, 2009). A key output from Validation Example A was the finding that variance in the errors of Phat is largely explained by variance in the values of Sp.

While the existence of a P-Sp trade-off has not been frequently cited, it sits in agreement with both medical literature (Leeflang *et al.*, 2013) that reports "*differences in prevalence mainly represent changes in the spectrum of people without the disease of interest*", as well as wildlife disease literature (Helman *et al.*, 2020), which reports that the optimal Se and Sp of a superior test is dependent on P. It is probable that there is a further but potentially less important P-Se trade-off among the other recognised trade-offs between Se and Sp, as well as between P and Sp.

Conclusion

What is a model validation? How should it be done? And can we learn more about how a BLCM may infer Se, Sp and P across parameter space? By addressing these questions in turn, this chapter responds to the lack of information on how to validate BLCMs and shows that model validation can be used to provide important insights into how latent parameters may be inferred. It is demonstrated that the dependencies between the accuracies of Phat, Sehat and Sphat, and the given modelling conditions, are complex.

The chapter should, however, be interpreted within the context of its limitations, which are chiefly a result of the small number of simulations and truths studied. It is hypothesised that variation in test outcomes, which creates noise among replicate simulations, should become less significant as the number of simulations increase. In turn, it is expected that the effect of position in parameter space will become more prominent as the number of simulations increase, and this has the potential to change or verify the trends described in this chapter.

That said, it is important to note that this present chapter has not found unidentifiable parameter space in modelling situations where the degrees of freedom rule has been satisfied, indicating that larger studies—embracing more simulations and more truths—would be required to locate these volumes of parameter space (if they exist).

Are the seven stylised facts presented generalisable? Do instances where parameter accuracy (and precision) improve with the number of diagnostic tests serve as a proxy to illustrate where the practical identifiability of a BLCM is

possible, and where it is not possible? When is a model identifiable?

Subsequent Chapters 5, 6 and 7 proceed to test these hypotheses.

Chapter 5

5. When are BLCM inferences uncertain?

Introduction

Critical to this chapter is the distinction between the metrics of BLCM performance—or power—calculated in terms of error, bias, and standard deviation; and the metrics of a diagnostic test, inferred as Sehat and Sphat.

This chapter concerns a simple problem that a researcher may wish to ask about parameter space: "*is there a region of my parameter space where condition X holds?*" (Chalom and de Prado, 2012); and responds to the same question as *op. cit*. Chalom and de Prado pose, where in this case condition X is the question of practical identifiability. To achieve this, the model validation methodology described in Chapter 4—which aimed to establish whether a BLCM can infer theoretical scenarios as expected, and for the correct reasons—is expanded to explore where BLCM inferences are uncertain across a wide range of possible diagnostic testing scenarios.

Specifically, this chapter expands on two key findings of Chapter 4. First, the finding that it is important to examine the error structures of simulation analyses, because the accuracies and precisions of inferred parameters are variable, even when small volumes of high-dimensional parameter space are studied. And second, the finding that the number of diagnostic tests available is a key driver of the performance of BLCMs, but that other drivers exist. One of these drivers is the apparent relationship (see stylised fact 6) between the inferred values of Se and Sp in terms of their respective accuracies. The aim of this

chapter is to develop understanding around combinations of parameter values that may lack practical identifiability (Munch, Poynor and Arriaza, 2017), specifically as a contingent of the presumed relationship between Se and Sp, which is now hypothesised to be a key artefact—i.e., a trend explainable by statistics rather than ecology—defining BLCM performance.

For a single diagnostic test, the widely reported reciprocal relationship between Se and Sp (for example, see Shreffler and Huecker, 2023) is a commonly cited reason for why a gold standard cannot be attained, and is also used to explain the reason why thresholds for positive diagnoses are disputed when serological data is used. Importantly, the work presented in this chapter does not repeat the creation of a classical Receiver Operating Characteristic (ROC) curve for *single* diagnostic tests (explained in Appendix 2: Key parameters, hyperparameters and functions), this is because the focus is on understanding the relationship between Se and Sp in the context of the better understanding of *batteries* of diagnostic tests, i.e. when two or more diagnostic tests are available, a critical consideration for ecologists wishing to adopt a BLCM approach.

To expand, the relationship between Se and Sp for single non-gold diagnostic tests is commonly represented on ROCs—i.e., plots showing the estimated Se and Sp for all cut-off values (Fischer, Bachmann and Jaeschke, 2003); with an assumed proportionately inverse relationship between Se and Sp, given that the parameters are normally distributed.

To better understand how identifiability changes across regions of parameter space, the statistical artefacts present need to be distinguished from the ecological artefacts—such as population level disease traits, or the dependency between the behaviour of diagnostic tests and the stage of disease (Ransohoff and Feinstein, 1978)—that the data given to a BLCM represents (Hallman and

Robinson, 2020). The accessibility of methods to explore high-dimensional parameter space is key to uncovering statistical artefacts and presents a problem relevant to the whole of systems biology (Vernon *et al.*, 2018). For this thesis, representing high-dimensional parameter space is critical for understanding where inferred parameters lack identifiability across a wide range of possible diagnostic testing scenarios.

BLCM performance can change with any factor that may also alter model identifiability, which could include the degrees of freedom available to the model, or the sample size available to infer the required statistics in real-world studies—particularly when the conditional independence of results is assumed (Dendukuri, Bélisle and Joseph, 2010). Importantly, model performance and the identifiability of posteriors are related but not mutually exclusive since it is possible for a BLCM to generate posteriors that do not identify the latent parameters. And for clarity, model identifiability is not explicitly quantified in this chapter; rather, the relationships between indicators of BLCM model performance are defined, and this information is used to question practical identifiability.

Validating models across high-dimensional space has been termed “uncertainty analysis” (Volodina and Challenor, 2021), which is a term adopted in this thesis, and is used to describe the variation in BLCM outputs given variation in BLCM inputs. In the present chapter, these measures of uncertainty are the errors Equation 16) and standard deviations of BLCM posterior inferences, as well as the global statistics, across a larger volume of parameter space than that explored in Chapter 4. These statistics are used to develop a series of heatmaps and regression analyses that represent the resulting uncertainty across parameter space as six conditions shown in Table 5-1 are varied.

These two methodologies—the development of heatmaps to visualise the uncertainty associated with predicted values across parameter space, and the specification of LMMs to quantify hypothesised dependencies between uncertainty, and modelling conditions (Table 5-1)—are used to indicate where practical identifiability exists across a range of infection scenarios. Identifiability issues established, researchers can then ascertain when tests stop being useful, when the interpretation of inferred parameters becomes tricky, and look for any resulting statistical artefacts. In this chapter, attention is given to a specific artefact—the relationship between Se and Sp across batteries of tests—which is hypothesised to drive identifiability issues.

Accordingly, this chapter examines the following three questions, which are underpinned by lower-level findings from the analysis of heatmaps and regression models.

1. Does the tendency to overestimate or underestimate Sehat and Sphat depend on the true value of Se, Sp and P?
2. Does the size of the absolute error of Sehat and Sphat depend on the true value of Se, Sp and P?
3. Does the standard deviation of the posterior inferences of Sehat and Sphat depend on the value of Se, Sp and P?

Ultimately, these questions are used to examine the hypotheses that:

1. Specific volumes of parameter space are associated with specific “uncertainties”, which may indicate identifiability issues.
2. The relationship between Se and Sp across batteries of tests is associated with identifiability issues across parameter space.

Methods

The hypothetical modelling environment

The overarching assumption within this environment is that if diagnostic accuracy varies across populations, diagnostic tests must be influenced by the heterogeneities between diseased populations. As a result, it is assumed that the position of any true value within parameter space will be associated with a parameter-specific error. Further, based on this logic, it is assumed that the errors of Sehat, Sphat and Phat will not improve uniformly across parameter space when the six conditions (detailed in Table 5-1) are applied; for example, when the number of diagnostics increase. Note, given that only one population is studied at a time in the hypothetical modelling environment relevant to this chapter, the assumptions of the Hui-Walter model are not violated.

A further 12 assumptions help to define the hypothetical modelling environment, and are as follows:

1. The error of global statistics—i.e., a mean statistic of Phat, Sehat and Sphat—is independent from the choice of truth, the error of an MCMC sampler, or model identifiability issues. This error describes the average error of Sehat, Sphat and Phat given any volume of parameter space.
2. Uncertainty is dependent on infection scenario.
3. In a parameter space, regions can be identifiable, other regions can be non-identifiable; and identifiability can be inferred.
4. The relationship between the errors of Sehat and Sphat influences practical identifiability.
5. Uncertainty can be quantified by proxy via understanding the accuracies (and precisions) of Sehat, Sphat, Phat and the global statistic.

6. It is hypothesised that the probability scale may be hard to trust when representing data on heatmaps.
7. It is hypothesised that while the relationship between Se and Sp may be reciprocal, it is complex.
8. It is possible to optimise the estimation of Se and Sp using batteries of diagnostics.
9. Diagnostic accuracy is not stable across testing environments.
10. The relationship between Se and Sp given P is not stable.
11. The ultimate reason behind variations in diagnostic accuracy is due to ecological factors, but first statistical artefacts must be known.
12. Based on the findings of Chapter 4, diagnostic tests one and two have similar variances, and so inferences for both diagnostic tests one and two do not need to be reported.

The truths of Se, Sp and P were systematically set to vary across simulations, while the truths of Sp for tests two to five, and the truths of Se for tests two to five, were set to known fixed values that remained the same across all simulations. The truths are as follows: `Se [2:5] = 0.71, 0.66, 0.52, 0.59` and `Sp [2:5] = 0.56, 0.91, 0.94, 0.72`, and were randomly chosen whereas `P = {0.05, 0.1 … 0.45}, Sp [1] = {0.55, 0.1 … 0.95}` and `Se [1] = {0.05, 0.1 … 0.45}`.

True values of P and Sp were restricted—as a separate process to applying prior constraints—to reflect a realistic modelling scenario. Values of P were restricted to between 0 – 0.5, and values of Sp to between 0.5 and 1. These restrictions reduce parameter space by a quarter of its former volume (from `10 x 10 x 10` to `10 x 5 x 5`) and create a smaller volume of posterior density for the MCMC algorithm to search within. These restrictions also ensure that

truths remain identical between scenarios where priors are either unconstrained or constrained, and that the volume of parameter space searched in every modelling scenario is equal.

To expand, constraints on the true values of Sp and P were applied using an assumption that most sustained wildlife infections have P of below 50%, and tests should be constrained to tailor to a scenario where most individuals—i.e., the greatest proportion—are not usually diseased (see Table 10-2 for full justification). This means that when Sp is constrained, its true values are limited to those greater than 0.5, and when P is constrained, its true values are limited to those less than 0.5.

Simulated dataset 3 (Table 10-1) contains results for four modelling scenarios, where each `0.1 x 0.1 x 0.1` voxel of parameter space is replicated 10 times:

1. Normal priors, constrained priors.
2. Normal priors, unconstrained priors.
3. Uniform priors, constrained priors.
4. Uniform priors, unconstrained priors.

Within each modelling scenario, the number of diagnostic tests, and the sample size of the target population was varied. Within modelling scenarios that use normal priors, prior precision was varied.

This modelling setup is specified to inform the Any-Test, Any-Population BLCM as described. In comparison to the BLCMs specified for Chapter 4, the BLCMs used in Chapter 5 are modified to allow the influence of six conditions (Table 5-1) on the accuracies of Sehat, Sphat and Phat to be tested.

Table 5-1: Modelling conditions referenced in Chapters 5, 6, and 7, and the levels of each condition. The rationale behind the levels chosen can be found in Table 10-2.

Condition	Levels
The number of diagnostic tests	2, 3, 4, 5
The sample size of the target population	500, 1000, 1500
The prior precision of informative priors	Imprecise, Precise
Uninformative and informative priors	Normal, Uniform
Constrained priors	Constrained, Unconstrained
Edge of parameter space	TRUE, FALSE

The "15% scenario" and its rationale

Note that the results for what this chapter terms the "15% scenario" were extracted from the four modelling scenarios described in the previous section.

Consider a scenario where 15% of individuals are infected. This is a typical modelling scenario in ecology, as most individuals within wild animal populations are expected to be healthy. For this situation, Sp must be maximised in preference to Se in order to avoid misclassifying more of the most abundant class of individuals, namely the uninfected (Rydevik, Innocent and McKendrick, 2018). However, as Sp is maximised, the behaviour of Se also

changes in response, and so the most suitable values of both Se and Sp need to be chosen.

While the correlation of diagnostic tests in terms of false positive and false negative rates is a well-known probability problem, no known study has evaluated the nature of the hypothesised relationships between Se and Sp when using the BLCM approach. To do this the 15% scenario is analysed—using a subset of data where P = 0.15 extracted from the four scenarios listed above—and compared with the results across parameter space where p assumes a range of values between 0 and 0.5. The purpose of this study is to test how “generalisable” the trends outlined in Chapter 4 are, using a probable modelling scenario.

Generating representative truths using grid sampling

Grid sampling is a methodology developed within this thesis to ensure that true values can be sampled across a parameter space without the introduction of sampling bias.

The BLCMs simulate batteries of diagnostic tests of up to five tests, and so a method to systematically sample across 11-dimensional parameter space—sampled in accordance with any constraints applied to true values—was required, in order to ensure that every `0.1 x 0.1 x 0.1 (Se1 x Sp1 x P)` space contains a defined number of simulations. Estimating probability distributions across 11 dimensions—and in a study demanding high levels of replication—is challenging, since the number of possible sequences that an MCMC sampler may follow grows exponentially as sequence length increases; and with the increasing dimensions come consequent reductions in the ecologically relevant space (Chen *et al.*, 2020). As a result, the estimation

problem must be (i) simplified without introducing further bias; (ii) a parameter space that can be efficiently traversed by the MCMC algorithm; (iii) and a parameter space that can be fully represented by the given parameters.

To do this, the simulations required a systematic way of generating true values for Se1, Sp1 and P. To be clear, the word systematic applies to the method—termed in this study grid sampling—used to ensure that every possible parameter value of each parameter does not need to be sampled; this experiment would be termed a “complete parameter space exploration”. Further, a random sampling approach was not adopted since it could introduce bias by oversampling some regions of parameter space and undersampling others.

When solving Equation 14 across 11-dimensional space it is advisable to hold parameters constant (Yang and Atkinson, 2008) in order to avoid having to simulate the entire volume of parameter space: and so, a variation on a method called Individual (Chalom and de Prado, 2012) or singular Parameter Perturbance (Watts, 2008) was explored.

A parameter space exploration of only test one parameters was done using the grid sampling approach, which simplified the full parameter space exploration by using “cells” to discretise, i.e. subsect the sampling problem, and generate areal, i.e. gridded data. The truths of parameters belonging to tests two to five were fixed to specific values as described. As well as implementing elements of Individual Parameter Perturbation, the grid sampling method also used a key aspect of Latin Hypercube sampling (McKay, Beckman and Conover, 2000) as it ensured that each cell was sampled with an equal intensity.

The grid sampling technique was developed to satisfy three criteria:

1. The sampling method must be capable of sampling across 11-dimensional space, and the consequent results must be able to be fixed across eight dimensions since it is only practical to represent up to three dimensions within a schematic.
2. True values of parameters must be fixed to ensure computational efficiency. This is because the amount of “noise” associated with 11 varying parameters is large, so a trade-off presents itself: with more varying parameters comes more noise, and this in turn demands more replication within the BLCM.
3. To enhance computational efficiency, P and Sp must be restricted as separate processes to applying constraints to the BLCM, in order to ensure that the results between scenarios are both relevant to most wildlife disease scenarios, and also directly comparable (see Figure 7-1).

Parameter space is divided into a 3D grid of $(N_P \times N_{Se_1} \times N_{Sp_1})$ voxels of width 0.1, where N_P is the number of voxels in the direction of P, N_{Se_1} is the number of voxels in the direction of Se1, and N_{Sp_1} is the number of voxels in the direction of Sp1. For all experiments in this chapter, N_P was given a value of 5 to create 5 voxels between 0 and 0.5, N_{Sp_1} was given a value of 5 to create 5 voxels between 0.5 and 1, and N_{Se_1} was given a value of 10 to create 10 voxels between 0 and 1. At each voxel, simulations were replicated 10 times.

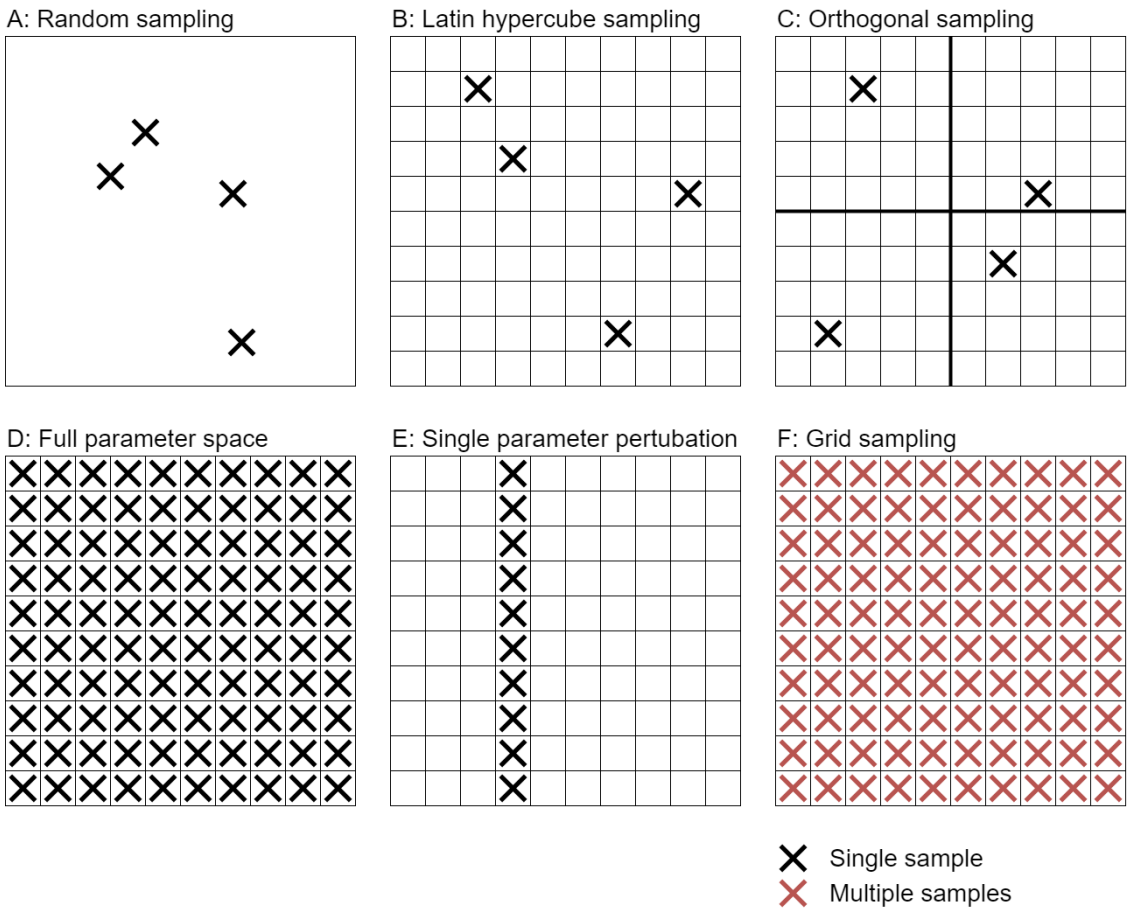

Figure 5-1: A schematic showing the sampling methods considered when selecting and developing the grid sampling method.

Manipulating the parameter space data using the special.melt functions

The challenge of manipulating 11 dimensions of simulated data into a format suitable for generating heatmaps of parameter space and specifying regression models was overcome by the creation of two complex functions—called `special.melt` and `special.melt2`, available at https://github.com/annabush/PhD—to manipulate and store data generated using normal and uniform prior distributions respectively. In short, `special.melt` and `special.melt2` initialise dataframes of the right number of columns, rows and dimensions to automatically organise the outputs of the MCMC sampling. The key difference between `special.melt` and `special.melt2` is the dimensions of the dataframe requiring initialisation, as

data generated using informative priors corresponds to a level of prior precision, and therefore requires an additional dimension of storage.

The general workflow to manipulate data for use in LMM's and heatmaps is as follows. This workflow allows a single dataframe to be generated, with columns for all variables that may need to be called, and rows for each observation. The size of the single dataframe containing the results analysed in this present chapter is 180,000 observations by 57 statistics.

1. Set the libraries and working directory.
2. Define `special.melt` and `special.melt2`.
3. Load results files in R Data File format from the required simulated dataset Table 10-1 and reformat the results using the `special.melt` functions.
4. Combine dataframes by row using the `rbind.fill` function of the `plyr` package (Wickham, 2011), which fills missing columns.

Generating the heatmaps of parameter space

The purpose of the heatmaps was to visualise high-dimensional numeric data, and to do this 11-dimensional space was condensed into three dimensions `P x Se1 x Sp1`. Heatmaps were used in order to explore the sensitivity of the response variables—error (Equation 16), bias (Equation 17), and standard deviation—across parameter space, which is graded according to the error specific to a `0.1 x 0.1` grid cell. This type of analysis is common when plotting landscapes of gene expression, which inspired the `shinyheatmap` package (Khomtchouk, Hennessy and Wahlestedt, 2017), though in this case heatmaps were created using the package `ggplot2` (Wickham, 2014) with the

generalised plotting code provided below, and full plotting manuscripts found on GitHub (https://github.com/annabush/PhD).

Overall, the bias, error, and standard deviation of inferences of P, Se, Sp—or inferences made regarding the global statistic—were plotted across matrices of heatmaps using the following facets via the `facet_grid` function of `ggplot2`:

1. Prior precision.
2. Constraint given normal data.
3. Constraint given uniform data.
4. Number of samples.
5. Number of diagnostic tests.
6. Prior distribution.

Overall, 54 heatmaps were produced to represent simulated dataset 3 (Table 10-1) across the three response variables (bias, error, standard deviation), the six facets, and for each of the four parameters under investigation (P, Se, Sp, global metric). Heatmaps are referenced numerically based on groups of 18 as shown in Table 5-2.

Table 5-2: For each response variable Phat, Sehat, Sphat and the global statistic, the heatmaps produced for Chapter 5 are numbered as follows. This full directory of 54 heatmaps can be found on GitHub (https://github.com/annabush/PhD).

	Scale of heatmap		
Facet of heatmap	Error	Bias	Standard deviation
Imprecise versus precise priors	1	2	3

Constrained versus unconstrained priors using data derived from normal distributions	4	5	6
Constrained versus unconstrained priors using data derived from uniform distributions	7	8	9
Number of samples	10	11	12
Number of tests	13	14	15
The prior distribution used	16	17	18

Importantly, the following three rules were used to format the colour scales of heatmaps. To create the colour scales, lists of manually specified colours—shown in pseudocode below—were passed into the `scale_fill_gradientn` function of the `ggplot2` package.

1. The scales of error are coloured `c("green", "pink", "red")` and forced to start at 0, since absolute values can be thought of as a distance from 0, with the colour red indicating the greatest distance from 0.

2. The scales of bias are coloured `c("blue", "White", "red")` and centered at 0 to enable inferences that are overestimates (coloured red) to be quickly differentiated from those that are underestimates (coloured blue).

3. The scales of standard deviation are coloured `c("green", "pink", "red")` and not forced since all values are relative: comparatively small standard deviations (coloured green) indicate precise inferences; and comparatively large standard deviations (coloured red) indicate imprecise inferences.

The code used to produce the heatmaps using `ggplot2` is available at https://github.com/annabush/PhD.

Specifying the Linear Mixed Effects Models

42 LMM's (Table 10-13) were used to study variations in the bias, error and standard deviation of Sehat, Sphat and Phat across up to 11-dimensional parameter space.

The LMM's were specified in accordance with the following pseudocode, with each variable described in Table 5-3:

```
value ~ prior.precision + constraint + n.samples + n.tests *
extreme + prior.distribution + (1 | P.truth) + (1 |
Se.truth) + (1 | Sp.truth),
```

where `value` is a metric of either accuracy or precision; the fixed effects are changed to specify model condition; the random effects are kept constant between LMM's; and the data used is the filtered data frame initialised by the `special.melt` and or the `special.melt2` functions described. All LMM's are fitted using Restricted Maximum Likelihood methods.

Table 5-3: A complete list of the fixed and random effects specified within the regression analyses conducted in Chapter 5 and Chapter 6. Column 2 shows how each variable was declared in R for use by the `lmer` function of the `lme4` package.

Variable name in full	**Variable name as declared**
Continuous fixed effects	
The number of diagnostic tests	`n.tests`
The sample size of the target population	`n.samples`
Categorical fixed effects	

The prior precision of informative priors	`prior.precision`
Uninformative and informative priors	`prior.distribution`
Constrained priors	`constraint`
Edge of parameter space	`extreme`
Random effects	
True values of P	`P.truth`
True values of Se	`Se.truth`
True values of Sp	`Sp.truth`
Response variables	
The accuracies and precisions of inferences of Se, Sp or P, or any inferences made regarding the global statistic.	`value`

Checking the assumptions of the Linear Mixed Effects Models.

The following four checks comprised the workflow for ensuring that the general assumptions (Schielzeth *et al.*, 2020) of LMM's were satisfied. The values required to complete the LMM model checking were obtained from the common return values of `merMod` objects (see Table 8, `lme4` vignette, Bates *et al.*, 2015).

1. Residual values—i.e., the level 1 variance not attributed to position in parameter space—were extracted and compared to the quantiles of a standard normal distribution using quantile-quantile plots to test for normally distributed residuals.

2. Residual values were extracted and plotted against the fitted values, i.e. predicted values of each independent variable to test for linearity (using boxplots for the categorical predictors).
3. Residual and fitted values were extracted and plotted against each other using quantile-quantile plots to test for non-constant variance.
4. Random effects values—i.e., the level 2 variance that is associated with position in parameter space—were extracted and compared to the quantiles of a standard normal distribution using quantile-quantile plots to test for normally distributed random effects.

Combinations of conditions tested by Linear Mixed Effects Models (1-7 repeated for 15% scenario, where P is restricted to 0.15).

1. The dependencies of all fixed effects on the variable `value` given data informed by normally distributed priors.
2. The dependencies of all fixed effects on the variable `value` given data informed by imprecise prior information.
3. The dependencies of all fixed effects on the variable `value` given data informed by uniform distributions.
4. The dependencies of prior distribution on the variable `value` given all data.
5. All manipulations on the variable `value` given all data regarding the inference of Se1.
6. All manipulations on the variable `value` given all data regarding the inference of Sp1.
7. All manipulations on the variable `value` given all data regarding the inference of P.

A note on the how the Linear Mixed Effects Models in Chapter 5 and Chapter 6 are reported.

Coefficients of the LMM regressors are reported on only in terms of their contrasting magnitudes and directions. This is because each fixed effect is dummy coded to represent the predicted difference between its reference level and its contrasting levels (Crawley, 2012), meaning that each fixed effect coefficient can be challenging to directly interpret, with the intercept representing the estimated response at the reference level for all categorical variables, and not a mean response.

The reference levels for the categorical variables used were automatically selected by the `lmer` function (i.e., alphabetically), and are as follows in the format of effect:reference level.

1. prior.precision:imprecise
2. constraint:constrained
3. extreme:FALSE
4. n.tests*extreme:FALSE
5. prior.distribution:normal

An effect-size parameter was not calculated since the data provided to the LMM's is synthetically generated and expected to be noisy (Gelman, 2019). Further, due to the coefficients being only "partially standardised"—i.e., not nested—with a maximum of three levels (for the categorical variable prior precision), any effect sizes could not be directly compared (Lorah, 2018). The variance explained by the LMM's in terms of the usually reported penalised R-squared values (Nakagawa and Schielzeth, 2013; Johnson, 2014) was not reported since this approach for non-nested LMM's is "*riddled with complication*"

(Bolker, 2020) and is not agreed amongst practitioners (Rights and Sterba, 2019). Moreover, since the number of observations, i.e. total sample size, used to calculate the regression coefficients always exceeded 12,000 samples, sampling bias was also not a concern to this study—despite the number of groups of observations being small, i.e. less than ten (Maas and Hox, 2005; Bell, Ferron and Kromrey, 2008)—due to the well-cited problem of diminishing p-values and small standard errors with large sample sizes (Halsey *et al.*, 2015; Amrhein, Greenland and McShane, 2019).

Results

Table 10-13 serves as a look up point for the reader for the LMM's referred to in this chapter, and the conditions they concern. Table 5-2 provides a look up point for all 54 heatmaps that informed this results section. For practical purposes, only select heatmaps are shown in this section, and all heatmaps are produced as multi-panel plots to facilitate quick visual comparisons between variables.

The following three sub-sections report on the accuracies and precisions of Sehat, Sphat and Phat across parameter space, in comparison to the accuracies and precisions of Sehat, Sphat and Phat across global parameter space.

The following comparisons are critical to the evaluation of the strengths and weaknesses of the global statistic. It is important to note that the scales of the heatmaps discussed throughout this results section are not always directly comparable (see the three colour coding rules above), and the x, y and z axes represent predicted rather than true values.

The accuracy (Figure 5-3) and precision of Phat versus the global statistic (Figure 5-2) across parameter space

When informative priors are provided to the BLCM, Phat is estimated to be more accurately inferred and less precisely inferred across parameter space than the heatmaps of global errors suggest. When prior constraint is applied, Phat shows a similar response. If values of P are low, i.e. less than 0.4, and the model has been provided with uninformative priors and priors are unconstrained, P is likely to be accurately inferred, and this is a similar finding within the heatmaps of global errors. For this same modelling scenario, estimates suggest that Phat cannot be precisely inferred if values of P are close to 0.5. Moreover, it is estimated that Phat is consistently underestimated when uniform priors are used and values of P are less than 0.4, whereas the heatmaps of global errors only show that the lower diagonal of parameter space is underestimated for this scenario. Interestingly, the maps of Phat error show no edge effects—i.e., statistically relevant changes in the accuracies or precisions of inferred values, which have been noted to occur when the truth lies within 0.1 units from the edge of a parameter space—when compared to the heatmaps of global errors when values of P are less than 0.3 when priors are constrained or unconstrained, or when the number of samples changed.

The accuracy (Figure 5-4

Figure 5-4) and precision of Sehat versus the global statistic (Figure 5-2) across parameter space.

In general, it is easier to precisely estimate Sehat across all modelling scenarios than the global statistic suggests. For example, the precision of Sehat is largely unaffected by changes in sample size, in contrast to the change in global error

as sample size changes. It appears that Sehat is easier to infer than the global statistic suggests when the value of P is higher; for example, the global statistic indicates that inferences of Sehat when values of P are greater than 0.3 are consistently overestimated. When imprecise priors are given to the BLCM, it appears to be difficult for the model to accurately infer Sehat when values of P are less than 0.1, and when values of Se are greater than 0.9. When informative priors are used, the errors of Sehat exhibit strong edge effects compared to the errors of the global statistic when the values of Se are greater than 0.9. However, there are no edge effects in the parameter spaces of Sehat errors, or global errors, when prior constraints are used, and this is in contrast to when precise priors are used.

The accuracy (Figure 5-5) and precision of Sphat versus the global statistic (Figure 5-2) across parameter space.

The level of information provided by an informative prior—imprecise or precise—does not significantly influence the accuracy of Sphat, and this contrasts with the impact of informative priors on the accuracy of the global statistic. In general, it is found that inferences of Sp are less accurate when inferred using a global metric, particularly when values of P are low. Also, when values of P are low, there is little difference between the accuracy of Sphat when the number of diagnostics used for inference is two, or three, suggesting a potential identifiability issue. When the errors of Sphat are plotted across parameter space there is only an edge effect when the values of Sp are greater than 0.9, and this edge effect is both greater than the corresponding edge effect on the map of global error, and is particularly visible when the values of P are high, i.e. when they are close to values of 0.5. These edge effects are also

visible when standard deviations are plotted across parameter space and show unusually precise inferences in these regions.

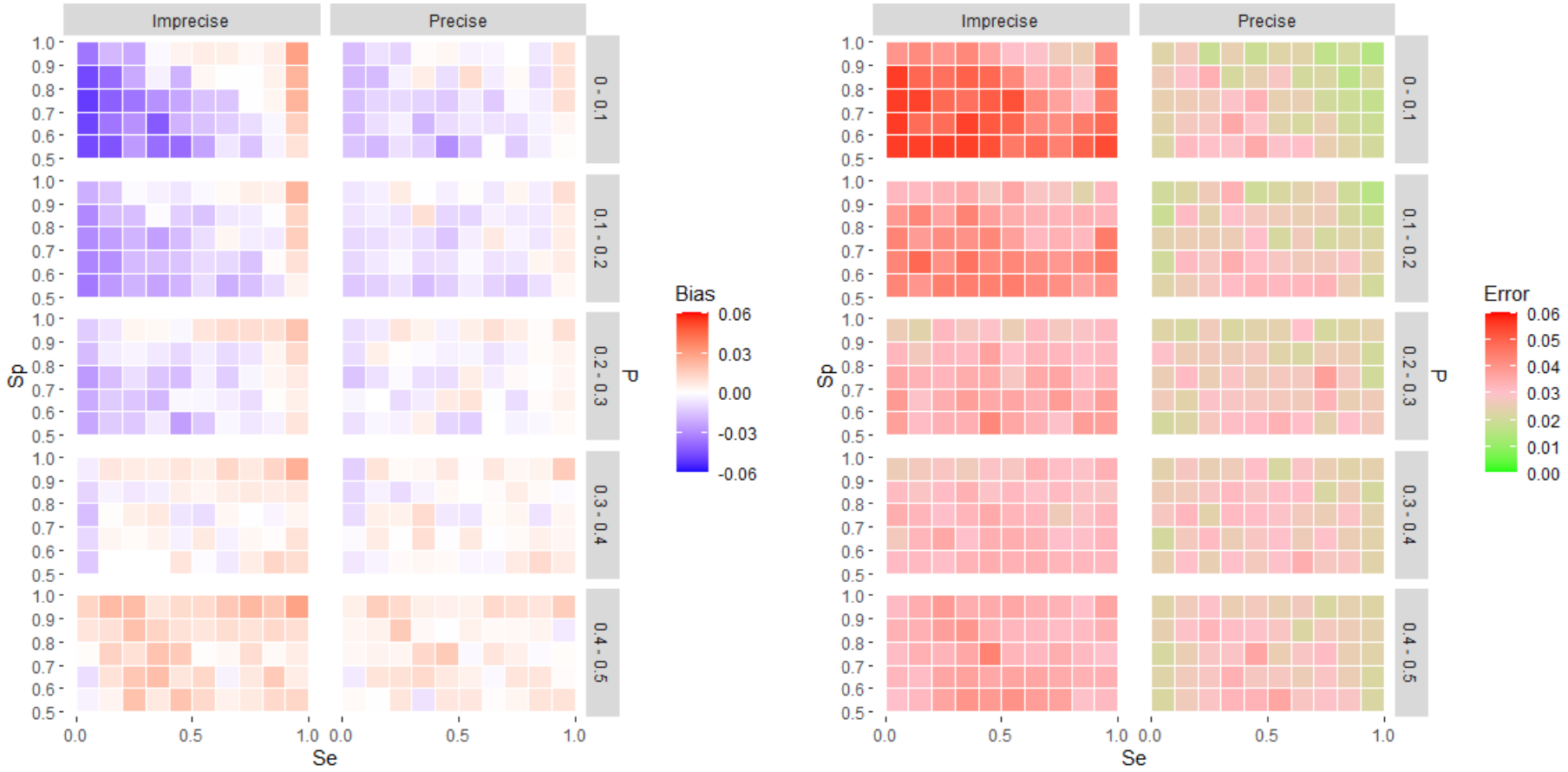

Figure 5-2: Heatmaps showing the bias (left panel) and error (right panel) of predictions of the global statistic given imprecise and precise priors.

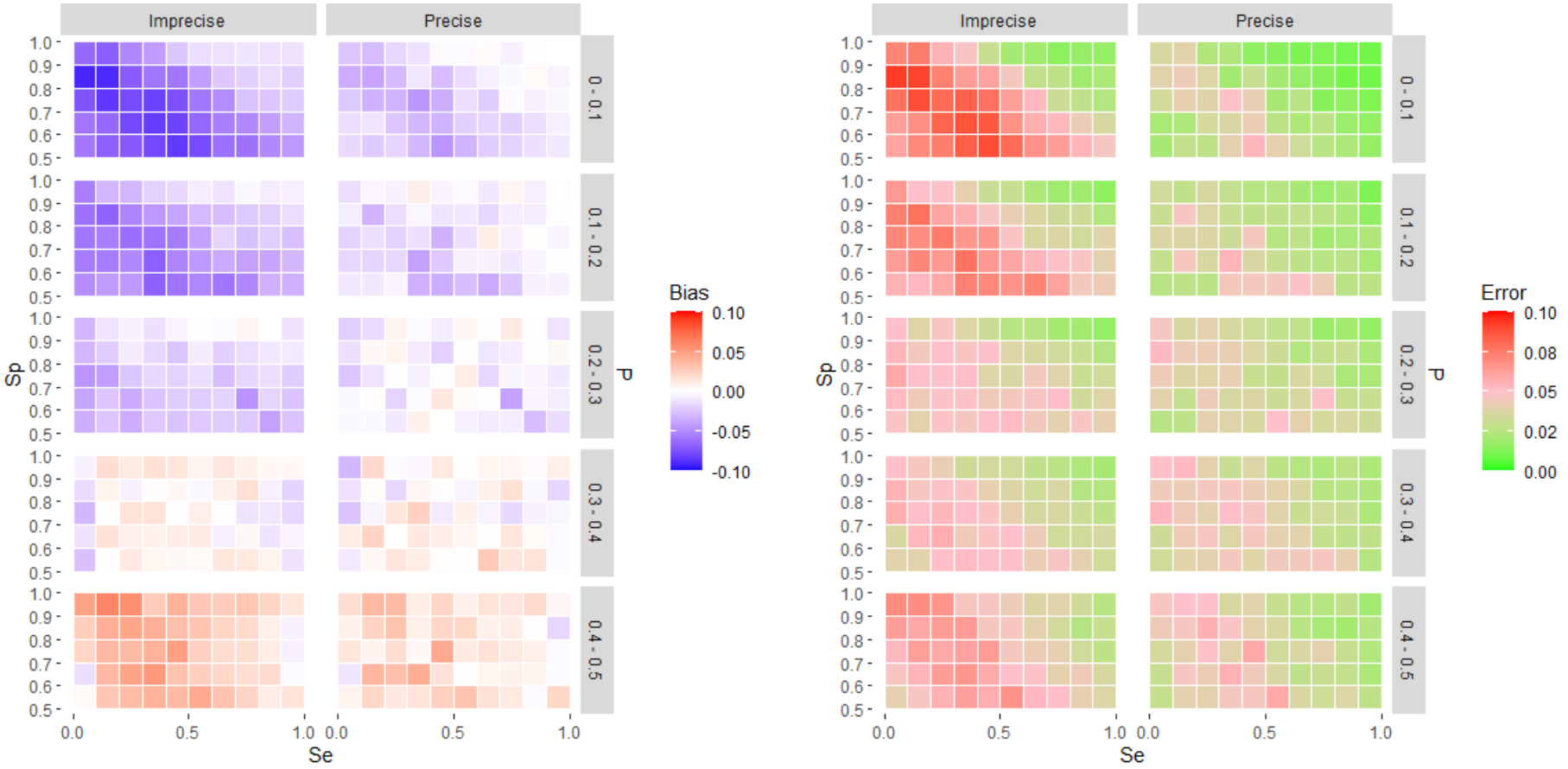

Figure 5-3: Heatmaps showing the bias (left panel) and error (right panel) of Phat given imprecise and precise priors.

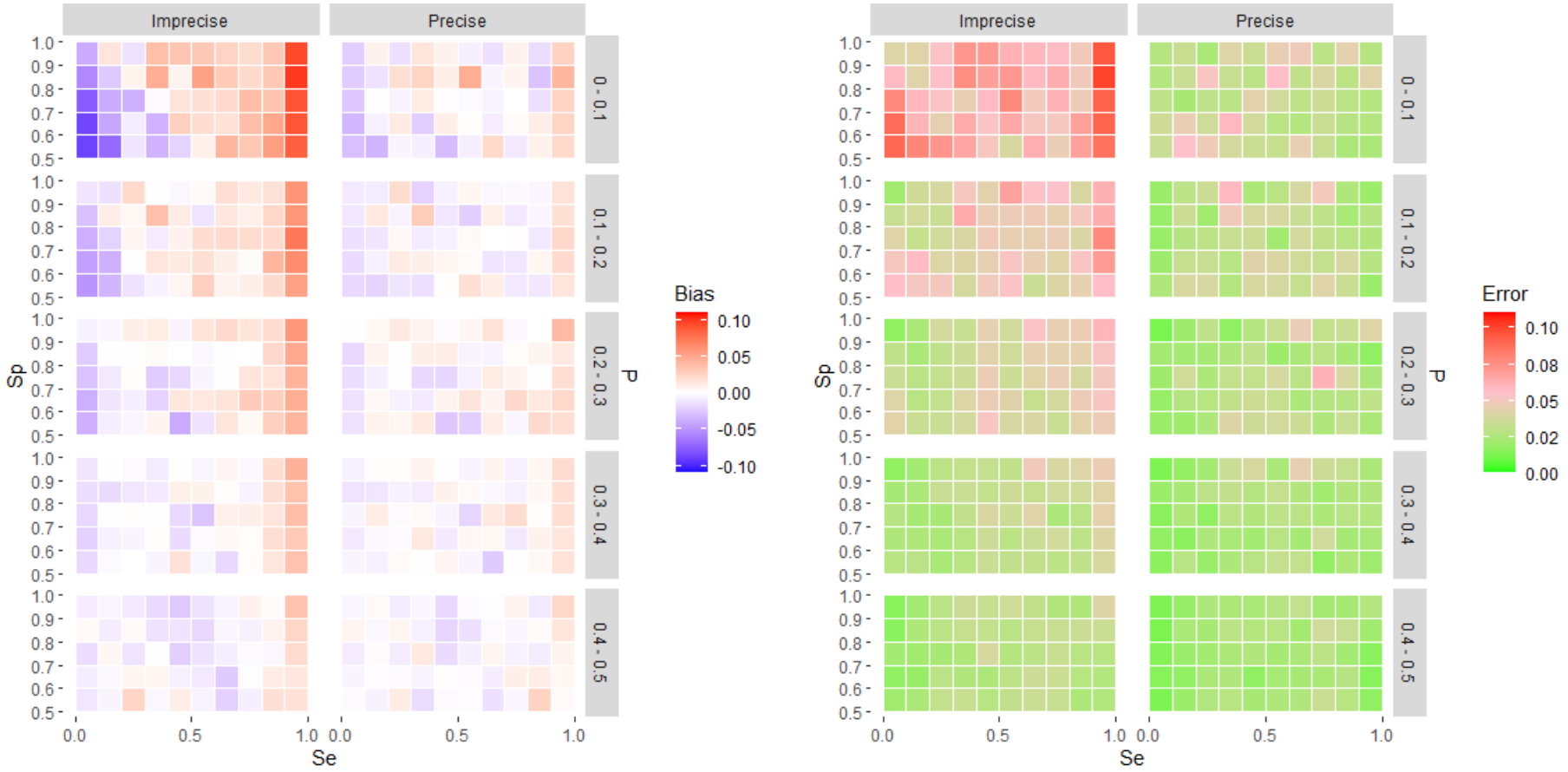

Figure 5-4: Heatmaps showing the bias (left panel) and error (right panel) of Sehat given imprecise and precise priors.

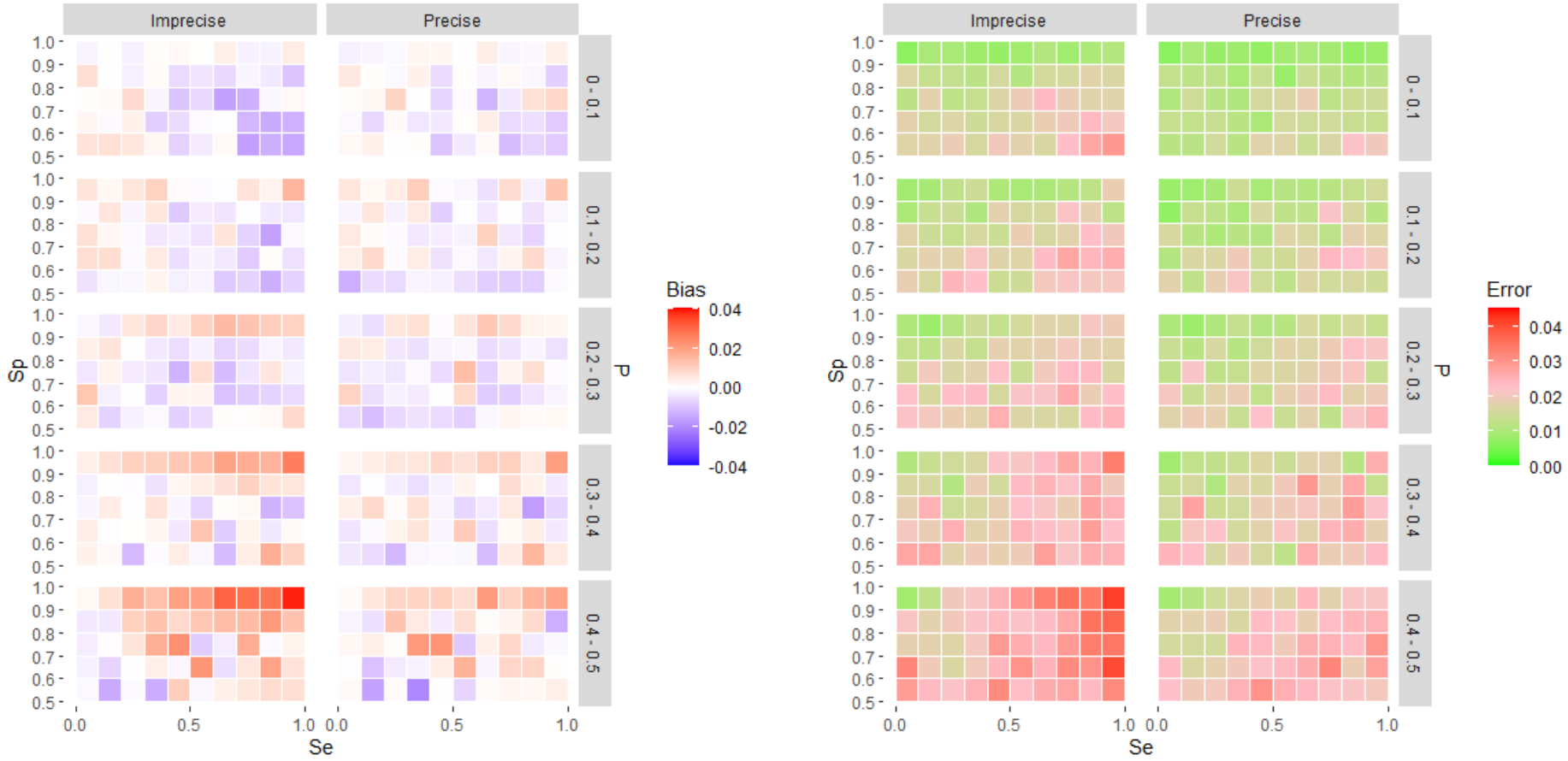

Figure 5-5: Heatmaps showing the bias (left panel) and error (right panel) of Sphat given imprecise and precise priors.

Four scenarios are now reported on: the 15% scenario; situations where parameters Sehat and Sphat are overestimated and underestimated; the artefacts present across parameter space; and the magnitude of global error across parameter space.

Analysis of the 15% scenario

When BLCMs are provided with normal prior distributions, prior precision is generally the most effective source of information for improving the power of a BLCM (LMM's 14, 35, 36, 38, 39, 41, 42). There is also evidence to suggest that providing precise priors is less important for accurately estimating Sphat by one order of magnitude, compared to when precise priors are used to infer Sehat or Phat (LMM 38). This evidence further suggests that providing more diagnostics is the best source of prior information for accurately estimating Sphat (LMM 38). The benefits of providing more diagnostics in comparison to improving other modelling conditions is not unique to the 15% scenario (LMM 38, LMM 29).

When uniform priors were provided to the BLCM (LMM 20 and 21), the application of constraint and increasing the number of diagnostic tests is estimated to result in the decreased error and increased precision of all inferred parameters by the greatest order of magnitude, compared to the effect of increasing sample size.

When normal priors were used to inform the BLCM, regression analyses show that improving prior precision—rather than increasing the number of diagnostic tests—is the best way to reduce the magnitude of errors associated with Sehat and Phat. But for Sphat, the magnitude of error was similarly influenced by prior precision and the number of diagnostic tests.

The variance in Se appears to be responsible for a large amount of variance in the errors of Sehat and this is the same for Sp and Sphat; however, P seems heavily influenced by variations in the accuracy of Sehat.

When is Se and Sp biased, i.e. overestimated or underestimated?

Comparing models informed by imprecise and precise priors.

When values of P are less than 0.3, Sehat and Sphat are more likely to be underestimated. In general, as P increases, Sehat and Sphat are more likely to be underestimated when Se and Sp take values of less than 0.8. When values of P are between 0.3 and 0.5 there are few notable differences between the inferences of Se and Sp in terms of bias, even when imprecise and precise modelling scenarios are compared. However, when values of Se are above 0.8, Sehat is most likely to be overestimated regardless of the value of P. In general, as values of P increase, Sehat and Sphat are more likely to be underestimated when Se and Sp take values of less than 0.8.

Comparing constrained and unconstrained priors given informative priors.

Sehat and Sphat are more likely to be overestimated in constrained scenarios when values of P are more than 0.3. And for scenarios where values of P are less than 0.3, the directionality and magnitude of the accuracies of Sehat and Sphat seem very similar between constrained and unconstrained scenarios.

Comparing constrained and unconstrained priors given uninformative priors.

When uniform priors are used and the model is unconstrained it becomes easy to decide when the errors of Sehat and Sphat are likely to be over- or underestimated, regardless of the value of P. As a rule, as Se and Sp become

larger, their errors are likely to be overestimated. And as P increases, the threshold at which the errors of Sehat and Sphat are overestimated occur at lower values of Se and Sp, in addition, the error associated with inferences at these lower values increase.

Comparing models given differing sample sizes.

The number of samples does not significantly influence the magnitude or directionality of the accuracies of Sehat or Sphat.

Comparing models given differing numbers of diagnostic tests.

When more than three tests are supplied to the model, the effect of overestimating or underestimating parameters becomes negligible compared to when two tests are used.

Further artefacts discovered.

Edge effects

It was found that regardless of the value of P, patterns in the errors of inferences (see plots 1, 4, 7, 10 and 13) occur at the edges of parameter space when values of Se are either low (less than 0.1) or high (close to 0.9) and Sp is high (close to 0.9). These “edge effects”—shown in Figure 5-6—are also present when the variance of error is plotted across parameter space (see plots 3, 6, 9, 12, 15, 18), for example, while sample size has a small effect on precision in general (plot 12), as sample size increases, unusually precise results occur at the edge of parameter space; though edge effects are less present when the number of diagnostic tests increase (plot 15).

These “edge effects” were not present when the biases of inferred parameters were plotted across parameter space, indicating that the edges of parameter space influence the magnitude of error more than its directionality.

Structured variance of error across parameter space.

It was expected that the variance of error across parameter space would be naturally constrained by a pattern of binomial variance.

It was found that the variance of error is always affected in a structured way (see plots 3, 6, 9, 12, 15, 18). When the data represented across parameter space was derived from normally distributed priors, or priors are constrained (or both), the structure can be described as a “ball of higher standard deviation” (Figure 5-7) in the centre of parameter space. When the data represented across parameter space is derived from uniform priors, or priors are unconstrained, the structure can be described as having a higher range of error values in the upper diagonal of parameter space (Figure 5-8).

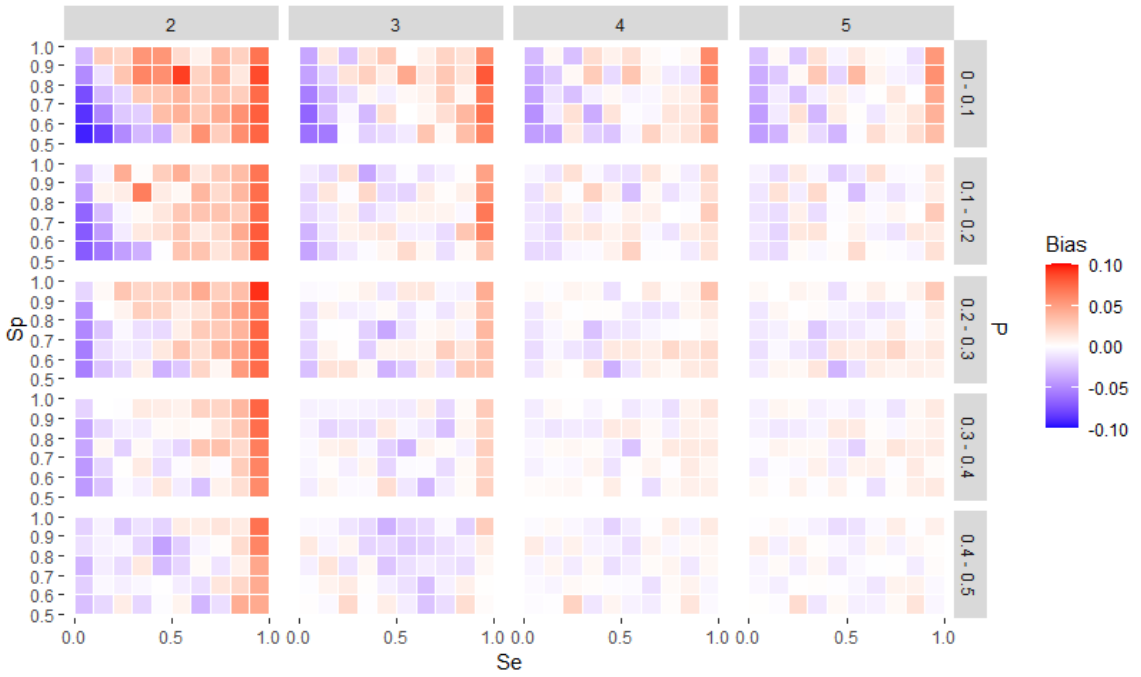

Figure 5-6: A heatmap showing that edge effects associated with the bias of Sehat (for this example) decrease as the number of diagnostic tests available increase from two to five.

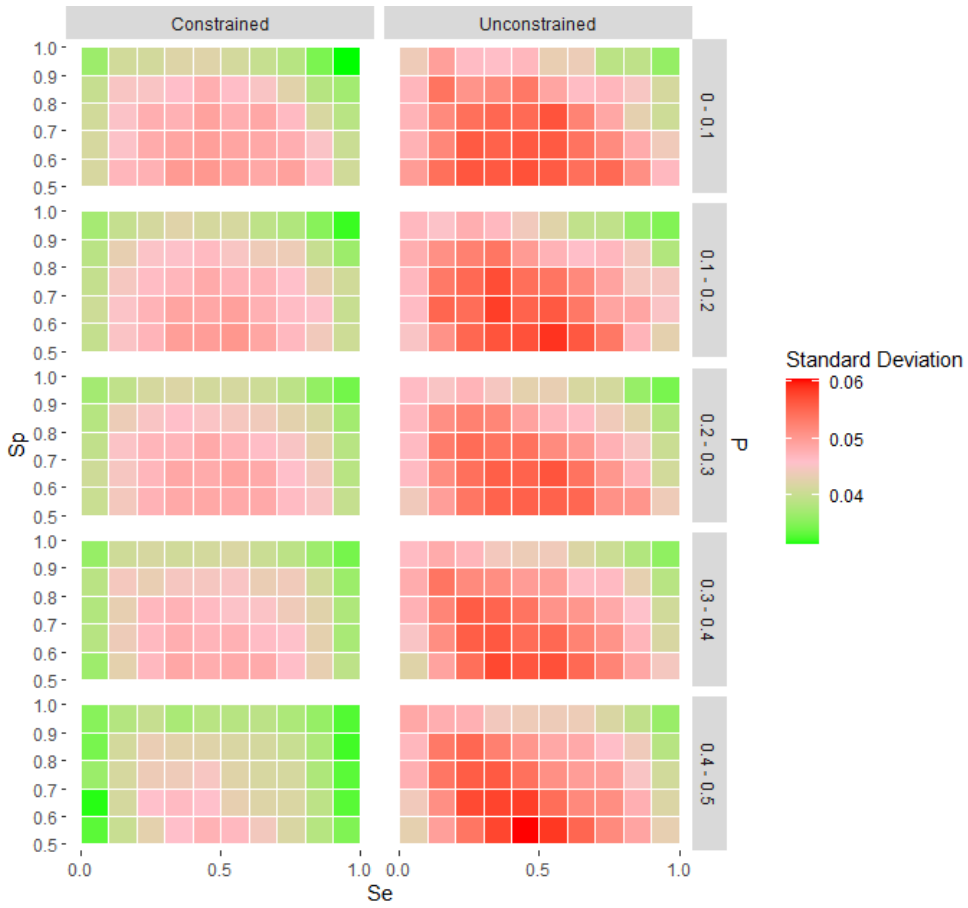

Figure 5-7: A heatmap showing what is described as a “ball of imprecision” in the middle of constrained parameter space, which for this example is associated with the standard deviation of the global statistic when priors (normal) are either constrained or unconstrained.

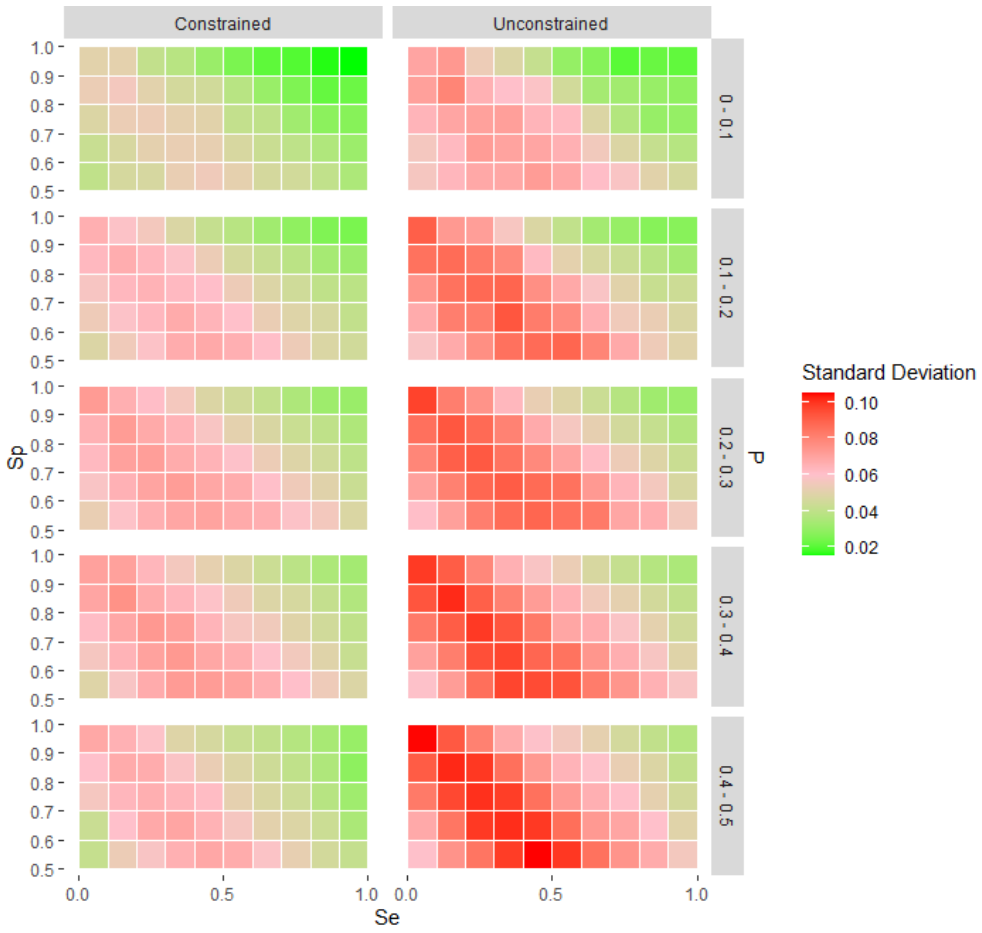

Figure 5-8: A heatmap representing the standard deviation of Phat given constrained or unconstrained priors (normal) to illustrate the “diagonal” pattern through parameter space.

On the magnitude of error across parameter space

Presenting error on heatmaps removes edge effects (plots 2, 5, 8, 11, 14, 17). When true Se and Sp is high, i.e. close to values of one, large errors can be expected when either uniform priors are used, or when the model is unconstrained. Errors are comparably small—even between treatment types such as between imprecise and precise priors—when the model is informed by normally distributed priors. Although in two-test parameter space when true Se and Sp is high, the magnitude of error is close to 0.

Discussion

There remains a discrepancy between studies that assume that the stability of Se and Sp changes with P (Brenner and Gefeller, 1997; Leeflang *et al.*, 2013) and studies that assume that diagnostic accuracy is stable across testing environments (Li and Fine, 2011) and species—given the lack of validation for testing protocols such as for bTB across species (Jia *et al.*, 2020). This is the first known study to evaluate how Sehat and Sphat can be optimised across hyper-dimensional parameter space.

This discussion investigates the assumption of an unstable relationship between Se and Sp given P using simulated data and experiments with two to five diagnostic tests. It is however recognised that a stable relationship may be justified—for example, when agreed cut-offs succeed in improving estimates of P (Helman *et al.*, 2020); or when “bronze” (Lynch *et al.*, 2010; Wu *et al.*, 2016) diagnostics such as a bacteriological culture tests with Sp values of 1 are included within a wildlife disease study.

This assumption that diagnostic accuracy varies across populations sits in agreement with research such as Bermingham *et al.*, 2015, with the ultimate explanation behind variations in diagnostic accuracy underpinned by ecology. For example, attributed to heterogeneities in life history factors such as age or immune status (Pollock, Welsh and McNair, 2005), for instance, Se is known to vary with calf age in tests for bovine cryptosporidiosis (De Waele *et al.*, 2011).

While research such as by Leeflang *et al.*, 2013 describe the assumption that diagnostic accuracy varies across populations as anecdotal, others assume (Gardner *et al.*, 2011), or find (Brenner and Gefeller, 1997), that Sehat and Sphat varies with P. To understand the ecology that prompts diagnostic

accuracy to vary, it first needs to first be understood how the accuracy and precision of inferred parameters varies across modelling situations, and how much variation in accuracy and precision can be accounted for by modelling bias rather than the causative ecology. In this study, many thousands of artificial populations are tested for disease across a variety of conditions.

Understanding how well BLCMs can infer Se, Sp and P is therefore the first step towards understanding the hypothesised reciprocal relationship between Sehat and Sphat, which can only be carried out once Sehat and Sphat are themselves robustly inferred.

General findings

This study finds that Phat has an intricate relationship with Sehat and Sphat which is not the same as between Phat and Sehat, and as between Phat and Sphat. Phat and Sehat were generally more accurate—and Sehat was generally more precise—when single parameter statistics were analysed (measures of accuracy and precision directly associated with the inferences of parameters Se, Sp or P, rather than a global statistic). However, global statistics of precision appear to overestimate the precision of Phat. The accuracies and precisions of Sphat were similar across parameter space when the inferences of Sp and those made regarding the global statistic was compared: it seems that the global metric is generally “good” at inferring P and Sp, and “less good” at inferring Se.

In response to the three high-level research questions listed within the introduction of this chapter, three following high-level dependencies have been found:

1. Using the metric global bias, Sehat and Sphat are likely to be underestimated if values of P are less than 0.3. In contrast, using parameter-specific metrics of bias suggest that while Sphat is likely to be underestimated, this is not true for Sehat.
2. The errors of Sehat and Sphat are dependent on Se, Sp and P and are structured across parameter space. Despite this, it was found that the errors of Sehat and Sphat are more dependent on prior information than position in parameter space.
3. The precision of Sehat and Sphat is strongly dependent on position in parameter space, and prior distribution.

Common to these three findings is a strong relationship between the error (Equation 16) and bias (Equation 17) of the mean posterior inference, and the position in parameter space. Accordingly, these three high-level findings suggest that P should only be inferred once the accuracy and precision of Sehat and Sphat has been determined, as well as the relationship between them.

The remainder of the discussion expands on these findings.

What can the 15% scenario tell us about diagnostic accuracy?

The 15% scenario examined the accuracies and precisions of inferences when values of P are 0.15. Analyses showed that Sehat and Phat generally react similarly to the given modelling conditions and in contrast to Sphat; this was reinforced by the finding that the accuracies of Sehat and Phat have similar dependencies (Table 10-13), which were edge effects and informative priors. This work sheds further light on a potential relationship between Sehat and Phat first reported in Chapter 4, which is not the result of collinearity integral to the BLCM (Figure 3-4), and that holds the potential to influence experimental

design. This finding is important to ecologists wanting to better inferences of P, since when infection rates are low, improving Sehat could be a viable strategy. And one way to do this could be to not trust global statistics of accuracy and precision as being representative of Se.

The 15% scenario also highlighted that the accuracy of Sphat was found to be less sensitive to precise priors than Sehat or Phat. When normal priors were provided, the accuracy of Sphat was more sensitive to the provision of further diagnostics. And when uniform priors were provided, the accuracy of Sphat increased when provided with constraints and further diagnostics, rather than when sample size increased. These findings suggest that when infection rates are low, including more diagnostic tests—probably by proxy—would be the best way to improve inferences of Sp, regardless of how the priors are specified, and regardless of whether it is accuracy or precision of the parameter that the researcher wished to improve. These findings are in agreement with the finding (Liu *et al.*, 2014) that the prior of Sehat is more important than Sphat when P has values of less than 0.5.

What can we learn from mapping across parameter space?

A notable outcome from the heatmap analysis using global statistics is that Sehat and Sphat are very likely to be underestimated if values of P are less than 0.3. This finding was unaffected by changes in how prior precision was specified, how prior constraint was specified, or the number of samples used to build the test array; and agrees with the findings of Helman *et al.*, 2020 in both their simulated study and wildlife case study. The only condition that contradicted this finding was the two-test scenario, where Sehat and Sphat were poorly underestimated (in scenarios when values of P are less than 0.3) when compared to the bias of Sehat and Sphat given the same modelling

conditions, and three diagnostic tests. However, it was shown that Sehat is not underestimated when values of P are less than 0.3, providing another example of when the performance of the global metric is suboptimal.

The finding that the global bias was largely independent of experimental conditions—and dependent on the value of P—is not replicated when the magnitude of error for Sehat and Sphat is considered, which seems very dependent on prior specification. It is found that the least accurate inferences generally occur when the values of Se and Sp are high, i.e. when better tests are used; and when either uniform priors are used, or when the model is unconstrained. Considering this, normal priors should always be preferred over uniform priors, even if they are imprecise. The single parameter experiments, i.e. those where global errors were not used, confirm that the least accurate inferences can be mostly attributed to edge effects. The finding that the biases in inferences of global error was largely dependent on P is also reflected in the single parameter studies, which find that P is dependent on prior specification to avoid the underestimation of error.

Heatmapping also highlighted a structuring of the variance of error across parameter space that is dependent on whether uniform or normally distributed priors were used. When priors are normally distributed, the variance of error is greatest in the centre of parameter space; and when priors are uniform then the variance of error is greater the closer the test is to a gold standard. These findings suggest that trusting precise values of Sehat and Sphat at the edges of parameter space should be avoided, and that prior information needs to be provided to the model to avoid overestimates of Sehat and Sphat when values of Se and Sp are above 0.5.

Further evidence of edge effects as a statistical artefact

On average, when error is absolute, the accuracy of Sehat and Sphat appears high. But underneath this general finding, there is evidence that a mean-variance relationship is partially responsible for the observed edge effects. In general, parameters seem most difficult to infer when on the edge of parameter space. For models provided with constrained truths this means when values of Se are either very low or very high (near values of 0 or 1), when values of Sp are low (near values of 1), and or when values of P are high (near values of 0.5). Uncovering edge effects supports existing findings such as that low Se and or Sp values reduce model identifiability (Bujang and Adnan, 2016); that the accuracy of Sphat is generally lower when P is higher (Leeflang *et al.*, 2013); and that Phat is inaccurate when Se is low (McDonald and Hodgson, 2018).

Are edge effects related to constraint?

This research supports the view that edge effects are directly influenced by the constraints applied to the priors that inform the BLCM. For example, when prior constraint is applied there are no apparent edge effects on heatmaps of global errors across parameter space or heatmaps of Sphat. However, when prior constraint is applied, edge effects are present on heatmaps representing global standard deviations. Edge effects on maps representing the standard deviation of Sphat across parameter space show unexpectedly precise inferences, indicating that these edges could be hard to trust. And for maps representing the errors of Sphat, edges are more inaccurate than can be seen on the heatmaps of global errors across parameter space, indicating strong edge effects. However, when values of P are low, i.e. less than 0.3, there are no edge effects in constrained or unconstrained models indicating that for most wildlife infections scenarios, edge effects of P are less relevant considerations.

This analysis of edge effects suggests that tests with high values of Se are particularly difficult to accurately infer using unconstrained models, that heatmaps of global errors across parameter space do not reflect the true edge effects for Sp, and that tests with high values of Se and Sp are particularly difficult to precisely infer, suggesting that this is a limitation to estimating Phat accurately within the BLCM structure specific to this chapter. It may be that P is the most “volatile” parameter within BLCMs, and this is supported by the finding by McDonald and Hodgson, 2018, that Phat is unreliable when diagnostic uncertainty is not taken into account.

Are edge effects related to prior distributions?

Every source of prior information acted to reduce error when prior precision was normally distributed, with the most significant reductions in error attributed to precise prior precisions and the least significant reductions attributed to sample size. In contrast, when priors were uniformly distributed, constraint became the most influential model condition. Although it has been suggested (Bujang and Adnan, 2016) that subjectively larger sample sizes are needed to estimate Sehat when values of P are low, and that larger sample sizes are needed to estimate Sphat when values of Sp are high, this study finds that—using relatively large sample sizes—sample size is largely irrelevant to these estimation problems.

Conclusion

This chapter has aimed to advance the understanding of combinations of parameter values that lack practical identifiability, and in particular, those that arise due to the hypothesised reciprocal relationship between Se and Sp. To do this, a method for exploring high-dimensional parameter space was described

and executed, and dynamics between Se and Sp are quantified. In addition, the chapter described and further explored three high-level trends that characterise the relationship between Se, Sp and P suggesting how the accuracy and precision of—and the relationship between—Sehat and Sphat can depend on model specification. This research is thought to be the first to suggest how information should be added into BLCMs to improve inference. In summary, it was found that there are structured patterns in the variance of error across parameter space. These mean-variance relationships were hypothesised to be another integral statistical artefact that will be critical to understanding how identifiability changes across regions of parameter space.

But can this hypothesised mean-variance relationship be distinguished from identifiability issues? And is this relationship simply due to less identifiable parameter space? In other words, are answers to the simultaneous equations that underpin Equation 14 more difficult for the MCMC algorithm to solve at the edges of parameter space? The following chapter, Chapter 6, turns to these important considerations.

Chapter 6

6. Investigating the interactions between edge effects, BLCM identifiability, and the mean and variance of error.

Introduction

The previous chapter showed that regions of extreme parameter space were "dependence structures"—places where dependencies existed between two or more variables of interest—that were critical to BLCM identifiability. Clearly, for real-world studies, the risk exists that if unaccounted for, the inaccuracies associated with the presence of such artefacts—that is, trends explainable by statistical vagaries such as edge-effects, rather than ecology—could result in flawed disease management decisions.

The relationship between a BLCM's power and its potential to influence wildlife disease management decisions is only recently being explored: for example, within Helman's recent study on Leptospira infection in California sea lions (Helman *et al.*, 2020). The present chapter contributes to this specific body of research by investigating the potential ramifications of addressing edge-effects, i.e. the mean-variance relationship identified across parameter space that describes the variance of error.

Understanding the biological (Horne and Schneider, 1995) or statistical meaning behind the non-constant variance of response variables across "space"—genetic, geographic, or even theoretical space, such as parameter space—is complex (McClintock *et al.*, 2010) and infrequently a core research

ambition of ecologists. Few examples of research with such ambition can be found in the fields of species distributions (Palmer, Hakenkamp and Nelson-Baker, 1997; Elith and Leathwick, 2009), disease, and survival (McDonald *et al.*, 2016). Importantly, to-date, inferences on the mean dominate the field of ecology, including wildlife disease research, despite it being understood that *"failing to account for [mean-variance relationships] appropriately can introduce serious artefacts into analysis"* (Warton and Hui, 2017).

Binomial data—such as those recording the presence or absence of disease, or count data summarising the infection statuses of a population as determined by physical diagnostic tests—often presents a relationship between means and variances; the most common in ecology being overdispersion, when the variance is generally greater than the mean (Lindén and Mäntyniemi, 2011; Conn *et al.*, 2018). Although studies on these mean-variance relationships are critical to analysing multivariate data (Warton and Hui, 2017), they have—since the publication of Taylor's Power Law, shown in Equation 19, in 1961 (Taylor, 1961)—been controversial: for example, see Warton, Wright and Wang, 2012, and subsequent responses by Roberts, 2017 and then Warton and Hui, 2017. While Taylor's Power Law is the mean-variance relationship that has dominated ecology, and which usefully describes population sizes for many species, its principles have been widely extrapolated (Eisler, Bartos and Kertész, 2008).

Equation 19

$$var(Y) = a\mu^b$$

In Equation 19, $var(Y)$ is the variance of the size of an insect population, μ is the population mean, and a and b are both positive constants.

A basic assumption of mean-variance relationships in ecology is that most "environmental factors", such as P, are bounded by 0 (Scheiner and Willig, 2013), and in the case of probability parameters, with an upper bound of 1. The heatmaps of absolute mean error put forward in Chapter 5 are bounded by 0 and 1, or 0.5, in all dimensions and show clear changes in inferences at "edges", i.e. the space between a boundary edge and a 0.1 unit from that edge, which in this thesis represents the "extreme" limits of possible values for inferences of Se, Sp or P.

Although Chapter 5 demonstrated the existence of a relationship between error and its variance across space, the exact shape of this relationship is unclear, as is the bias that it could represent. Further, the effect of the mean-variance relationship on error needs disentangling from the presence of identifiability issues, which could also force edge effects. This chapter therefore investigates whether edge effects are statistical artefacts that ecologists must understand prior to correctly interpreting model uncertainty.

Edge effects in ecology are well-studied (Ries and Sisk, 2004) when they concern population-level responses to an environmental boundary, and the term "edge effect" in theoretical ecology is sometimes associated with graph theory. However, classical ecological definitions of an edge effect which invariably relate to changes in ecosystems at boundaries due to environmental factors (Ries *et al.*, 2004), are still meaningful in studies of theoretical space because their definitions are researcher-dependent (Strayer *et al.*, 2003), and their study relates to the prediction of population-level traits, in this instance P.

How the variance of BLCM inferences is interpreted can directly influence wildlife disease management decisions. For example, Helman's 2020 study on Leptospira infection in California sea lions was one of the first to be informed by

how the accuracy and precision of parameter inferences can change across parameter space as modelling conditions vary (Helman *et al.*, 2020). Chapter 5 noted that there is little guidance for ecologists on analysing BLCM inferences, this chapter focuses on how statistical artefacts such as heteroscedasticity—the increasing spread of residuals as the fitted values of the response value change—may be interpreted, and explores whether it can be separated from other biases such as model identifiability issues. Ultimately the ability to understand how statistical artefacts change with modelling conditions could have direct disease management implications, such as those indicated by Helman *op cit*.

This chapter also advances the argument presented in Chapter 5 that the edges of parameter space—modelled in LMM's by the variable "`extreme`"—are associated with dependence structures and statistical artefacts critical to both BLCM identifiability and parameter interpretation. In this chapter the variance of error across parameter space is analysed, and how the error structures of Sehat, Sphat and Phat may be interpreted across a range of modelling conditions are discussed. To accomplish this, this chapter first examines whether mean-variance relationships are present across parameter space, and second, explores how much distortion the mean-variance relationship causes in terms of the inference that should be expected when the mean-variance relationship is removed.

Assumptions and Methods

This chapter asks two key questions:

1. Are mean-variance relationships present across parameter space?

2. How much distortion does the mean-variance relationship cause at edges?

These explorations of the mean-variance relationship and its impact—or otherwise—on inferences are carried out by testing five hypotheses. These focus on exploring the *accuracy* of mean error across space, rather than the *precision* of mean error across space (which would be an analysis of the variance of variance estimates). Simulated dataset 3 (Table 10-1) is used, continuing the same basic hypothetical modelling scenario as described in Chapter 5, unless explicitly stated otherwise.

The answers to these two questions lie in the construction and testing of 5 hypotheses explained within this section.

Are mean-variance relationships present across parameter space?

To explore the existence of mean-variance relationships across parameter space, it is necessary to look for difficult-to infer parameter space, which involves making some basic assumptions about where the edge of parameter space is, the nature of the relationship between prior information and difficult-to-infer parameter space, and probable impacts of edge-effects.

Where is unidentifiable parameter space?

When the number of diagnostic tests is five, and eight parameters are free to vary (Table 3-1), it can be assumed that more inferred values are closer to the edge of parameter space than if mapped using smaller batteries of tests. It is also assumed that a relationship exists between the prior information provided, and unidentifiable parameter space; as intuitively, prior information should improve inferences.

HYPOTHESIS 1: Inferences are associated with less accuracy as truths approach the edges of parameter space, and prior information affects this relationship.

HYPOTHESIS 2: Edge effects cause a reduction in model power.

Chapter 5 demonstrates that Sehat, Sphat and Phat do not behave similarly, so in this chapter their errors are modelled as independent response variables, in addition to the mean global error within each `0.1 x 0.1 x 0.1` cell of a maximum `1 x 1 x 1` parameter space. The decision to avoid total reliance on global statistics is further supported by the body of literature on Small Area Estimations (such as Jiang and Lahiri, 2006), in which each cell in parameter space may be considered one. The decision is also supported by Waller and Carlin, 2010 who remark that "*such smoothing* [i.e. taking averages across space] *may not be appropriate if the goal is instead to identify boundaries or regions of rapid change* [i.e. edge effects] *in the response surface* [i.e. the surface of parameter space]*, since smoothing is antithetic to this purpose.*"

Chapter 5 demonstrated that presenting the errors of Sehat, Sphat and Phat on heatmaps can remove edge effects. While there is evidence to support the assumption that when error is absolute, on average its value is not overestimated or underestimated, ecologists must be confident that absolute errors—when inputted into LMM's—are representative of true accuracy. In the absence of further evidence to the contrary, mean-variance relationships automatically violate the familiar model assumptions (Gelman and Hill, 2006) behind LMM's by introducing heteroscedasticity.

Assumptions about edge effects

Chapter 5 provides evidence that in some instances edge effects represent a decrease in a BLCM's ability to infer the truth, possibly due to unusual diagnostic outcomes at the edges of parameter space; for example, positive diagnoses are rare when Sp and P are low.

HYPOTHESIS 3: The variance of the error of Sehat, Sphat and Phat across parameter creates edge effects.

HYPOTHESIS 4: Both model constraints, and whether the errors of inferred values are absolute or not, influence the homogeneity of variance across parameter space.

Prior to investigating hypotheses 3 and 4 it was confirmed that a correlation between the MCMC samples used to infer Sehat, Sphat and Phat was not present (Figure 3-4).

Modelling the mean-variance relationship

A modelling scenario with three diagnostic tests, 1000 individuals, and normal priors was generated, and then compared to a similar modelling scenario informed by uniform priors. This subset of simulated dataset 3 ensured that comparisons among and between modelling scenarios were manageable, and that all results could be directly comparable with Chapter 5 since the datasets used were of the same seed.

The mean inferred values of Sehat, Sphat and Phat were plotted over five conditions, generating 15 mean-variance relationships for study shown in Figure 6-1 to Figure 6-5. These conditions were:

1. Uniform priors (control).

2. Precise priors and constrained parameter space.
3. Imprecise priors and constrained parameter space.
4. Precise priors and unconstrained parameter space.
5. Imprecise priors and unconstrained parameter space.

Mean-variance relationships of the mean posterior inferences were plotted—rather than the error of mean posterior inferences calculated as per Equation 16—to indicate the variance of parameter predictions rather than of error predictions. Note, the mean-variance relationship of the global statistic was not studied. The `geom_smooth` function of the `ggplot2` package (Wickham, 2014) was used to provide a trend line through the point data, where each point represents a single posterior inference.

How much distortion does the mean-variance relationship cause at edges?

To explore the distortion that mean-variance relationships present across parameter space, it is necessary to identify whether transformations of error can address this distortion, to correctly specify and interpret LMM's accordingly, and to be able to robustly check for the homogeneity of error across parameter space.

Specifying the logit (i.e. log-odds) transformation of error

So far, changes in error with respect to position in parameter space have only been examined in "absolute terms", i.e. using comparisons of differences. In contrast, examining logit-transformed errors (Equation 20) across parameter space shows relative changes, and this has two purposes.

1. To distinguish between edge effects and non-edge effects.

2. To investigate whether a logit-transformation of error removes edge effects by accounting for the heteroscedasticity in models where data can't be represented on the probability scale.

When the errors of Sehat, Sphat and Phat are transformed by the log-odds (Equation 20 and Equation 21), it is hypothesised (Hypothesis 5) that the variance of error is flattened across parameter space, leaving behind the edge effects, potentially removing and therefore explaining the edge effects.

HYPOTHESIS 5: Edge effects can be removed from analyses by transforming errors using the logit link function.

Logit-transformed errors across parameter space are not directly interpretable beyond this high-level trend, and further transformations to try and counter this—such as absolute logit-transformed errors, and inverse logit absolute logit-transformed errors—suffer from similar interpretability issues. It is assumed that symmetrical errors on the logit scale become asymmetric on the probability scale unless values of P = 0.5.

Equation 20

$$Logit\ error = \ log\left(\frac{\hat{y}}{1-\hat{y}}\right) - log\left(\frac{y}{1-y}\right)$$

Equation 21

$$Absolute\ logit\ error = \left|log\left(\frac{\hat{y}}{1-\hat{y}}\right) - log\left(\frac{y}{1-y}\right)\right|$$

Coding the linear mixed effects models

As Hayes and Cai, 2007 write, "*linear regression is a foundation upon which more complex models can be constructed.*" The direction and magnitude of relationships between logit-transformed errors and predictor variables (those

considered are listed in Table 5-3) were analysed using LMM's. 16 LMM's were specified (Table 10-14) using the pseudocode as follows:

```
value ~ prior.precision + constraint + n.samples + n.tests *
extreme + prior.distribution + (1 | p.truth) + (1 |
se.truth) + (1 | sp.truth),
```

where the regression coefficients remain as defined within Chapter 5 (Table 5-3), with the following three changes.

1. The variable "`value`" can be any metric of accuracy for four parameters—global metric, Sehat, Sphat and Phat—with four variations: absolute and logit-transformed; not absolute and logit-transformed; not absolute and not logit-transformed; and absolute and not logit-transformed. These four variations are referred to in accordance with the terms shown in Table 6-1. Note, the fixed effects are not changed.
2. The random effects are kept constant between LMM's and remain crossed, i.e. non-hierarchical in models that contain parameter-specific responses, as the relationship between Se, Sp and P is intrinsic to the BLCM.
3. Data is the filtered data frame initialised by the `special.melt` and or the `special.melt2` functions first described in Chapter 5. For this chapter *data* is filtered by normally distributed data only.

Table 6-1: The four transformations of "error" analysed within Chapter 6.

	No logit transformation	**Logit transformation applied**

No absolute transformation	Bias (Equation 17)	Logit error (Equation 20)
Absolute transformation applied	Error (Equation 16)	Absolute logit error (Equation 21)

Checking the homogeneity of the variances of errors across parameter space

Standard plots (Crawley, 2012) of the fitted values, i.e. the predicted values of the LMM, versus the residual values, i.e. the difference between the predicted and actual values of the dependent variable, were created. The dependent variables are the measures of error, bias and standard deviation associated with the posterior inferences of a BLCM.

Standard “fitted versus residual” plots were created to assess the homogeneity of the LMMs for three reasons.

1. To ensure that the spread of residual variance is normally distributed.
2. To ensure that the mean of the residuals is constant across space.
3. To ensure that the fitted values for each regression analysis are associated with known errors.

Accordingly, fitted versus residual plots are used in this thesis to examine where the biases of Sehat, Sphat and Phat may be overestimated or underestimated by the LMM.

The following assumptions were applied when conducting the residuals analysis:

1. The difference between the fitted and true values can be used to quantify the power of the inferences obtained from a BLCM.
2. In a “good” LMM, residuals will randomly deviate from zero in a symmetric way and be close to zero to demonstrate low variability.
3. By plotting the residual versus fitted values, models can be visually examined for heteroscedasticity.

For all modelling conditions, the residuals versus the fitted values of all LMM’s were extracted from the model output using the `lme4` package (Bates *et al.*, 2015), and were colour-coded according to the following two rules, which provide the plots with differing information.

RULE 1: Colouring residuals if they are associated with an edge of parameter space, where Se > 0.9 and Se < 0.1 and Sp > 0.9 and P < 0.1.

RULE 2: Colouring residuals based on their position in parameter space, where “`upper`” = (Se > 0.9 and Sp > 0.9); “`lower`” = (Se < 0.1 and P < 0.1); and “`middle`” = all other space.

The `reformat` and `run_model` functions specified within the online code repository (https://github.com/annabush/PhD) were constructed to enable plotting.

Results

Findings are reported in three sections. First the mean-variance relationships of Sehat, Sphat and Phat are described (see Figure 6-1 to Figure 6-5); second, regressions using logit-transformed Sehat, Sphat and Phat are reported (Table 10-14), and third findings, from the fitted versus residuals plots are presented

(only select plots are reproduced in this section, for all plots see https://github.com/annabush/PhD).

The mean-variance relationships of Sehat, Sphat and Phat

The "control" modelling condition, Figure 6-1.

When the mean-variance relationships of Sehat, Sphat and Phat were generated from uniform priors, new statistical artefacts were generated. The variance of Sehat and Sphat peaked twice at mean values of ~0.3 and ~0.6, and the mean-variance relationship of Phat was found to be highly skewed, where models with mean values of Phat between ~0.3 and ~0.6 exhibited a negative binomial correlation with the variance of Phat.

In precise and constrained modelling conditions, Figure 6-2.

The variance of Phat peaked between mean values of Phat of 0.3 and 0.4, and significantly dropped towards 0 when mean Phat became less than 0.2, or more than 0.4. The average variance of Phat was twice that of Sehat and Sphat under the same (precise and constrained) modelling conditions.

The variance of Sphat was found to be highly dependent on the number of diagnostic tests and increased as mean Sphat increased. The variance of Sphat was smaller than Sehat or Phat by one order of magnitude. The variance of Sehat decreased at mean values of Sehat of less than 0.1 and at mean values of Sehat of more than 0.9, whereas the variance of Sphat decreased as mean Sphat increased. This latter finding is true for all mean-variance relationships that were based on data from models given informative priors.

In precise and unconstrained modelling conditions, Figure 6-3.

The variance of mean Phat is greater than mean Sehat or mean Sphat, however (unlike when the constrained modelling condition was used) the variance of mean Phat does not drop towards 0 at mean values of 0.5: as mean Phat increases the variance increases.

The mean-variance relationship of Sehat and Sphat in precise and unconstrained modelling conditions is similar to the mean-variance relationship given precise and constrained modelling conditions.

In imprecise and constrained modelling conditions, Figure 6-4.

The variance of mean Phat peaks between 0.2 and 0.3 and drops rapidly towards 0 at 0.3. The variance of mean Sehat and Sphat is generally greater by one order of magnitude than in precise and constrained modelling conditions. Values of mean Sehat are associated with the most variance, compared to mean Sphat or mean Phat, and this is also true when the precise and constrained modelling condition is used.

In imprecise and unconstrained modelling conditions, Figure 6-5.

In this modelling scenario the mean-variance relationships of Sehat, Sphat and Phat were skewed by outliers, and values of mean Sphat are associated with the greatest variances. As reported for the precise and unconstrained modelling condition, the variance of mean Phat does not drop towards 0.5.

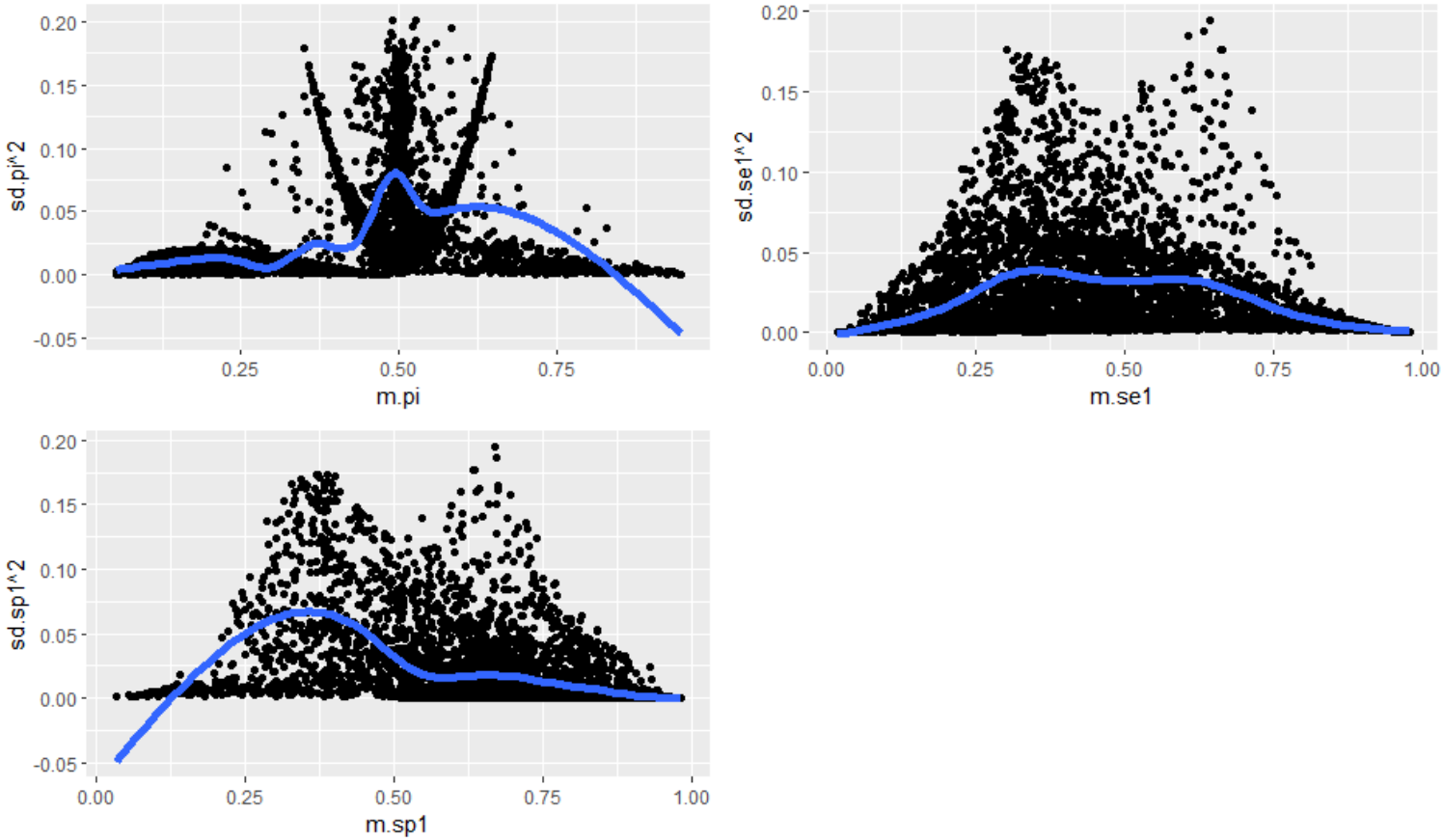

Figure 6-1: The relationship between the mean (`m.variable`) and variance (`sd.variable`) of the posterior inferences of the replicated posterior means of P (top left), Se (top right) and Sp (bottom left) given data from a BLCM informed with uniform priors.

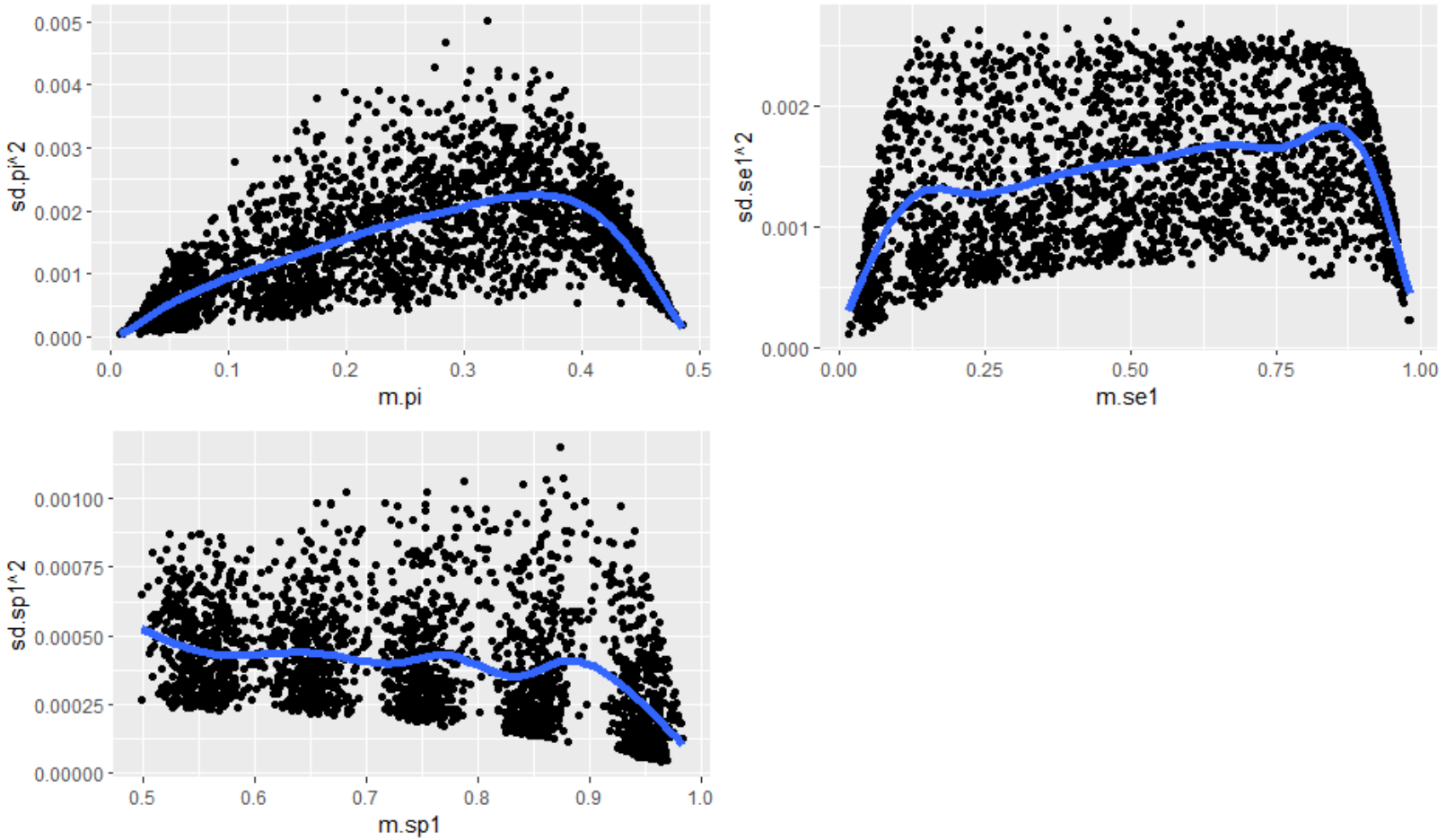

Figure 6-2: The relationship between the mean (`m.variable`) and variance (`sd.variable`) of the posterior inferences of P (top left), Se (top right) and Sp (bottom left) given data from a BLCM informed with precise priors and constrained truths.

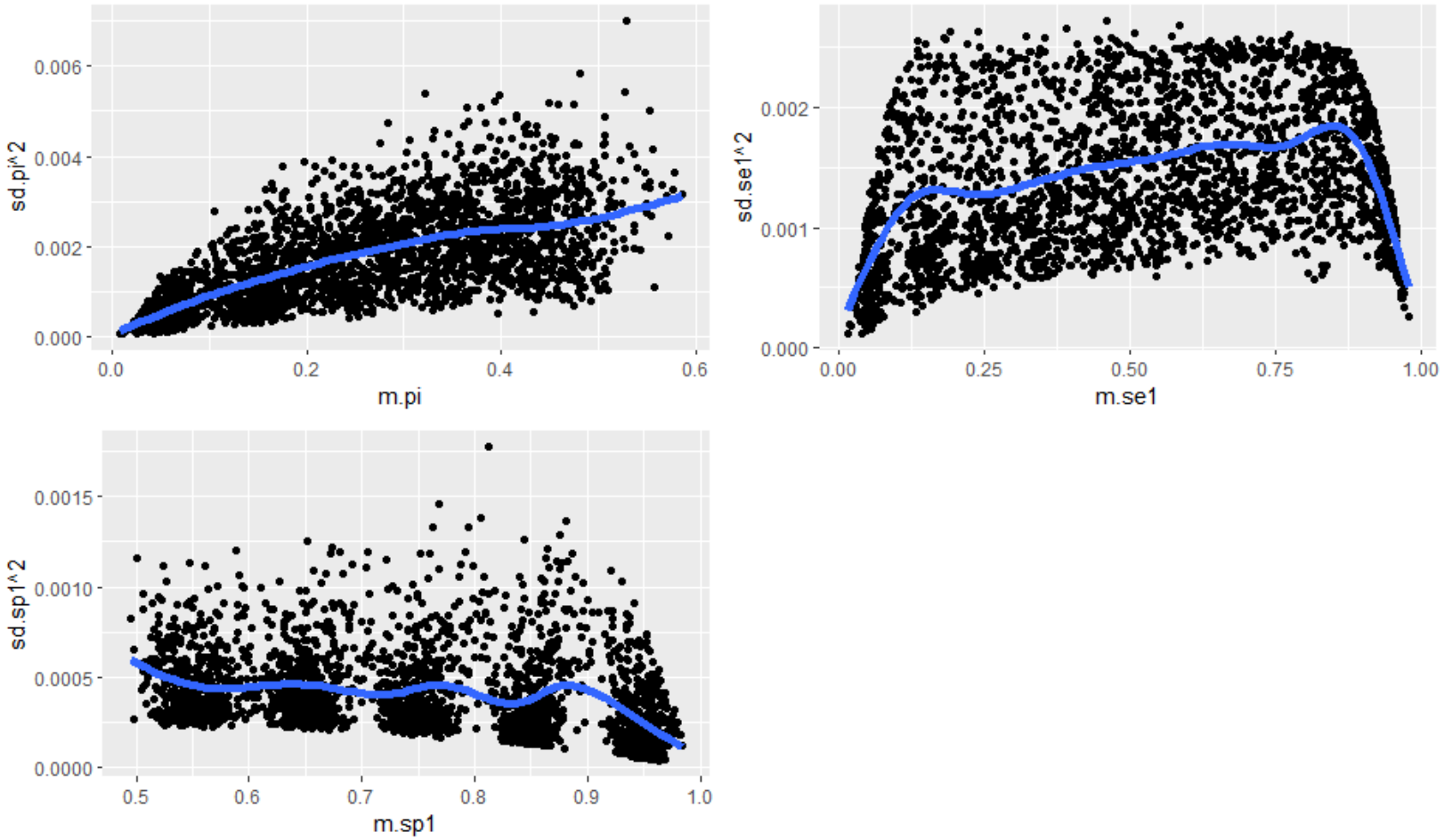

Figure 6-3: The relationship between the mean (`m.variable`) and variance (`sd.variable`) of the posterior inferences of P (top left), Se (top right) and Sp (bottom left) given data from a BLCM informed with precise priors and unconstrained truths.

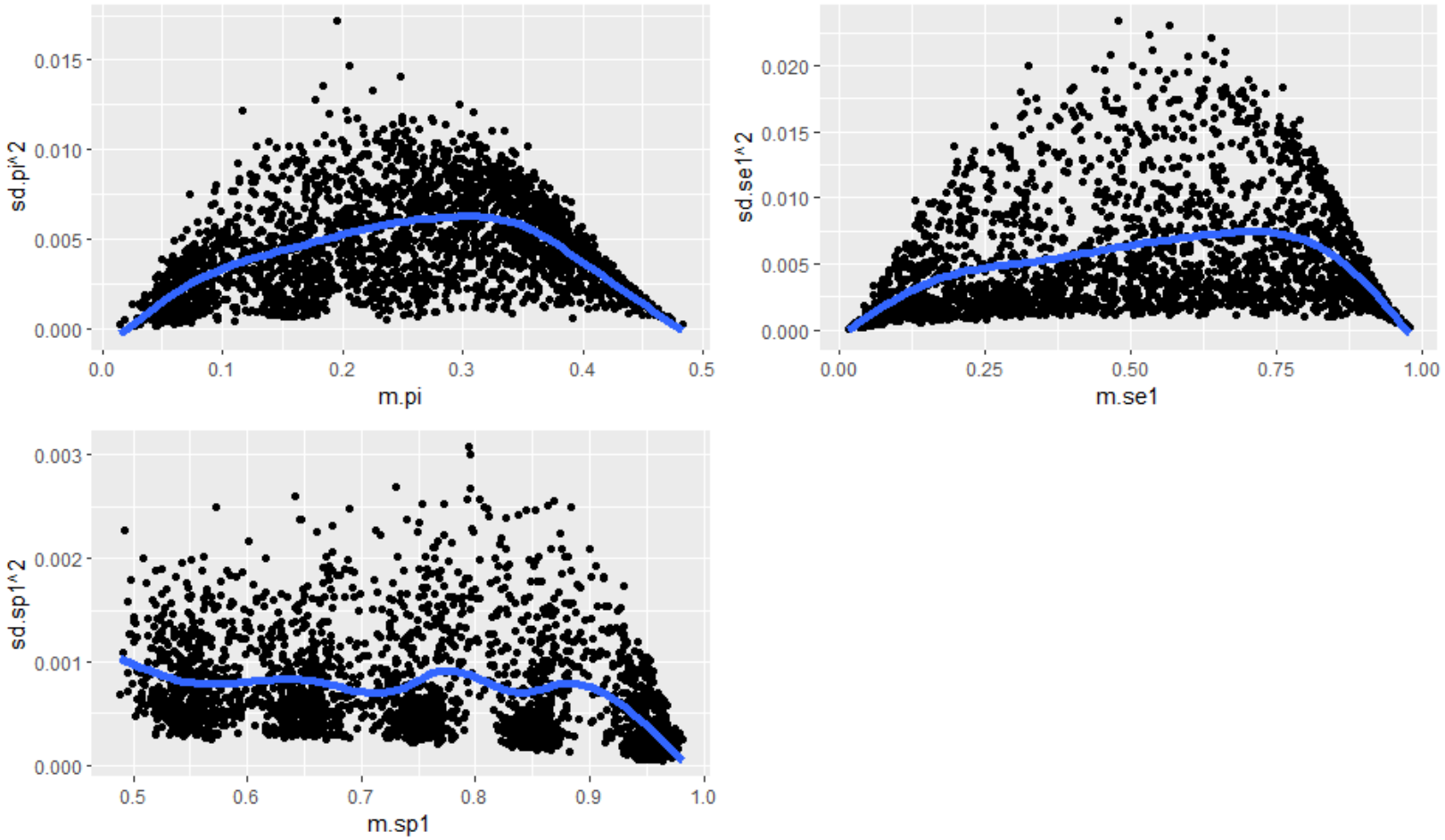

Figure 6-4: The relationship between the mean (`m.variable`) and variance (`sd.variable`) of the posterior inferences of P (top left), Se (top right) and Sp (bottom left) given data from a BLCM informed with imprecise priors and constrained truths.

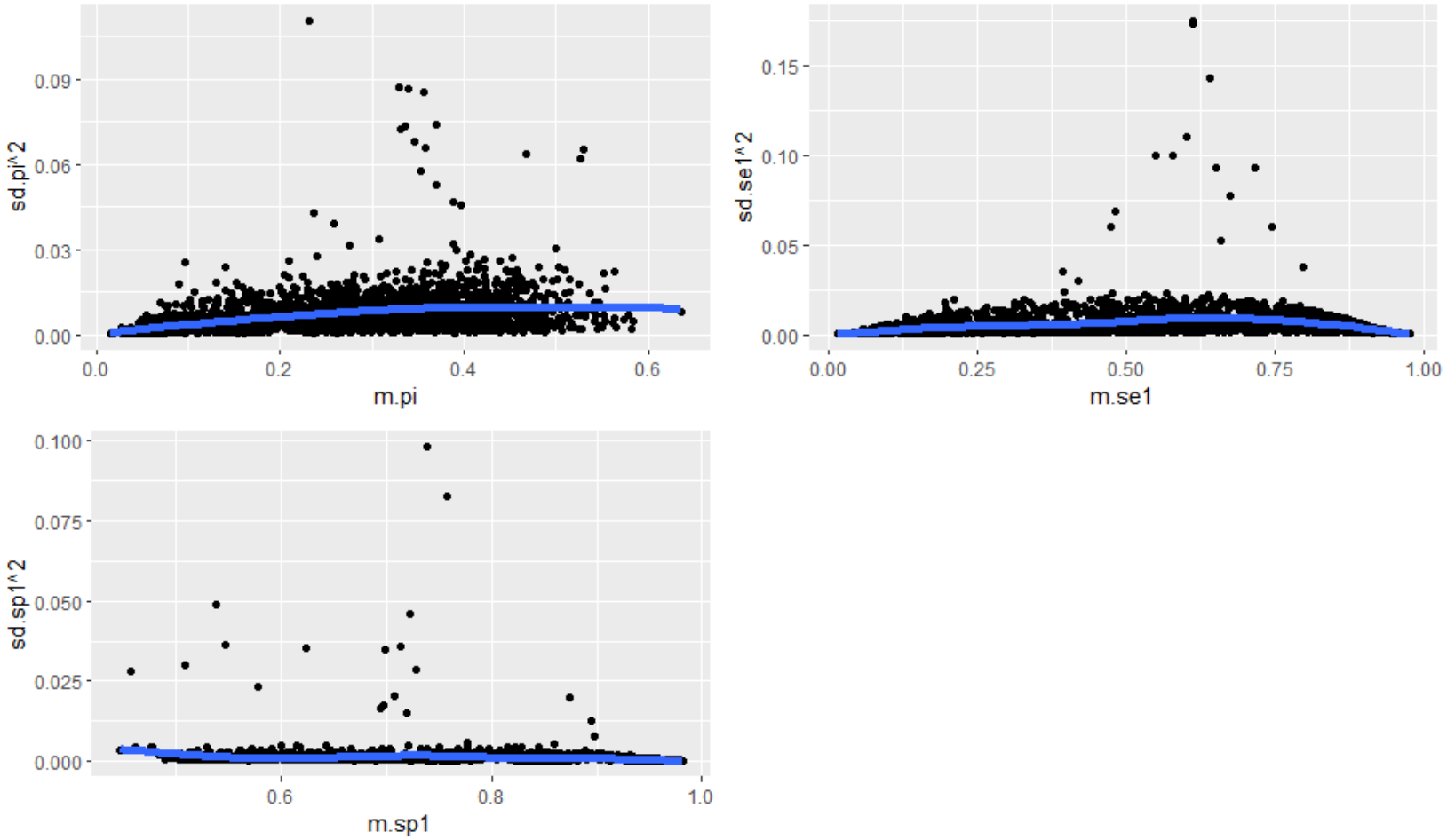

Figure 6-5: The relationship between the mean (`m.variable`) and variance (`sd.variable`) of the posterior inferences of P (top left), Se (top right) and Sp (bottom left) given data from a BLCM informed with imprecise priors and unconstrained truths.

Regressions using logit-transformed errors.

The LMM's referred to in this section are summarised in Table 10-14. General findings (a) to (f) are identified and discussed.

(a) Not absolute error values appear unreliable.

Three key observations have informed this finding:

1. In models that do not use absolute error values, increases in sample size also result in an increase in global errors and in the errors of Sehat, Sphat and Phat (LMM 1a-d, 2a-d).
2. The trend where error decreases as the number of diagnostic tests increase—termed the "n.tests trend" for brevity—is not always apparent in regressions when error is not absolute. For example, when prior precision is imprecise, the error of Sehat increases as the number of diagnostic tests increases (LMM2c), and this association is stronger by one order of magnitude than when priors are uniform (LMM3c).
3. When errors are absolute, and prior information is normally distributed, the association between the response variable, and the resulting interaction between the number of diagnostic tests and extreme parameter space is always negative in direction; yet when errors are not absolute, the relationship is either positive or negative with no clear pattern.

(b) Applying the logit function influences how error is interpreted in extreme parameter space.

In every LMM that uses absolute values (LMM 1e, f, g, h, 2, 32, 26, 29) error decreases as the number of diagnostic tests increases. However, logit transformations of the errors of Phat appear to have positive relationships with

extreme parameter space, and breakdowns in the n.tests trend are apparent in this space.

(c) Logit-transformed errors possess complicated dependencies with constraint and informative priors.

Examples of these complicated relationships include the positive relationship between the logit-transformed global error and the logit-transformed errors of Phat and constraint (LMM1a and 1b), and the positive relationship between the similarly transformed errors of Sehat and Sphat with prior precision (LMM1c and 1d). Although it may be possible to avoid these associations by altering how prior information is given to model, logit-transformed errors are not straightforward to interpret in this “latent space”—i.e., a multi-dimensional space containing transformed parameter values that cannot be directly interpreted, but that encodes a meaningful representation of a parameter space (Hoff, Raftery & Handcock, 2002). Consequently, it is not recommended that ecologists explore latent spaces to remove edge effects, as they are not readily interpretable.

(d) The n.tests trend is dependent on whether the inferred parameter value is in extreme space.

The n.tests trend breaks down with the presumed difficulty in sampling and or inferring extreme volumes of parameter space. Although the n.tests trend has already been established (Goodman, 1974) and discussed in previous chapters, this present chapter shows that this dependency is affected by extreme parameter space, and is not directly influenced by the logit-transformation of error (LMM’s 1a-h, 2a-h 3a-h). The n.tests trend is found to weaken at the edges of parameter space: for example, regression analyses which detect the n.tests trend also show that Phat increases in error when its values are

associated with extreme volumes of parameter space (e.g. LMM32, LMM1b; LMM1f).

(e) Relying on global errors can lead to incorrect conclusions about extreme parameter space.

The direction of association between the number of diagnostic tests available and extreme parameter space is chiefly influenced by parameter-specific errors. For example, the direction of association between the number of diagnostic tests and extreme parameter space is not the same between Phat, and Sehat and Sphat, when logit errors are used. Utilising global errors removes these important directions of association.

(f) Error increases as the number of diagnostic tests available increase when uniform priors are used, trumping the need to consider edges.

In general, this statement (statement f) is true when:

1. Global errors are absolute (LMM20).
2. The errors of Sehat are absolute (LMM3g).
3. Global errors and the errors of Sehat and Sphat are not absolute, but logit-transformed (LMM 2c, 3a, 3c, 3d).

Fitted versus residuals

In total, fitted versus residual plots were constructed for each parameter Sehat, Sphat, Phat and global error, for each transformation of error described in Table 6-1, and for Rule 1 and Rule 2. For illustrative purposes, all fitted versus residual plots for the not absolute and not logit-transformed errors can be found in Figure 6-6 and Figure 6-8 (where the Rule 1 method is applied) and Figure 6-7 and Figure 6-9 (where the Rule 2 method is applied). All remaining fitted

versus residual plots can be found on GitHub at https://github.com/annabush/PhD.

The fitted versus residual plots show that the mean of the residuals generally remains constant, confirming that the key artefact is the distribution of variance. In general, when error values are absolute, heteroscedasticity is inevitable, and the logit transformation has limited impact on this observable non-constant variance. This suggests that the rules of LMM's are violated when error is absolute, however it is reasonable to conclude that the LMM's in this thesis that employ absolute error are still reliable for the following reasons:

1. Most real-world data is heteroscedastic, and the goal of thesis is to validate BLCMs for use in the real-world.
2. The sample size is large enough to ensure that the regression fit is precise enough.
3. There was no assumption that errors remain constant across parameter space.
4. Unbiased estimates for the relationship between the predictor and response are still provided, as only relative relationships are examined, and no frequentist tests of significance are relied upon to produce these estimates.

It appears that the logit transformation of error deals with a proportion of the variance that makes analyses on the probability scale hard to trust, i.e. it provides a degree of normalisation. Consequently, it is believed that the logit transformation improves homogeneity of variance to some extent. For example:

1. When error is absolute, the logit transformation helps visually distinguish between values on edges, and those that are not on edges. For example, see Figure 6-8 and in Figure 6-9.
2. The logit transformation also helps visually discern differing patterns of variance between the errors of Sehat, Sphat and Phat.
3. The logit transformation removes positive relationship between fitted and residual values that exists for mean global and Phat errors when this transformation is not applied.
4. The logit transformation removes the peculiar, clustered relationships between fitted and residual values that exists when the errors of Sehat and Sphat are response variables in non-logit situations. However, relying on the global statistic also removes need to consider clustering.

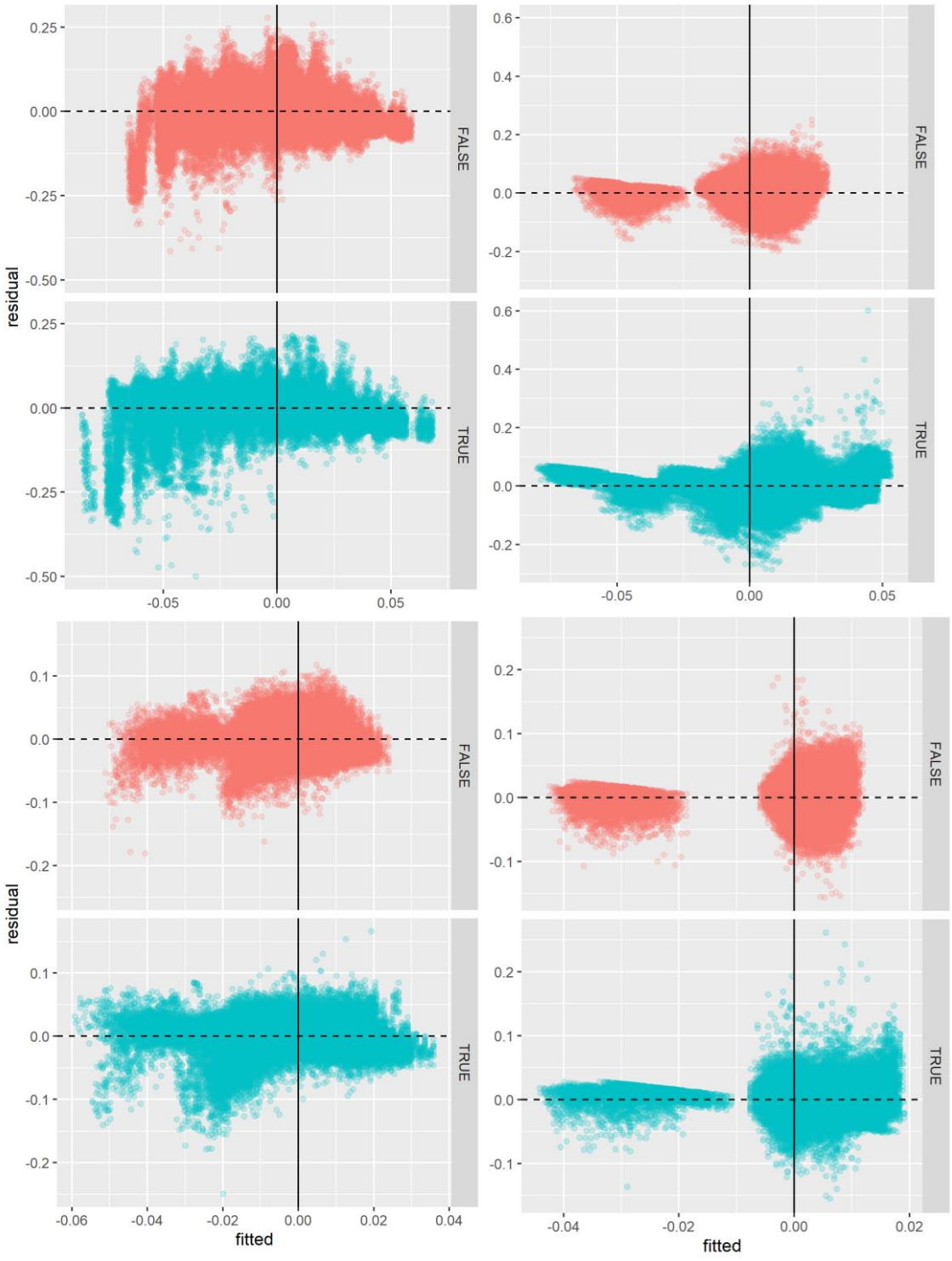

Figure 6-6: The fitted versus residual values for P (top left), Se (top right), the global statistic (bottom left) and Sp (bottom right) when error is not absolute and not logit-transformed, drawn using Rule 1, where TRUE and FALSE indicate whether the data point sits near the “edge” of parameter space.

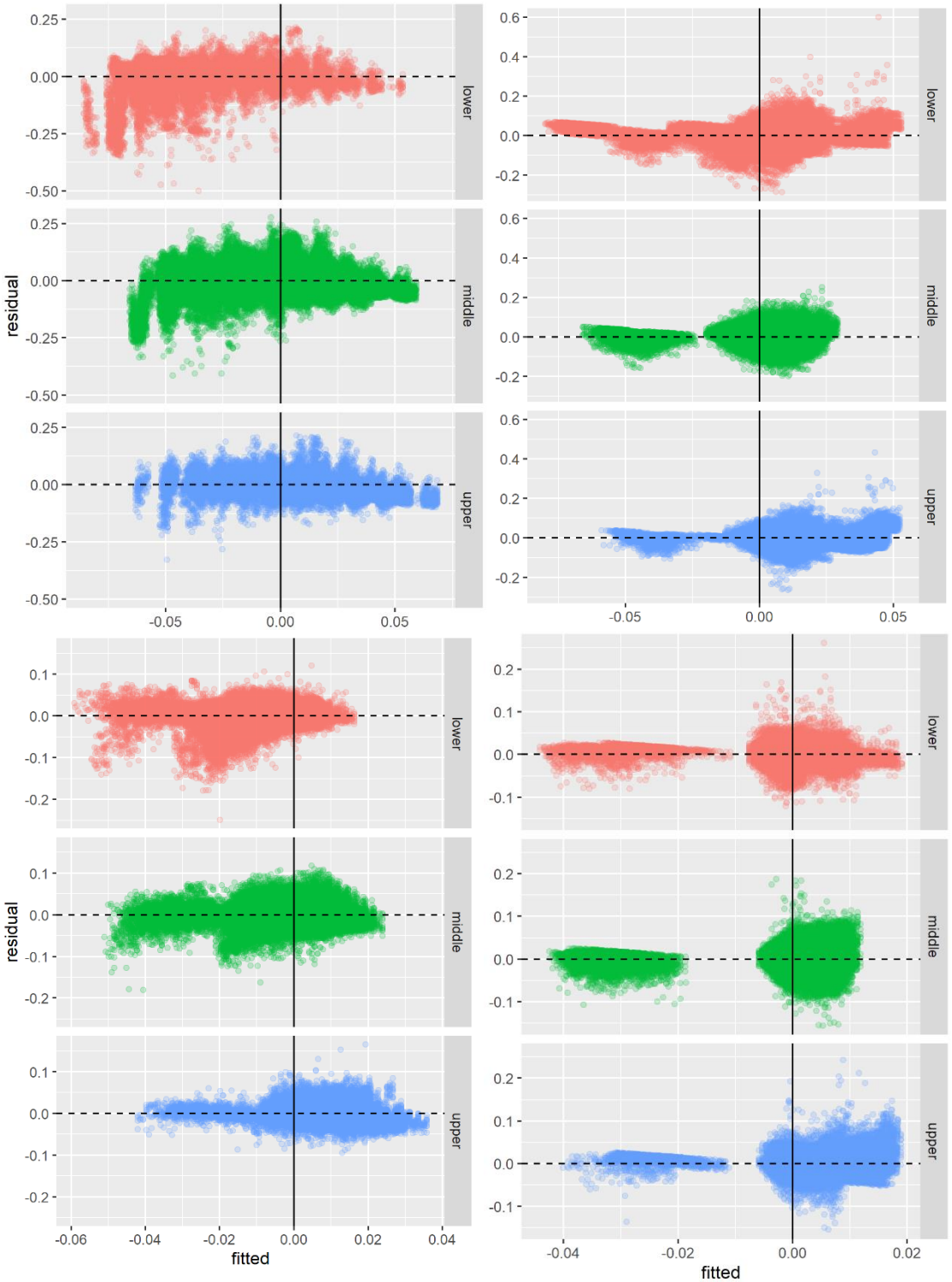

Figure 6-7: The fitted versus residual values for P (top left), Se (top right), the global statistic (bottom left) and Sp (bottom right) when error is not absolute and not logit-transformed, drawn using Rule 2, where "`upper`" = (Se > 0.9 and Sp > 0.9); "`lower`" = (Se < 0.1 and P < 0.1); and "`middle`" = all other space.

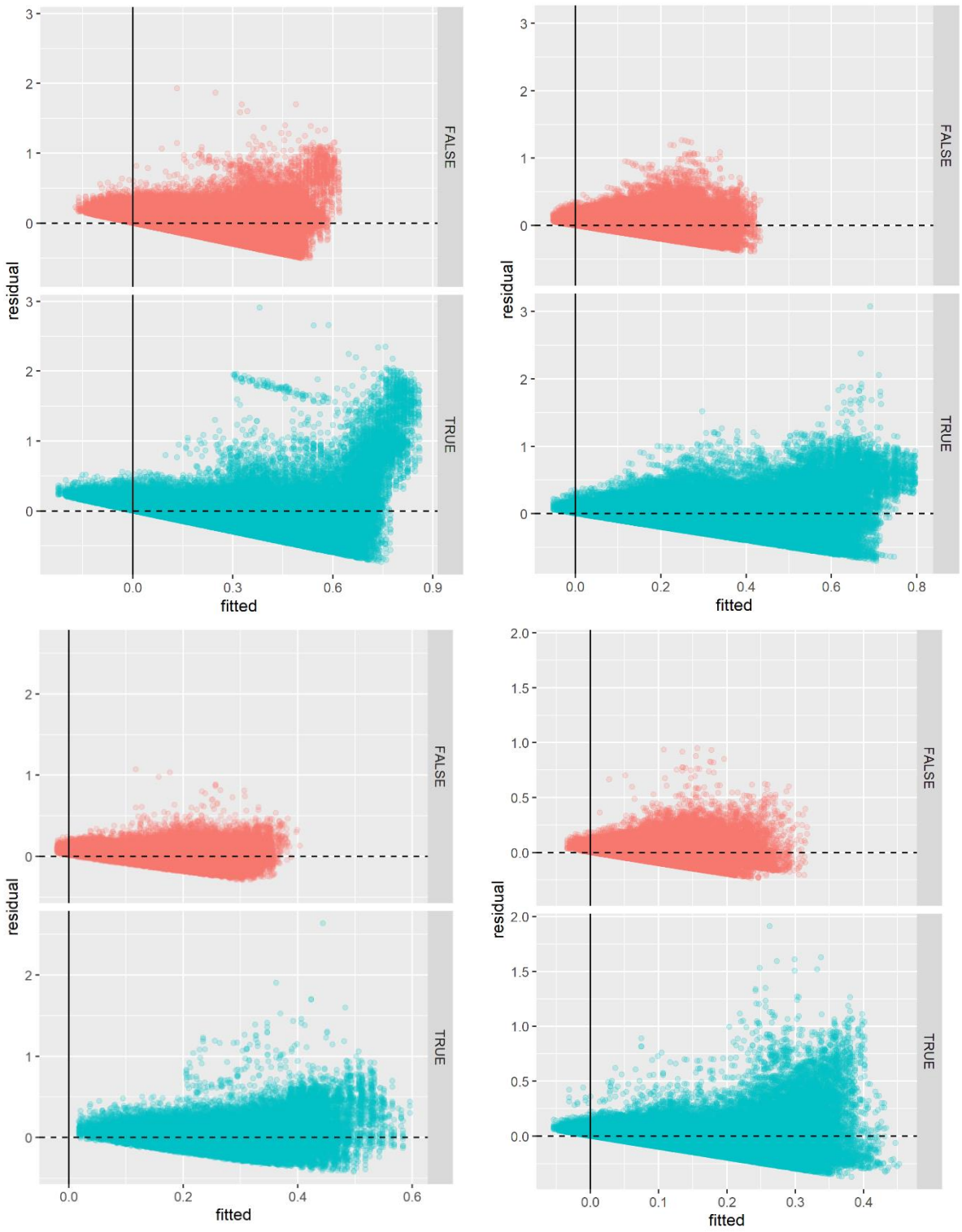

Figure 6-8: The fitted versus residual values for P (top left), Se (top right), the global statistic (bottom left) and Sp (bottom right) when error is absolute and logit-transformed, drawn using Rule 1, where TRUE and FALSE indicate whether the data point sits near the “edge” of parameter space.

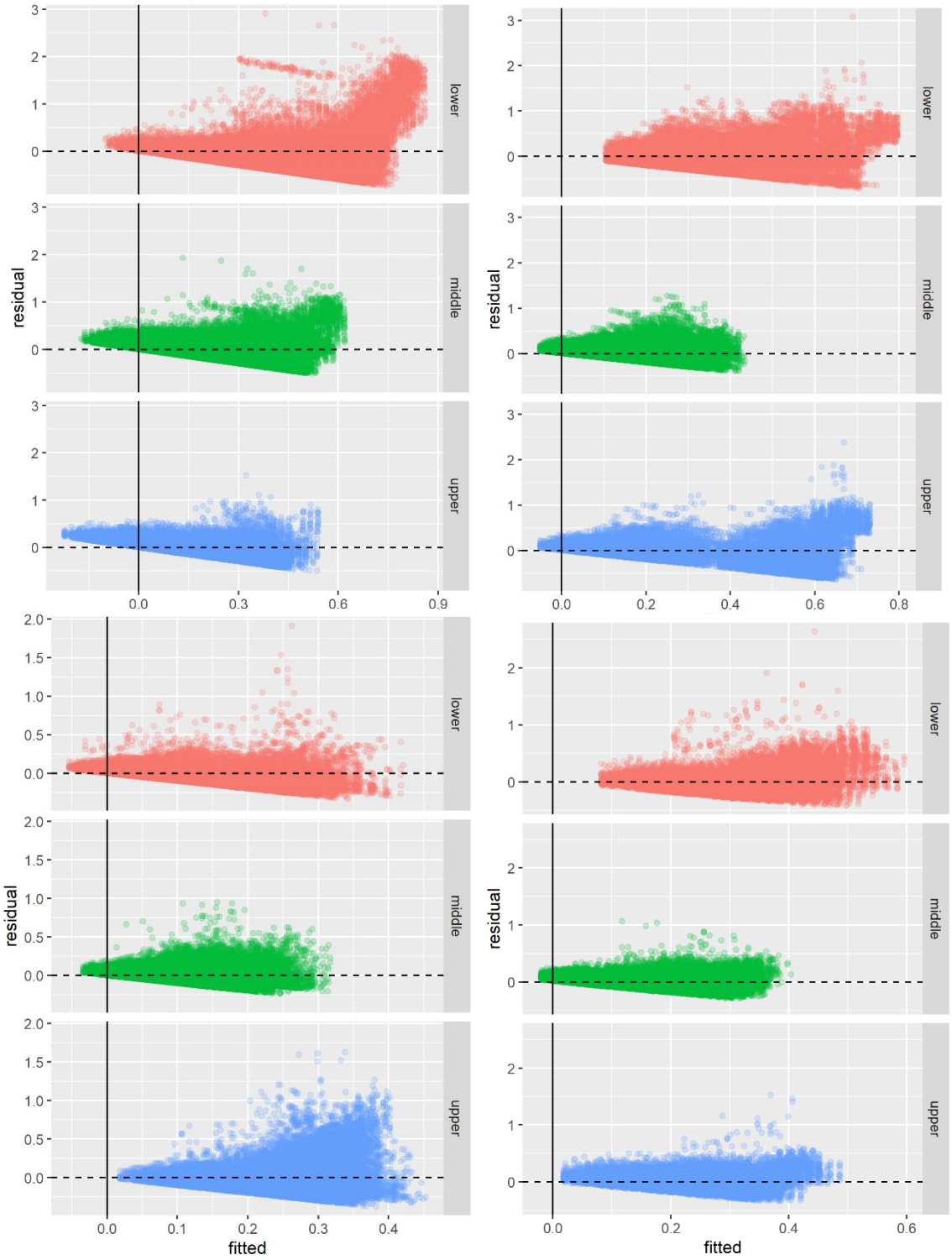

Figure 6-9: The fitted versus residual values for P (top left), Se (top right), the global statistic (bottom left) and Sp (bottom right) when error is absolute and logit-transformed, drawn using Rule 2, where "`upper`" = (Se > 0.9 and Sp > 0.9); "`lower`" = (Se < 0.1 and P < 0.1); and "`middle`" = all other space.

Analyses of hypotheses 1 to 5.

Each of the five hypotheses outlined above is now discussed in turn, in the context of these results.

HYPOTHESIS 1: Inferences are associated with more error as truths approach the edges of parameter space, and prior information affects this relationship.

The results show that the relationship between error and the number of diagnostic tests available is dependent on both extreme parameter space and an interaction effect between the number of diagnostic tests available and extreme parameter space. These dependencies are not constant between modelling scenarios, including those that have varied prior information, and are parameter-specific. For example, the relationship between the errors of Phat and the number of diagnostic tests is opposite to the relationship between the errors of Phat and the interaction effect between the number of diagnostic tests available and extreme parameter space, suggesting identifiability issues in extreme parameter space which result in the collapse of the n.tests trend.

Notwithstanding the above, prior information in general is found to affect the relationship between edges and the n.tests trend. For example, when prior information is removed, the n.tests trend reverses, suggesting the collapse of the n.tests trend in non-identifiable situations, but also that given enough prior information edge effects may be identifiable.

HYPOTHESIS 2: Edge effects represent a reduction in model power.

When models are uninformed, the need to address edge effects seems to be overridden by the need to address more systemic identifiability issues. When prior information is provided, the variance of errors appears parameter-specific.

For example, when models are informed by precise rather than imprecise priors, the errors of Phat exhibit the most variance whereas when models are constrained, rather than unconstrained, the errors of Sehat and Sphat exhibit the most variance. It is suggested that a lack of prior information contributes to the presence of edge effects.

HYPOTHESIS 3: The variance of the error of Sehat, Sphat and Phat across parameter creates edge effects.

The mean-variance relationships of the errors of Sehat, Sphat and Phat do not exhibit a constant variance at the edges of parameter space. And although the definition of an edge used in this thesis captures the most pronounced of the non-constant variance, the relationships plotted suggest that edge effects are larger when less information is provided.

HYPOTHESIS 4: Both model constraints, and whether the errors of the inferred values are absolute or not, influence the homogeneity of variance across parameter space.

It is found that decisions on whether models should be informed by constraint or prior precision, and whether the resulting inferred parameters should be interpreted using the various transformations described in Table 6-1, are not straightforward and do not follow consistent rules. This is because it is suspected that the variance in error is affected by dependencies—such as between the n.tests trend and extreme parameter space; and how differently Sehat, Sphat and Phat react to prior precision and or constraint—in complex ways. Moreover, when conducting regression analyses there is a well-cited risk that linking predictors to response via a function—in this case via the logit and

absolute transformations—can easily result in complications (Bolker *et al.*, 2009; Harrison *et al.*, 2018)

HYPOTHESIS 5: Edge effects can be removed from analyses by transforming errors using the logit link function.

Logit-transformed errors offered these analyses some improvements in the homogeneity of variance, but it did not eliminate heteroscedasticity. Despite this, the LMM's used are still trusted not least because frequentist metrics such as p-values and 95% confidence intervals are not analysed or reported on (Cleasby and Nakagawa, 2011). And since the underlying mechanisms contributing to the observed heteroscedasticity were unclear, no further manipulations—such as investigating heteroskedasticity-consistent standard errors—were carried out. Despite these limitations, the linear regressions presented remain useful for understanding the relationships between error and predictors, with the caution that dependencies are difficult to interpret when logit-transformed.

Discussion

Twenty years ago, it was purported that "*biological studies, even experimental ones, will often only explain a very small amount of variance*" (Møller and Jennions, 2002) and that "*ecologists using statistical models are explaining roughly half of the variability in dependent variables in their studies*" (Peek *et al.*, 2003). To-date, the need to better understand different types and reasons for variance in ecological models still remains (Mitchell, Beckmann and Biro, 2021), and there is a widely-shared view in ecology that "*a major unsolved problem in ecology is resolving the relative importance between different types and scales of variability to ecological processes*" (Holyoak and Wetzel, 2020). Accordingly,

this chapter has attempted to contribute to the understanding of variance in Sehat, Sphat and Phat across parameter space.

The experiments presented here examine the shape of the mean-variance relationships belonging to Phat, Sehat and Sphat, as well as the direction and magnitude of the predictors for these values. The conclusion is that at edges, ecologists should consider the errors belonging to the inferences of Phat, Sehat and Sphat separately; cautioning that while the analyses on global error reported on within Chapter 5 remain relevant, the global statistic is not to be used in extreme parameter space. Furthermore, the evidence presented here also suggests how the absolute and logit-transformed errors of Sehat, Sphat and Phat behave as modelling conditions vary, concluding that these transformations do not resolve issues of heteroscedasticity (for detailed figures see https://github.com/annabush/PhD).

By looking at the inferences of Sehat, Sphat and Phat separately, it has been demonstrated that edge effects are relevant statistic artefacts for ecologists to consider in their analyses. The *shape* of the mean-variance relationships belonging to Sehat, Sphat and Phat all exhibit some degree of heteroscedasticity, and were found to be highly distinctive given any form of prior information. It has been shown that the mean-variance relationship of Sehat is only heteroscedastic at edges; the mean-variance relationship of Phat is positively correlated and clustered by the number of diagnostic tests; and the mean-variance relationship of Sphat values are negatively correlated. These relationships indicate that extreme values of Sehat may suffer from non-standard error structures; that precise inferences of Phat are dependent upon the number of diagnostic tests available; and that as the mean of the value of Sphat increases, the inference of Sphat tends to become more precise.

Yet the *size* of the edge effects relating to inferences of Sehat, Sphat and Phat is dependent on the amount of prior information provided. With less information, inferences of Phat suffer the most variance compared to inferences of Sehat and or Sphat—and in precise scenarios, this effect size is several orders of magnitude—providing solid evidence that compared to Sehat and Sphat, Phat is the most difficult parameter to infer accurately particularly since inferences of Phat in extreme parameter space are most affected by constraint. Moreover, discovering that the shape of the mean-variance relationship is heavily biased, including by outliers, when none to very little prior information is provided supports the theory that heteroscedasticity is very context specific, and dependent on random effects (Schielzeth et al 2020).

The dependency between edge effects and prior information, however, sits in agreement with the findings presented in Chapter 5, namely that when errors of Phat are small, they are not associated with edge effects; that errors of Sehat are associated with edge effects when prior precision is precise but not when constraint is applied; and that the errors of Sphat are most affected by edge effects when the values of both Se and Sp are comparatively high, i.e. close to the value of one.

Broadly, the experiments presented have shown that analysing the magnitude and direction of the dependencies on error using regressions is crucial. And, that the logit transformation reduces some of the heteroscedasticity associated with the regressions that model responses on absolute error, but that logit-transformed errors create difficult to decipher interactions, particularly between prior information and error. For instance, it was found that the effect of extreme parameter space on the errors of Sehat, Sphat and Phat, and the interaction

effect between extreme parameter space and the number of diagnostic tests available, was affected by the logit transformation in complex ways.

Importantly, in studies that seek more accurate inference at the upper edges of parameter space, improving the number of diagnostic tests is particularly important. Three key findings support this suggestion and are essential takeaways from this chapter. First, the examinations of not absolute errors found that the n.tests trend—which this thesis assumes to be a proxy for model identifiability—disappears. Second, the examinations of absolute errors found that the n.tests trend conflicted with extreme volumes of parameter space, and that the errors of Phat are particularly influenced by this conflict. Finally, when there is no prior information, the n.tests trend reverses and overrides the importance of considering edge effects to achieve model identifiability.

Overall, ecologists often rely on mechanistic models to justify their findings, and rely on ecological justifications (Lindén and Mäntyniemi, 2011) to justify their model specifications. In contrast, this chapter uses LMM's as a mechanistic way to understand a dataset of errors across space, yet the justifications for the findings, and model specifications are largely statistical in nature.

Conclusion

So, what are the interactions between edge effects, BLCM identifiability, and the mean and variance of error? This chapter tested the assumption that extreme volumes of parameter space present identifiability issues using 5 hypotheses developed from the findings presented in Chapter 5. The resulting experiments investigated whether mean-variance artefacts exist across parameter space and examined the error structures at the edges of parameter space. Importantly, it was found that the shape of the mean-variance relationships of

the errors of Sehat, Sphat and Phat were each uniquely identifiable, but that the heteroscedasticity present was highly dependent upon modelling conditions, and random effects, and interpretability could not be substantially improved using transformations of error. It was found that identifiability issues were usually indicated by the absence of the n.tests trend, and that it difficult to interpret parameter space often occurs in extreme volumes of parameter space. Overall, the findings presented contribute to the understanding of variance in the errors of Sehat, Sphat and Phat across parameter space, and emphasise that patterns in these errors should not be neglected.

Where next? So far in this thesis, parameter space has been constrained by only allowing the true values of P and Sp to take certain values. This means that parameter spaces where Sp is less than 0.5, and where P is greater than 0.5 have not been examined in the experiments so far. Chapter 7 goes on to explore whether the findings of Chapters 4, 5 and 6 still hold when truths are simulated across unconstrained parameter space.

Chapter 7

7. Generalisability across parameter space

Introduction

In Chapters 4, 5 and 6, model validation exercises were demonstrated with—among other things—the aim of achieving “better” inferences of P using BLCMs. These model validation methods examined existing model fits based on the data available. In contrast to a model validation, a sensitivity analysis looks at the robustness of a model’s results given new information, or changes to its assumptions; and is accordingly the focus of Chapter 7.

Consider the four types of experimental design portrayed in Figure 7-1. Chapters 5 and 6 of this thesis have so far explored options C and D of those four options under the overarching assumption that the parameter space (as defined within the model) should be constrained to values between `0.5 (Sp1) x 0.5 (P) x 1 (Se1)` (see Table 10-2). And yet it is possible that in some instances, these restrictions may not be sufficient to allow the MCMC algorithm to yield enough solutions from the parameter space available, particularly if the cut-off for positive test results is disputed (Akobeng, 2007; Habibzadeh, Habibzadeh and Yadollahie, 2016), or if most of a population are infected.

This chapter goes on to test the performance of stochastic BLCMs under option A (Figure 7-1) when the assumption of a constrained parameter space is removed, meaning that true values are located within a much larger volume of parameter space.

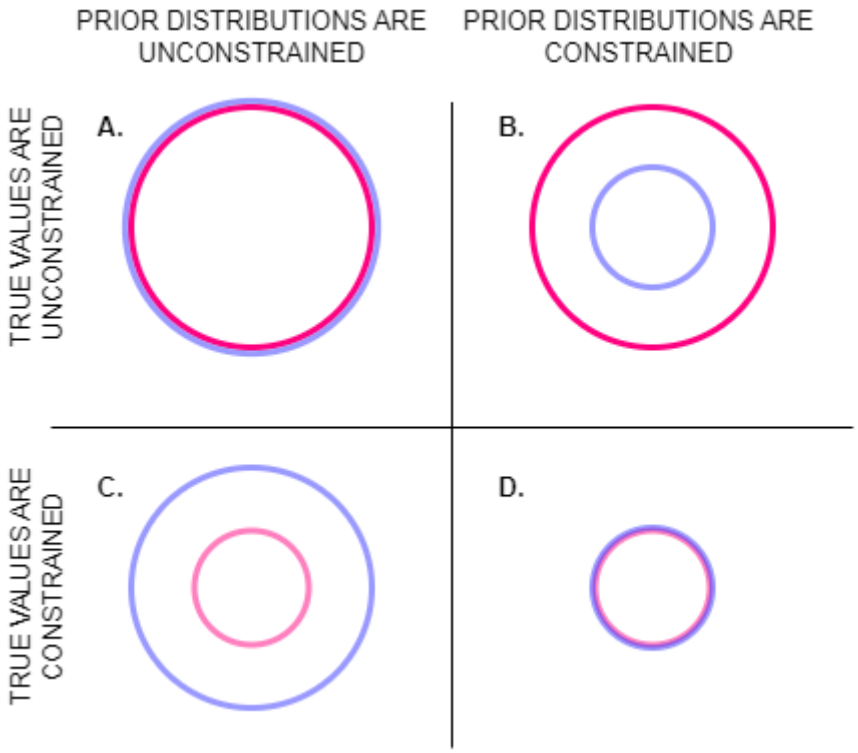

Figure 7-1 The relationship between the prior constraint applied (blue) and the true values that may be selected (red). For example, in this thesis an unconstrained three-dimensional parameter space has a volume of `1 x 1 x 1`, and a constrained parameter three-dimensional parameter space has a volume of `0.5 x 1 x 0.5`.

In Figure 7-1, the red lines show the volume of parameter space available as a proportion of the space where the model is informed to search for a correct inference (blue lines). Following this logic, the smaller circles are one quarter of the size of the large circles. In Chapters 5 and 6, hypotheses were investigated using partial parameter space, thus, options C and D were the experimental designs considered. In the present chapter, option A is investigated, and so a sensitivity analysis is conducted across global parameter space. Option B is not a model setup worth investigating because the truth may lie outside of the given prior distributions.

Why generalise?

In ecology, model identifiability is generally considered in terms of unique sets of parameter values that have been calibrated to maximise the likelihood of a model under certain assumptions. But for models generated within a stochastic framework, not defining the "generality" of the findings (Spake *et al.*, 2022) can be a dangerous practice, as partial observations often result in misleading likelihoods (Vernon *et al.*, 2018; Stocks, Britton and Höhle, 2021).

To generalise their findings, ecologists are often familiar with conducting "local" sensitivity analyses by varying one parameter and or value at a time (Naujokaitis-Lewis *et al.*, 2009; Olsen *et al.*, 2022) while others are fixed (Xu *et al.*, 2004), as shown in Validation Example A of this thesis in Chapter 4. However, in studies across high-dimensional space, confidence in BLCM specifications and assumptions should be generalisable across the wide range of possible truths that may be encountered in nature; and to achieve this, Global Sensitivity Analyses are required. Accordingly, the Global Sensitivity Analysis presented in this chapter is highly relevant for ecologists wanting to determine whether their simulation models are sufficiently robust to new information, or to changes in model assumptions.

Prior to using BLCMs on diseased populations of wildlife, it is therefore essential to subject the model to a Global Sensitivity Analysis (Wagner, 1995), which tests parameters across the full range of "total predictive uncertainty" (Cariboni *et al.*, 2007), i.e. new information, or changes in model assumptions. However, it has been noted that "*the widespread application of GSA [Global Sensitivity Analyses] in ecological models has been hindered because the model output can be unwieldy and methods of analyzing these data can be computationally intensive.*" (Harper, Stella and Fremier, 2011) The term

“generalisability” is itself loosely defined in ecology (Spake et al 2022); and in this thesis is used synonymously with the term Global Sensitivity Analysis.

As described in Figure 7-1, so far in this thesis, only one quarter of the possible three-dimensional parameter space of Se, Sp and P have been examined due to the constraint—of Sp to values greater than 0.5 and of P values to less than 0.5—within each model. And in general, accepted practice among ecologists is that for most diseases, the constraint of truths avoids misclassifying more of the largest group, infected or uninfected, whichever that may be (Rydevik, Innocent and McKendrick, 2018).

Consequently, this thesis has not yet reported on how the Any-Test, Any-Population model will perform in “any” testing scenario and cannot determine whether the BLCM is robust until the error structure of an up to 11-dimensional parameter space has been examined. Furthermore, understanding uncertainty in the 3/4 of remaining parameter space is likely to be important for ecologists who wish to study diseases with high values of P—i.e., P values above 0.5—that threaten extinctions, such as DFTD (McCallum *et al.*, 2009). This is because the diagnostic test array will be dominated by true positives and false negatives, with fewer true negatives and false positives, forcing a higher estimate of Se and a lower estimate of Sp (Helman *et al.*, 2020). Consequently, it is necessary for this chapter to explore if the methodology demonstrated up to this point can be applied more broadly to the wide spectrum of wildlife diseases that exist in the wild.

Global Sensitivity Analyses of BLCMs

As ecologists, if the disease systems that we observe in ecology are only a subset of those that could possibly exist, our models must be practical, but

suitably flexible to deal with the diversity of testing scenarios that may be encountered. So far in the research underpinning this thesis, constraining the model using reasonable assumptions freed enough of the available computational capacity to allow the MCMC to complete the required number of iterations in order to ensure convergence.

This chapter now questions the assumptions behind the constrained parameter space used, by asking the higher-level question “when is it valid to fix and/or constrain truths?”. To do this, two experiments are reported on as follows:

EXPERIMENT 1: where constrained parameter space—i.e., Sp values greater than 0.5 and P values less than 0.5—is modelled using “new” truths. The purpose of this experiment is to evaluate the extent to which the findings reported in Chapter 4 to Chapter 6 are specific to the “original” truths, or generalisable given “new” truths.

EXPERIMENT 2: where unconstrained parameter space is explored by allowing the model to search for values of Se, Sp and P between the limits of 0 and 1 and using the same true values as assigned in Chapters 5 to 7.

So, what exactly do these two experiments test?

Experiment 1 tests the bias introduced by deciding the values of truths by searching a very different 3D “slice” of up to 11-dimensional constrained parameter space. And Experiment 2 tests how our conclusions on parameter uncertainty are affected by searching global parameter space, i.e. a fully unconstrained parameter space, where a truth may be any value on the probability scale.

Similarly, what are the specific questions examined?

1. How much influence do the fixed truths for tests two to five have on the error, bias, and standard deviation of Sehat, Sphat and Phat?
2. How does the error, bias and standard deviation of Sehat, Sphat and Phat change when values of Sp can be less than 0.5, and values of P be greater than 0.5?

Methods

Hypothetical modelling scenario

Experiment 1 tests a constrained BLCM under new truths (simulated dataset 4,Table 10-1), and Experiment 2 tests an unconstrained BLCM under the same truths (simulated dataset 5, Table 10-1) as deployed in Chapters 5 and 6. In Experiment 1, the new fixed truths for tests two to five were set to `Se2 = 0.33, Se3 = 0.68, Se4 = 0.44, Se5 = 0.43, Sp2 = 0.94, Sp3 = 0.56, Sp4 = 0.80`, and `Sp5 = 0.87`. For Experiment 2, the truths for tests two to five are the same as Chapters 5 and 6, and Se1, Sp1 and P were set to randomly selected values between 0 and 1. Simulated datasets 4 and 5 record how the model performs using normal and uniform priors, and the resulting dimensions of datasets 4 and 5 can be found in Table 10-1.

The intensive computational effort workload to generate the 720,000 observations (Table 10-1) across unconstrained high-dimensional parameter space for Experiment 2 was mitigated by running the required simulations over 20 cores on each of two servers of Exeter University's High Performance Computer at the same time.

Experiment 1

How much influence do the fixed truths for tests two to five have on the error, bias, and standard deviation of Sehat, Sphat and Phat?

For experiment 1, 4 × 18 heatmaps were plotted to demonstrate how statistics describing the global and parameter-specific values for error (Equation 16), bias, (Equation 17) and standard deviation respond to changes in four modelling conditions—constraint, prior precision, sample size, and the number of diagnostic tests available—with levels as previously defined (Table 5-1).

So far in this thesis, the research into the effect of sample size on error, bias, and the metric for precision has been limited, but it remains in the study design because Chapter 5 found that sample size has impacts on the variance of error and given that the availability of sample data from real-world wildlife disease studies is limited, any findings that could expand on this would be both relevant and informative.

It was suspected that changing the fixed truths for tests two to five to different values will disrupt the generalisations made across constrained parameter space in Chapter 6.

Experiment 2

How does the error, bias, and standard deviation of Sehat, Sphat and Phat change when values of Sp can be less than 0.5, and values of P be greater than 0.5?

For experiment 2, 4 × 12 heatmaps were plotted to demonstrate how the statistics describing the global metric, error (Equation 16), bias (Equation 17), and standard deviation change across parameter space given three modelling

conditions: prior precision, sample size and the number of diagnostic tests available, with levels as described in Table 5-1.

It is suspected that the bias of the global statistic will not be consistently overestimated or underestimated across global parameter space; that edge effects will remain across global parameter space; that the value of P as well as the number of diagnostic tests available dictates the error, bias, and standard deviation of the inferred values. This chapter reports on the hypothesis posed in terms of these four predictions.

Plotting

Heatmaps were scaled and plotted using the same methodology as described in Chapter 5 to provide a consistency across analyses. The full directory of 72 heatmaps used to generate the findings of this chapter can be found on GitHub (https://github.com/annabush/PhD).

Results

Note, values described as "high" or "low" are indicatory of their exact position in parameter space. A low value indicates a position closer to 0, and a high value indicates a position closer to 1.

How much influence do the fixed truths for tests two to five have?

The following results were obtained from analysing the 72 heatmaps generated; the following text summarises the high-level findings.

It was hypothesised that changing the true values of tests two to five would change the errors of Sehat and Sphat, and therefore contradict the generalisations stated earlier in this thesis regarding the original set of true values. Yet the findings from Experiment 1 support those reported and

discussed in Chapter 6, with no contradictions so far discovered. For example, see Figure 7-2, which shows the bias of Phat given the original fixed truths for tests two to five compared to Figure 7-3, which shows the bias of Phat given the new fixed truths of Experiment 1.

Experiment 1 provided the following high-level findings:

1. The value of P dictated how constraint affected inferences regarding the global statistic, and the errors of Phat, in complex but structured ways. In addition, the value of P also affected the size of the errors of Sehat and Sphat. For example, when the values of P were low, the errors of Sehat and Sphat were high. These findings support research by Berkvens *et al.*, 2006 suggesting that the only way to infer P is by introducing external knowledge through constraint.
2. Overall, constraint did not significantly decrease the errors of Sehat, Sphat and Phat in comparison to the provision of informative priors. However, when priors are uniform, constraint appears important to achieving accurate inferences regarding global errors. These findings are supported by Chapter 6, which reports that prior precision is the best way to quickly improve accuracy of inferences in normal models, and that constraint is the best way to quickly improve accuracy of inferences in uniform models.
3. For global inferences, when priors are normally distributed, Phat is consistently overestimated when values of P are close to 0.5, and consistently underestimated when values of P are close to 0; and this finding is consistent across modelling conditions. This finding is consistent with the finding in Chapter 6 that the biases in the inferences

of global error is largely independent of experimental conditions—and significantly, dependent on P.

4. Global standard deviation is dependent on the number of diagnostic tests and sample size, though the standard deviations of Sphat, Phat, and particularly Sehat are dependent on the number of diagnostic tests only (see Figure 7-4 and Figure 7-5). This finding was also corroborated by Chapter 6 where it was discovered that increasing the number of diagnostic tests available to the model was the best way of improving the precision of Sehat, Sphat and Phat for the 15% scenario described there. For parameter-specific inferences, precision is generally not affected by sample size, but is affected by the number of diagnostic tests available.

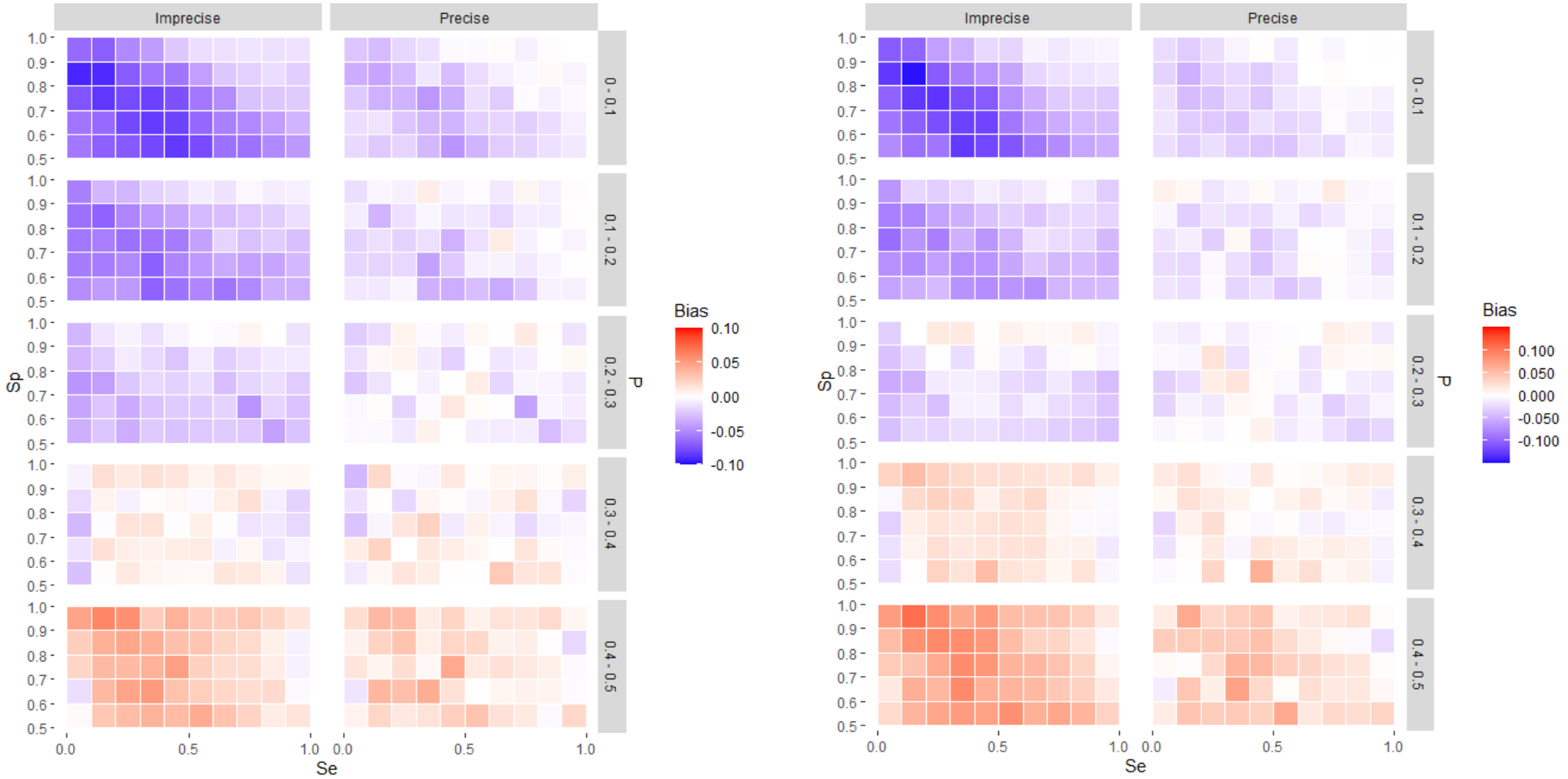

Figure 7-2: The heatmap panels on the left show the bias of Phat across constrained parameter space given the original truths, and the heatmap panels on the right show the bias of Phat across constrained parameter space given the new truths of Experiment 1.

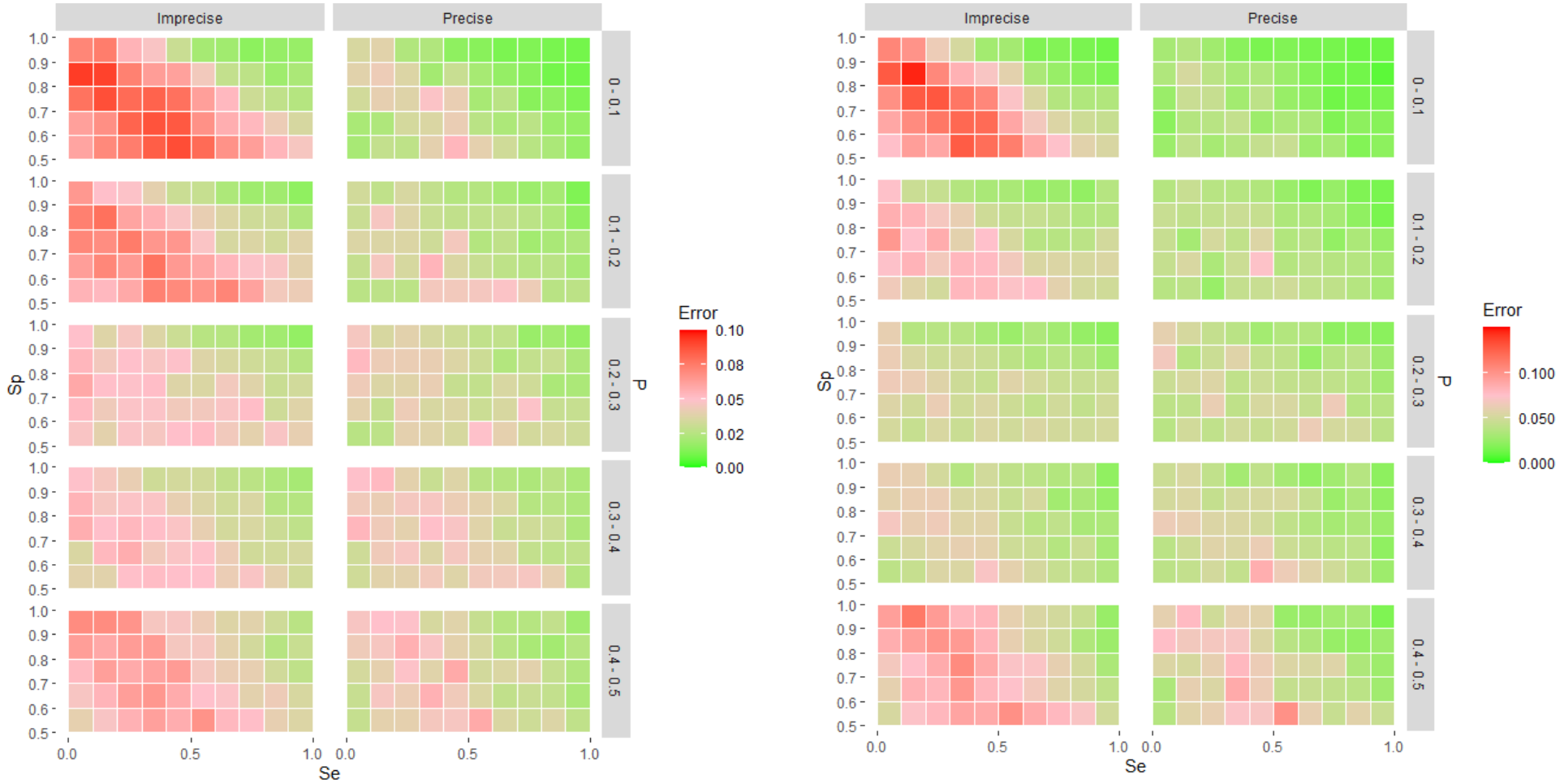

Figure 7-3: The heatmap panels on the left show the error of Phat across constrained parameter space given the original truths, and the heatmap panels on the right show the error of Phat across constrained parameter space given the new truths of Experiment 1.

Figure 7-4: A panel of heatmaps showing the standard deviation of Phat across batteries of two to five diagnostic tests in constrained parameter space given the original truths.

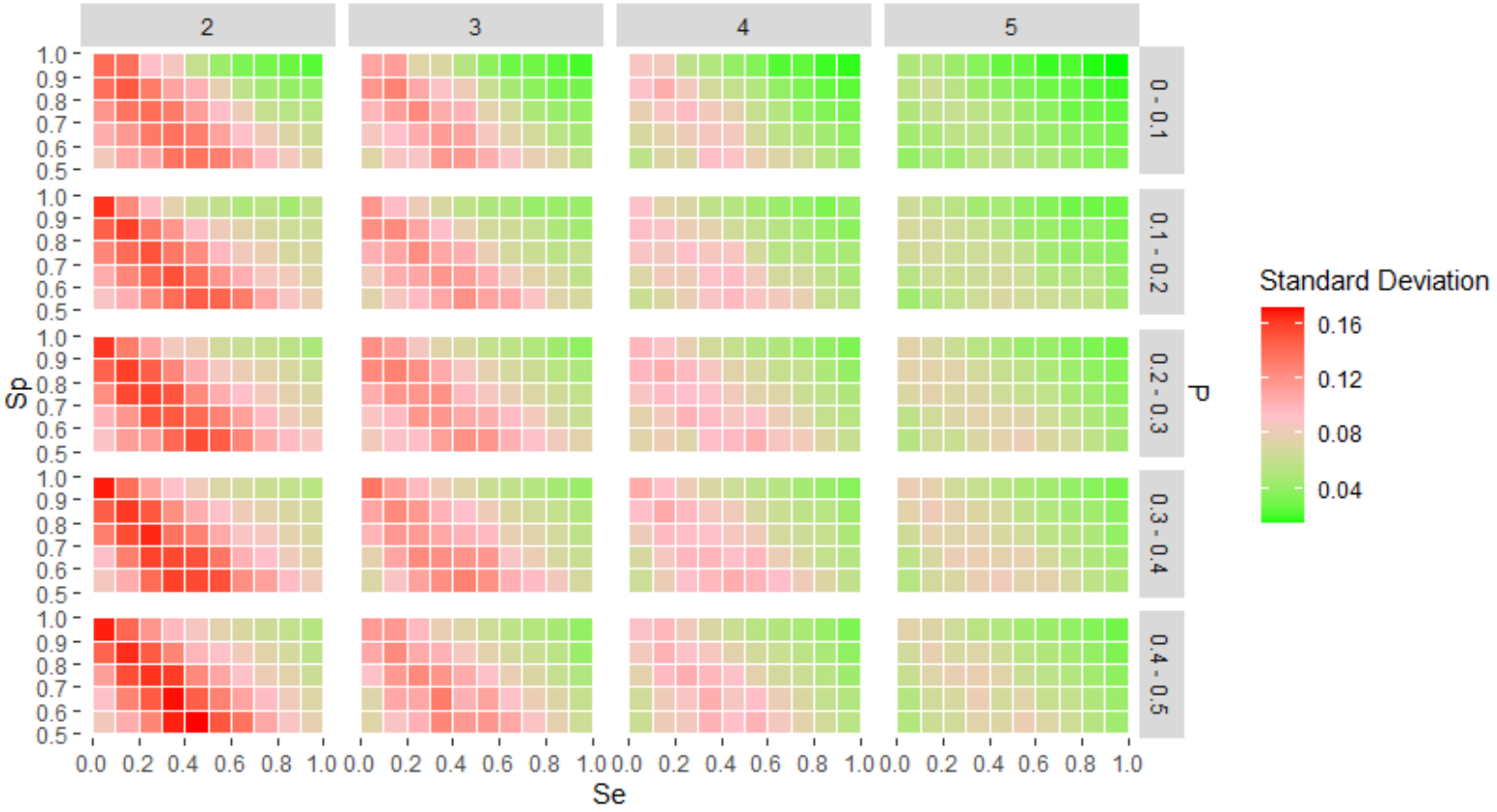

Figure 7-5: A panel of heatmaps showing the standard deviation of Phat across batteries of two to five diagnostic tests in constrained parameter space given the new truths of Experiment 1.

How does the error, bias, and standard deviation of Sehat, Sphat and Phat change when Sp is less than 0.5, and P is greater than 0.5?

The answers to this research question, which form the Global Sensitivity Analysis, are provided to address the predictions that the bias of global statistics will not be consistently overestimated or underestimated across global parameter space; that edge effects will remain across global parameter space; that the value of P as well as the number of diagnostic tests available dictates the error, bias, and standard deviation present.

Accordingly, four sub-questions are reported on that are now listed:

1. Where is global parameter space being overestimated and underestimated?
2. Do edge effects exist in global parameter space?
3. Does P dictate error, bias, and standard deviation?
4. How much influence does the number of diagnostic tests have on error, bias, and standard deviation?

Where is global parameter space being overestimated and underestimated?

Across unconstrained parameter space, four observations can be made from the 48 heatmaps that were plotted from the inferences associated with Experiment 2:

1. When P is greater than 0.6 global error is overestimated, and when P is less than 0.4, global error is underestimated.
2. The observation described in (1) is even stronger when the errors of Phat are plotted.

3. The observation described in (1) does not exist when the errors of Sehat and Sphat are plotted.
4. When P is greater than 0.6 the errors of Sehat are more likely to be overestimated or underestimated, however when P is less than 0.6 the errors of Sphat are more likely to be overestimated or underestimated.

Do edge effects exist in global parameter space?

The linear edge effects that this thesis reports on are in relation to constrained parameter space. In unconstrained parameter space, the errors of Sehat and Sphat have a horizontal symmetry on a plane of P = 0.5, with a 90° rotation of each panel within a facetted heatmap (for example, see Figure 7-6 in comparison to Figure 7-7), which is an observation that involves but is not exclusive to “extreme” inferred values.

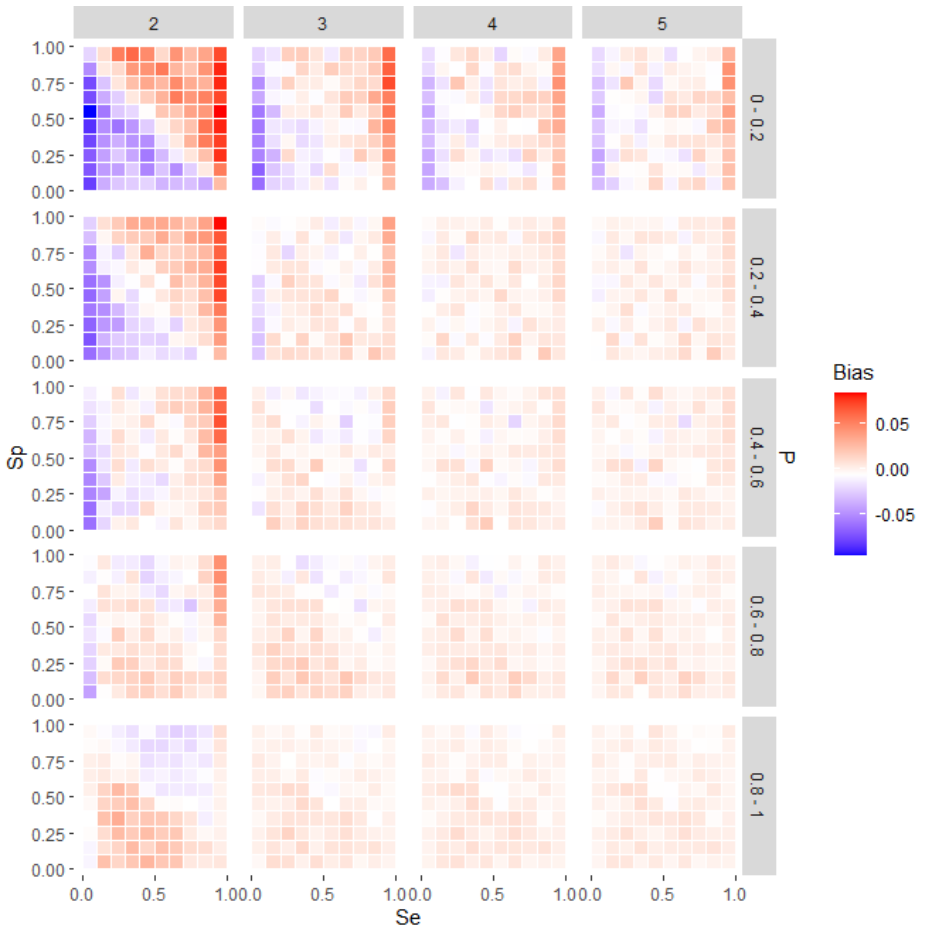

Figure 7-6: The bias of Sehat across unconstrained parameter space for batteries of two to five tests. This figure relates to Experiment 2.

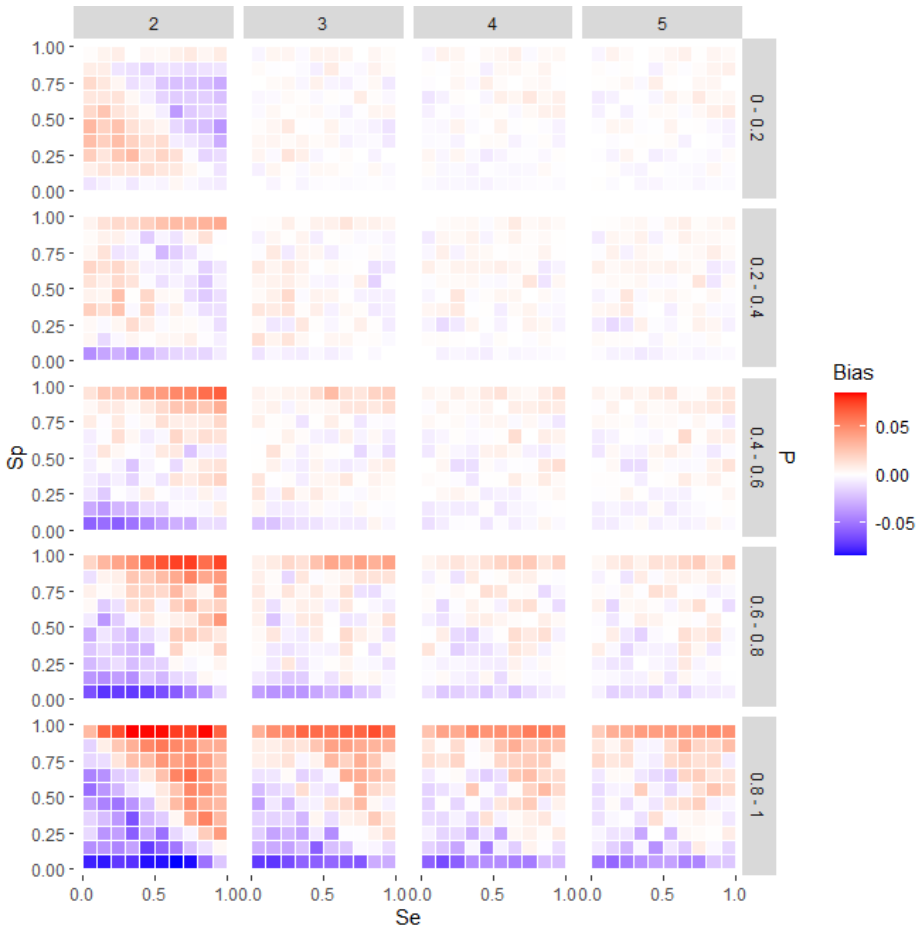

Figure 7-7: The bias of Sphat across unconstrained parameter space for batteries of two to five tests. This figure relates to Experiment 2.

Does P dictate error, bias, and standard deviation?

The findings of Experiment 2 indicate that the uncertainty of Phat when inferred using uniform priors are similar to those obtained using normal priors when the value of P is close to 0.5 (Figure 7-8 and Figure 7-9). However, an analysis across the full directory of heatmaps (Plot A to Plot L) associated with inferences of Phat for Experiment 2 more generally suggest that Phat cannot be correctly inferred if the sum of Se and Sp are close to 1, possibly due to the label switching problem discussed in Chapter 3.

How much influence does the number of diagnostic tests have on error, bias, and standard deviation?

While the errors of Phat are highly dependent on the number of diagnostic tests available, the errors of Sehat and Sphat are highly dependent on both the number of diagnostic tests available and values of P that are less than 0.2 and greater than 0.8. The errors of Sehat are higher when P is lower, and the errors of Sphat are higher when P is higher.

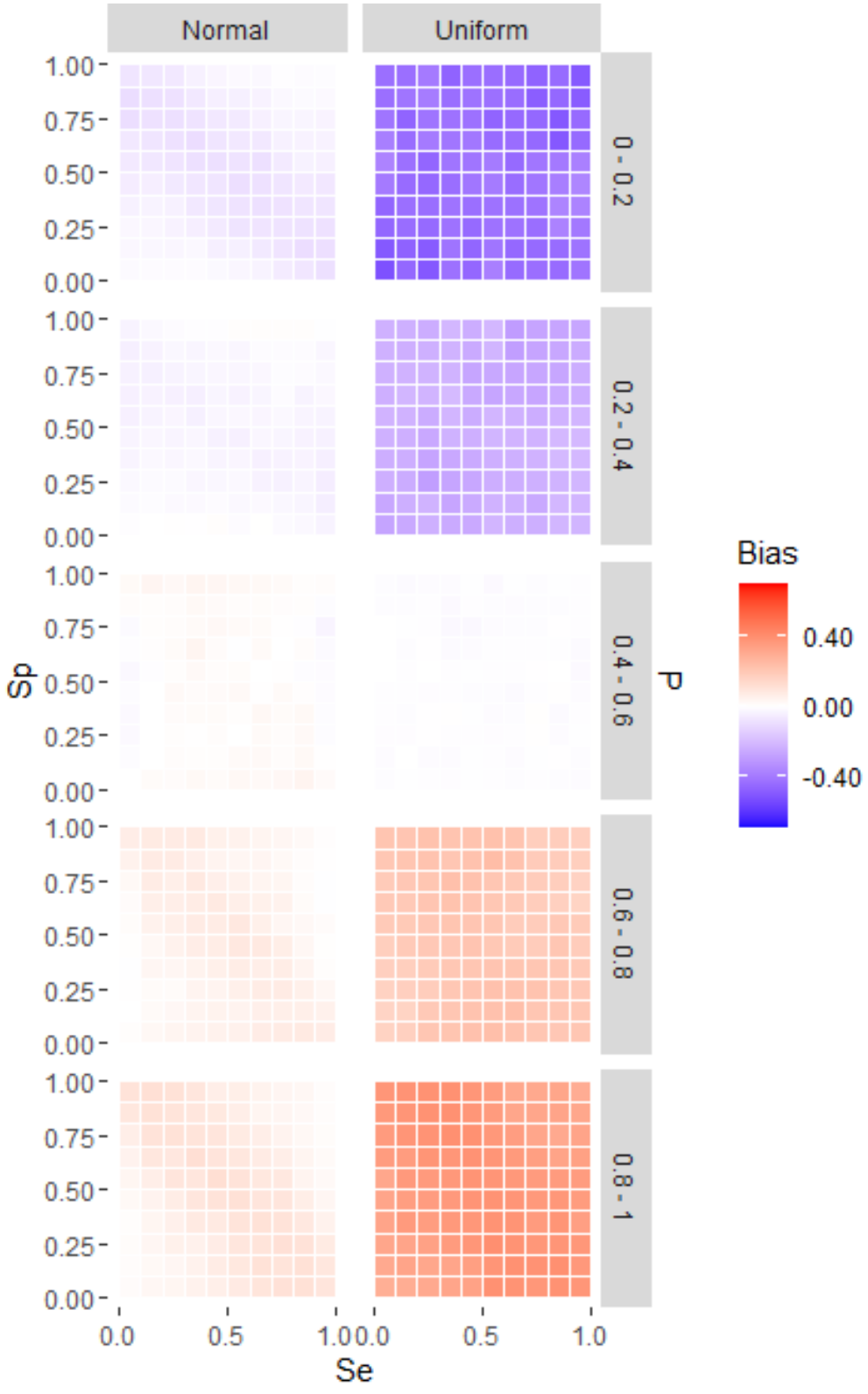

Figure 7-8: The bias of Phat across unconstrained parameter given either informative or uninformative priors. This figure relates to Experiment 2.

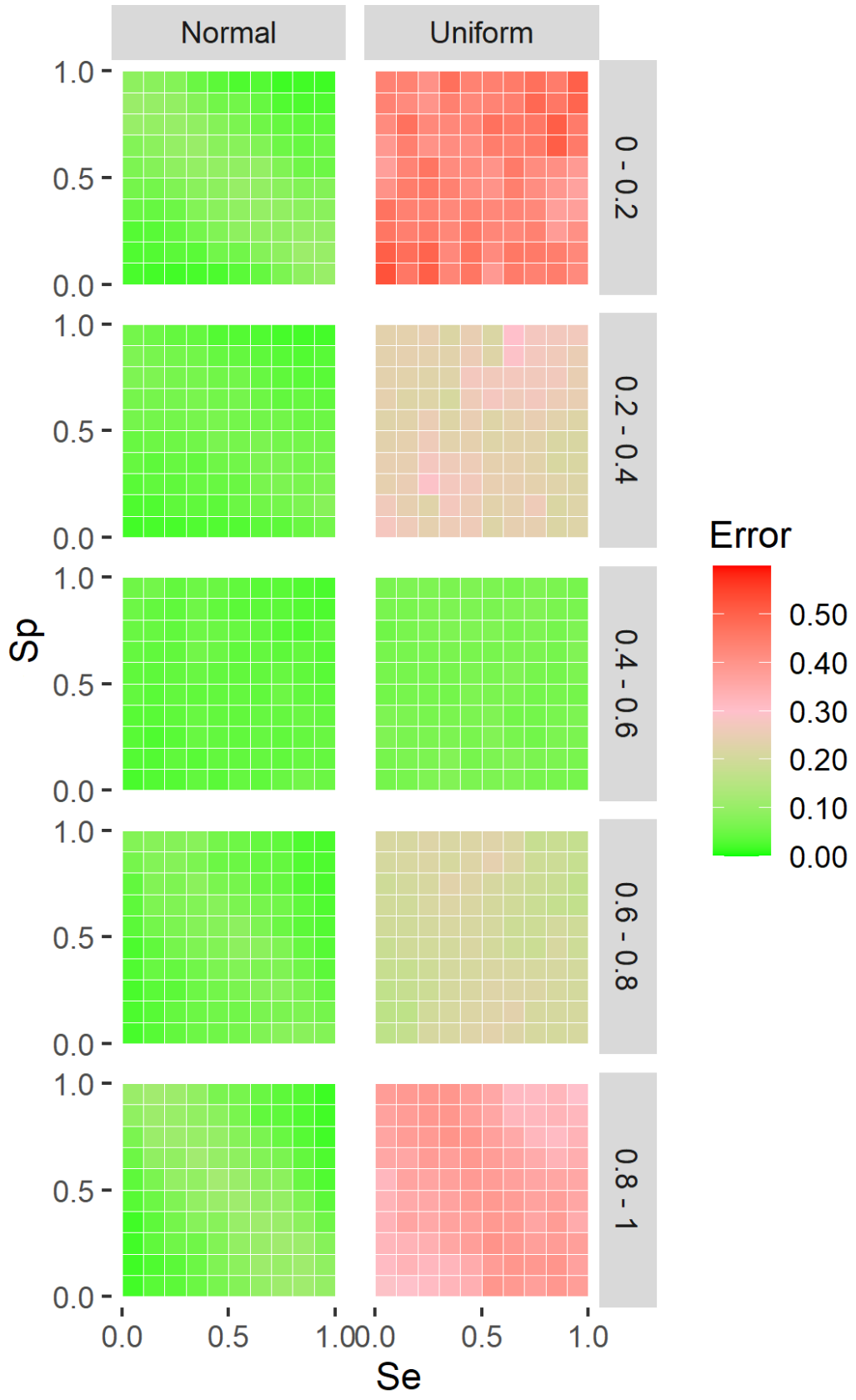

Figure 7-9: The error of Phat across unconstrained parameter given either informative or uninformative priors. This figure relates to Experiment 2.

Discussion

There is a need to conduct sensitivity analyses when using BLCMs (McAloon *et al.*, 2019, Beguin *et al.*, 2012). Indeed, the requirement to do a Global Sensitivity Analysis is not only important to the use of BLCMs, but also key to any ecological study that deals with "imperfect" data, such as those studies which sample rare species or deal with imperfect detections (Belmont *et al.*, 2022). This chapter presents a Global Sensitivity Analysis of the Any-Test Any-Population model by first varying the battery of fixed diagnostic tests (tests two to five)—in addition to Se1 and Sp1—and then by removing the assumption that "useful" parameter space would generally be within the confines of a constrained model. Since the ecological literature on BLCM sensitivity analyses are extremely limited, and furthermore largely confined to "local" rather than "global" analyses, this chapter's results are mainly discussed in the context of previous chapters within this thesis.

Experiment 1 confirms that the conclusions of Chapters 4 to 6 are generally robust against changes to the fixed diagnostic accuracies of tests 2 to 5. Consequently, this discussion focuses on the addressing the four predictions that underpin the question posed by Experiment 2: how does the error, bias, and standard deviation of Sehat, Sphat and Phat change when values of Sp can be less than 0.5, and values of P be greater than 0.5?

The experiments within this chapter did not find evidence to show that parameter space has a limited use outside of the constraints of P and Sp enforced in previous chapters, and consequently this chapter does not advise ecologists to avoid it. Although it was initially hypothesised that identifiability problems would appear in parameter space when values of Sp are less than

0.5, Experiment 2 reveals that P is likely to be unidentifiable within spaces where values of P are greater than 0.4 and values of P are less than 0.6. Figure 7-9 provides an example of this phenomenon, where BLCMs given uniform priors unexpectedly predict parameters with the same accuracy as those given normal priors when values of P are greater than 0.4 and values of P are less than 0.6.

Where is global parameter space overestimated and underestimated?

Experiment 2 reports on where global parameter space is overestimated and underestimated across four findings, and this section discusses whether these four findings can be generally applied. Overall, the four reported findings may be taken as a basic set of "rules" for interpreting bias in global parameter space.

While Chapter 6 finds that in constrained parameter space the errors of Phat are more accurately inferred than global errors, Experiment 2 presents a crucial caveat to this finding with the discovery that whether Phat is overestimated or underestimated is dependent on the value of P, meaning that global statistics could in some instances be a more robust measure of accuracy.

In addition, findings support the theory of a complex trade-off between the errors of Phat and Sphat, a finding that is further supported by stylised fact 7 in Chapter 5, which reports a heavy dependency between the errors of Phat and Sphat. In addition, Chapter 6 also supports the theory of a complex trade-off between the errors of Phat and Sphat, where in instances where the value of P is greater than 0.3 global errors are overestimated, the errors of Sehat and Sphat are unaffected. Importantly, the theorised reciprocal relationship or "Se-Sp trade-off" first reported in this thesis in Chapter 5 (stylised fact 5) is supported across global parameter space.

Are edge effects relevant in global parameter space?

Previous analyses in Chapter 6 generally show that edge effects are not present when parameter space is constrained (with an exception to this being in relation to the errors of Phat when values of P are low). Given the finding that there is a relationship between edge effects and constraint, a key aim of this chapter was to find out if edge effects persist in a completely unconstrained environment. Heatmaps show edge effects when the values of Sehat and Sphat are low and underestimated, as well as when they are high and overestimated. And the heatmaps of global errors across parameter space indicate that edge effects occur when the value of Sp is low.

This chapter also confirms that establishing the directionality of error is not simple, a finding first advanced in Chapter 6 where it was observed that the errors of Sehat and Sphat exhibit the same directionality (i.e. either overestimated or underestimated) when given either high or low values of Se and Sp respectively. In addition, this chapter confirms that the directionality of error is dependent on the value of P, a finding also first advanced in Chapter 6 where it was observed that the errors of Sehat are inaccurate when values of P are over 0.9, but also that the errors of Sehat and Sphat are likely to be underestimated when values of P are less than 0.3.

Chapter 6 cautioned against forming conclusions about extreme parameter space using global errors in isolation. Experiment 2 supports this theory, particularly given that edge effects are not clearly visible on heatmaps of global errors across parameter space.

Chapter 6 also theorises that the type of prior information that causes edge effects and advances the argument that the mean-variance relationships of the

errors of Sehat, Sphat and Phat are highly distinctive regardless of mean-variance relationships. This present study across global parameter space indicates that both these assertions are true.

The edges of parameter space are also important to consider when investigating the precision of BLCM inferences, particularly given that Chapter 5 reported that the precision of parameter inferences at edges where values of Sp are greater than 0.9 are “overly precise”, raising questions about the trustworthiness of inferences in extreme space. Experiment 2 confirms that estimates on the edge of parameter space where values of Sp are greater than 0.9 are not always “overly precise”, given the different structuring of variance between global and single parameter inferences, and the finding that the variance of error, bias and precision is structured differently across parameter space in unconstrained models.

What happens to the n.tests trend across global parameter space?

The n.tests trend was initially described in Chapter 5, stylised fact 3, which reports that increasing the number of diagnostic tests significantly reduced the errors of Phat compared to the errors of Sehat or Sphat. The simplicity of the n.tests trend means that it is powerful: it provides ecologists with a simple way of obtaining better inferences.

The heatmaps of global errors across parameter space visually highlight dependencies between the errors of Sehat, Sphat and Phat and the number of diagnostic tests available, and provides further evidence that the definition of extreme parameter space (as a 0.1 unit from the edge of parameter space) is too prescriptive, and that extreme space may also occur in regions of parameter space that are not edges. Chapter 7 provides an analysis of the interaction

between extreme space and the n.tests trend as well as on the reversal of the n.tests trend given uniform data; this chapter indicates that both trends are supported across global parameter space.

Can we trust inference when P is close to 0.5?

There is a high possibility that inferences of P around the value of 0.5 may suffer from the label switching problem, given that all three rules for avoiding this problem as reported in Chapter 3 are violated within Experiment 2 due to the lack of constraint.

The relationship between generalisability and identifiability.

Overall, given that test data as well as model constraints and priors interact via a complex function to enable identifiability (Joseph, Gyorkos and Coupal, 1995), the findings of this chapter will not generalise to all testing scenarios. For ecologists wishing to conduct a sensitivity analysis of their BLCM, there remain good reasons to work in a constrained parameter space where the information exists to make assumptions about truths. However, this chapter shows that when using batteries of diagnostic tests, the use of tests that a typical ROC analysis would consider as no better than chance alone does not automatically prevent identifiability.

Conclusion

So, are the conclusions made in this thesis on constrained parameter space generalisable across unconstrained parameter space? The Global Sensitivity Analysis conducted in this chapter found a high level of consistency between the findings of the constrained and unconstrained analyses, suggesting that the conclusions made in this thesis on constrained parameter space are indeed

generalisable across the spectrum of testing scenarios that may be faced in the wild, and were not dictated by the choice of truth. The caveat to this is that edge effects are only obvious statistical artefacts on heatmaps in constrained experiments, and so conclusions regarding edge effects may not apply to unconstrained parameter space; and the described symmetry between the errors of Sehat and Sphat indicate another type of statistical artefact that requires further investigation. Critically, the finding that parameter space is useful to ecologists outside of the constraints of P and Sp is exciting: it supports the use of diagnostic tests with a low Sp to bolster the battery of tests available to a BLCM; and suggests that the methodologies developed within this thesis are applicable to any wildlife infection scenario. The following chapter now goes on to examine whether BLCMs can infer Se, Sp and P through time, and for the first time in this thesis, applies a new tranche of time-dependent BLCMs to real-world test data.

Chapter 8

8. BLCMs can be used to infer diagnostic accuracy and prevalence through time from historic datasets.

Introduction

The theory tested in this chapter is that in the real-world, Sehat, Sphat and Phat change as a function of the period across which they are being inferred. Based on this theory, the assumption posited in Chapter 5—that diagnostic accuracy is heterogeneous across populations—is expanded to allow diagnostic accuracy to be heterogeneous across time. Given that Sehat, Sphat and Phat are known to have latent dependencies on latent variables—such as changes in demographics (McDonald *et al.*, 2016) or strain of pathogen (Strelioff *et al.*, 2013)—which change over time, this assumption clearly demands investigation.

Ecologists need to understand artefacts of time series data, and a swathe of methods to do this are conveniently at hand. Less common mechanistic methods include the use of ecological diffusion theory to forecast disease spread spatiotemporally (Hefley *et al.*, 2017), while more ubiquitous probabilistic tools for investigating time series data include naïve Bayes models (Lau *et al.*, 2017) such as Generalised Additive Mixed Models (von Brömssen *et al.*, 2018), the application of conditional heterogeneity—the assigning of statistical rules to define variance between timesteps—and state-space models that can distinguish process errors from observational errors. This useful property of state-space models is important for wildlife disease testing, as it permits the error from imperfect testing to be distinguished from the error of imperfect

trapping. Consequently, this chapter describes a new tranche of BLCMs that combine state-space theory with Bayesian latent class theory, enabling a new modelling environment where epidemiological and diagnostic parameters can vary through time.

In a time series, or an “antecedent analysis” as it is sometimes called (Bell *et al.*, 2018) a response at time t, or the mean response, is related to preceding responses. To account for a change through time, a statistical task called a “decomposition through time” (Tuncer, Tanik and Allison, 2008) can be used to manipulate longitudinal data into categorical time-dependent components—such as days, months, or years—increasing the degrees of freedom available to a BLCM (see Table 8-1). The main cost of decomposing a BLCM through time is a reduction in the number of test outcomes belonging to each element of a three-dimensional test array with dimensions as follows:

1. The diagnostic test outcomes.
2. The battery of diagnostic tests available.
3. The number of timesteps included within the sample.

Consequently, in a time-dependent BLCM, each possible testing scenario is informed by less data than is available to a time-independent BLCM.

It is logical that a latent interaction effect between years may explain any change in the Sehat and or Sphat and or Phat belonging to a diagnostic testing regime that would remain undiscovered in time-independent BLCMs. Furthermore, these latent effects may change the ability of a model to detect infection. One reason for this could be a non-trivial probability that disease statuses of individuals change between each testing point. Most BLCMs in the wildlife disease literature to date assume that each test has the same diagnostic

accuracy each time that it is used (for example Drewe *et al.*, 2010a and Buzdugan *et al.*, 2017), and this present chapter tests the impact of making this assumption.

It is already established that BLCMs can be used to detect change over time for Se, Sp and P, with compelling findings. For example, Helman *et al.*, 2020 report that estimates of P made using Bayesian Latent Class Analysis may be more robust to changes in P across cyclical epidemics than estimates made using a single test. And Patel *et al.*, 2022 find that the Sehat and Sphat of tests for Rabbit Haemorrhagic Disease viruses changed in response to changes in Phat over time. Even in the medical literature, for example, it has been discovered that time-varying values of Sehat are linked to mother-to-child HIV transmissions (Brown, 2010). Despite these findings, at present, BLCMs do not widely account for time-varying effects as the result of environmental drivers, changes in test manufacturing and or procedures, the availability of new types of diagnostic tests, or biological complexities such as a varying levels of immunity among individuals.

This chapter posits that evaluating how Se and Sp change over time is critical for maximising model power, and also for understanding why Se and Sp may change over time. As such, it shifts the focus of this thesis away from theoretical testing scenarios with known “truths” by: (a) developing and validating novel and temporally-explicit BLCMs, i.e. BLCMs capable of estimating Se, Sp and P within a time series; and (b) using these validated temporally-explicit BLCMs to infer, with credibility, the Se, Sp and P for each year of a test array formed of ten years of data collection efforts at Woodchester Park, Gloucestershire.

This chapter therefore poses and answers two key research questions:

1. Does decomposing through time enable BLCMs to improve their inferences of Se, Sp and P?
2. Can the temporally-explicit BLCM infer Se, Sp and P through time given real-world test data?

Methods

Two methodologies are now described, which address the two research questions posed. The initial “validation experiment” uses simulated test data, the findings of which are used to inform the subsequent “real-world study”, which applies a longitudinal diagnostic test array from the Woodchester Park study on bTB infected badgers.

The code for both experiments can be found on GitHub (https://github.com/annabush/PhD), with core functions printed in Table 10-4.

The overarching experimental design.

To ensure that the validation experiments could be usefully applied to the Woodchester Park data, it was important that the design of the validation experiments reflected the following three criteria:

1. The simulated data must reflect the general dimensions of the subset of the available Woodchester Park test array.
2. The BLCMs must be capable of identifying the types of trends through time that might be uncovered when decomposing the Woodchester Park test array.
3. The BLCMs must be capable of handling different types of trends through time in the same simulation.

Accordingly, for the validation and real-world studies presented in this chapter, three diagnostic tests are modelled, and all three tests are reported on—unlike the previous chapters of this thesis which assumed that the results for diagnostic test one were representative of the complete battery of tests.

Moreover, in comparison to the previous empirical chapters of this thesis, global statistics are not reported on in this chapter for two reasons. First, global statistics are mean functions and cannot be used as a proxy to infer Se, Sp and P over time since they don't depend on time (either of all timesteps or each timestep). Second, the global statistic also cannot be used as a mechanism to provide a reduction in a models' degrees of freedom requirements, which is a key purpose of a time decomposition.

The following eight assumptions guided the designs of both the validation and real-world studies:

1. P changes through time in wild populations.
2. Se, Sp and P can be inferred at distinct points through time.
3. There is a trade-off between the degrees of freedom available, the number of diagnostic tests available, and consequently the amount of data available to enable a time-dependent analysis (see Table 8-1). This is an assumption since $n - 1$, where n is the number of diagnostic outcomes (Siegel and Castellan, 1988), is not the only way to consider degrees of freedom (Bolker, 2020).
4. A single timestep within a theoretical model is representative of an annual change within a real-world model. This chapter has focussed on modelling and describing temporal patterns in Se, Sp and P across years, yet the methods presented are not restricted by how a time interval may be defined. For instance, if the diagnostic data under

investigation reflects very short periods of rapid testing, the units of change may be swapped with hours, days, or weeks.

5. For the real-world study—and as assumed in related research (McDonald and Hodgson, 2018)—the three available tests are assumed to be fully independent of each other, and able to diagnose infection at any infection stage.
6. Adding the dimension of time to parameter space requires a new validation methodology to evaluate how robust the BLCM is to different patterns of change through time.
7. The “Any-Test and Any-Population” model produces identifiable results for three-test situations over independent timesteps.
8. Model power depends on how the time effects are specified.

The detailed methodology for the validation experiment is described next, followed by the detailed methodology for the real-world study.

Table 8-1: The degrees of freedom available when up to five diagnostic tests are decomposed across up to three timesteps, under the assumption that Se and Sp can change across timesteps.

Number of tests	**Number of timesteps** $nsteps$	**Number of parameters** $nsteps(2D+1)$	**Number of test outcomes** $nsteps(2^D)$	**Degrees of freedom** $nsteps(2^D) - nsteps$	**Are the degrees of freedom ≥ parameters?**
1	1	3	2	1	N
1	2	6	4	2	N
1	3	9	6	3	N
2	1	5	4	3	N
2	2	10	8	6	N
2	3	15	12	9	N
3	1	7	8	7	~
3	2	14	16	14	Y
3	3	21	24	21	Y
4	1	9	16	15	Y
4	2	18	32	30	Y
4	3	27	48	45	Y
5	1	11	32	31	Y
5	2	22	64	62	Y
5	3	33	96	93	Y

Validating the power of the time-dependent BLCMs.

In this chapter, the diagnostic test results, and the total sample size of results available to the model, are explicitly defined across time. To do this, the general structure of the BLCMs used so far in this thesis is modified to update the model at each user-defined timestep by iterating for time around the likelihood function.

The validation experiments therefore test whether different patterns of change in Se, Sp and P can be detected over time before the model is applied to real-world test data. Regressions integral to the JAGS model are defined (Equation 22 and Equation 23), which are shown to detect these patterns of change across time.

In the validation experiments, each timestep is assumed to be one year, one population is studied, and the population is replicated. For comparison purposes, in the real-world study, each timestep represents one year of the trapping and testing cycle used to generate the data available to this study.

For the validation experiments, uniform priors were used instead of normal priors for two key reasons:

1. To isolate the noise of each model to the time effect, and the noise of that effect, only.
2. To enable a series of seven of the simplest time decompositions to be investigated—which did not include the additional possible sources of bias from the need to provide informative priors to up to seven regression coefficients (Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat, Phat; where Se1hat, for example, denotes the inferred value of Se for

diagnostic test 1 within a battery of diagnostic tests) across time—which were used to direct the choice of model for the real-world simulation.

Therefore, the only prior information given to the models were parameter constraints, in addition to the time effect integral to each time decomposition.

Each model was run using a simulated diagnostic test array with a sample size of 300 individuals, over five timesteps, and repeated across 50 simulations. These modelling conditions were chosen based on the following reasons:

1. Around 300 badgers are trapped and tested in the Woodchester Park study per year.
2. Limiting the experimental design to five timesteps across 50 simulations enabled a clear trend through time to be plotted, while avoiding the simulation of lengthy time series, which is computationally intensive.

The following subsections describe how the true values—and the applicable time decompositions—were specified, before presenting the seven “scenarios” that form this validation methodology. In general, each scenario details a different method of generating the true values for Se1, Se2, Se3, Sp1, Sp2, Sp3 and P, which becomes successively more complex. And truths from each of the seven scenarios are applied in turn to each time decomposition.

How the truths are selected

For each scenario, changes in Se1, Se2, Se3, Sp1, Sp2, Sp3 and P across time may be specified as being “constant”, “linear”, “independent”, or “mixed”, where the truths may therefore be categorised as follows:

Independent: Se, Sp and P change independently through time.

Constant: Se, Sp and P do not change across timesteps.

Linear: Se, Sp and P change linearly across timesteps.

Mixed: Se, Sp and P can each have an independent, constant, or linear change across timesteps.

It was considered important to select the true values of each timestep in different ways given that the relationship between truth and time is a latent variable in the real world. In addition, the linear and constant truths can be specified as having “noisy” relationships with time, given some random noise drawn from the gaussian distribution with a mean of 0 and a chosen standard deviation of 0.02.

How the time decompositions were specified.

Within the validation experiment, the time decomposition models were used to investigate model performance given a known trend through time. The knowledge gained from this investigation was then used to justify the choice of models used to infer Se, Sp and P through time given the Woodchester Park test data.

Three types of time decomposition were specified as regressions within the JAGS code as follows.

For the Three-Test, Five-Timestep constant model, all parameters were assumed to remain constant throughout all timesteps, such that:

Equation 22

$$\hat{y}_t = \hat{y}.$$

For the Three-Test, Five-Timestep linear model, all parameters were assumed to have a linear relationship with respect to time, such that:

Equation 23

$$\hat{y}_t = \hat{m}t + \hat{c},$$

where $\hat{m}$ (gradient) and $\hat{c}$ (intercept) are inferred by the BLCM.

For the Three-Test, Five-Timestep independent model, all parameters were assumed to vary independently of time, this equates to the BLCM structure used in previous chapters being repeated for each timestep.

The following JAGS code provides an example of the Bayesian specification of the Woodchester_linear_linear_independent model identified in results section 2. Within this model, values of P are assumed to have a linear relationship across time, values of Sp are assumed to have an independent relationship across time, and values of Se are assumed to have a linear relationship across time.

```
model {
  # Set P
  pi.m.prior ~ dunif(-(pi.limit[2] - pi.limit[1])/n.time, (pi.limit[2]
- pi.limit[1]) / n.time)
  pi.m <- pi.m.prior
  pi.c.prior ~ dunif(pi.limit[1] + max(0, - pi.m * (n.time - 1)),
pi.limit[2] - max(0, pi.m * (n.time - 1)))
  pi.c <- pi.c.prior
  for (t in 1:n.time){
    pi[t] <- pi.m * (t - 1) + pi.c
  }

  # Set Se
  for (t in 1:n.time){
     for (i in 1:n.diag){
      se.prior[i, t] ~ dunif(se.limit[1], se.limit[2])
      se[i, t] <- se.prior[i, t]
     }
  }

  # Set Sp
  for (i in 1:n.diag){
    sp.m.prior[i] ~ dunif(-(sp.limit[2] - sp.limit[1]) / (n.time - 1),
(sp.limit[2] - sp.limit[1] ) / (n.time - 1))
    sp.m[i] <- sp.m.prior[i]
    sp.c.prior[i] ~ dunif(sp.limit[1] + max(0, - sp.m[i] * (n.time -
1)), sp.limit[2] - max(0, sp.m[i] * (n.time - 1)))
    sp.c[i] <- sp.c.prior[i]
      for (t in 1:n.time){
    sp[i, t] <- sp.m[i] * (t - 1) + sp.c[i]
    }
  }

  for (t in 1:n.time){
```

```
    p[1, t] <- pi[t] * (1-se[1, t]) * (1-se[2, t]) * (1-se[3, t]) +
(1-pi[t]) * sp[1, t] * sp[2, t] * sp[3, t]
    p[2, t] <- pi[t] * (1-se[1, t]) * (1-se[2, t]) * se[3, t] + (1-
pi[t]) * sp[1, t] * sp[2, t] * (1-sp[3, t])
    p[3, t] <- pi[t] * (1-se[1, t]) * se[2, t] * (1-se[3, t]) + (1-
pi[t]) * sp[1, t] * (1-sp[2, t]) * sp[3, t]
    p[4, t] <- pi[t] * (1-se[1, t]) * se[2, t] * se[3, t] + (1-pi[t])
* sp[1, t] * (1-sp[2, t]) * (1-sp[3, t])
    p[5, t] <- pi[t] * se[1, t] * (1-se[2, t]) * (1-se[3, t]) + (1-
pi[t]) * (1-sp[1, t]) * sp[2, t] * sp[3, t]
    p[6, t] <- pi[t] * se[1, t] * (1-se[2, t]) * se[3, t] + (1-pi[t])
* (1-sp[1, t]) * sp[2, t] * (1-sp[3, t])
    p[7, t] <- pi[t] * se[1, t] * se[2, t] * (1-se[3, t]) + (1-pi[t])
* (1-sp[1, t]) * (1-sp[2, t]) * sp[3, t]
    p[8, t] <- pi[t] * se[1, t] * se[2, t] * se[3, t] + (1-pi[t]) *
(1-sp[1, t]) * (1-sp[2, t]) * (1-sp[3, t])
    y[t, 1:8] ~ dmulti(p[1:8, t], n[t])
  }
}
```

Accordingly—and separate from how the true values are selected—models were specified to fit four categories as follows.

Independent: Se1, Se2, Se3, Sp1, Sp2, Sp3 and P are independently inferred for each timestep. This model is not strictly a time decomposition.

Constant: For each timestep, the Three Test, Five Timestep constant model directly infers one value for each of Se1, Se2, Se3, Sp1, Sp2, Sp3 and P. This scenario tests what happens to the inferred values when the truth is inferred to not change across timesteps.

Linear: For each parameter Se1, Se2, Se3, Sp1, Sp2, Sp3 and P, the Three Test, Five Timestep linear model infers a gradient and intercept of a linear relationship with respect to time. This scenario tests what happens to the inferred values when the truth changes linearly with timesteps.

Mixed: A model that can infer Se1, Se2, Se3, Sp1, Sp2, Sp3 and P through time, where each parameter may be associated with a different trend through time, that can be independent, constant, or linear.

The combinations of truths and models that were investigated, and why.

Scenario 1: Se, Sp and P are randomly generated.

Scenario 1 serves as the “control” study, providing baseline inferences when there is no trend through time to detect. A time-independent Three-Test, One-Population model, as specified within Chapter 3, is used.

Scenario 2: Se, Sp and P are constant through time.

Scenario 2 is the first “time decomposition” experiment, and tests whether a constant trend across time can be detected using the Three Test, Five Timestep constant model.

Scenario 3: Se, Sp and P have a noisy constant relationship with time.

Using the Three Test, Five Timestep constant model, Scenario 3 investigates what happens to inferred values when the truth changes slightly between timesteps.

Scenario 4: Se, Sp and P have a linear relationship with time.

Scenario 4 tests whether a linear trend across time can be detected using the second time decomposition model, the Three Test, Five Timestep linear model.

Scenario 5: Se, Sp and P have a noisy linear relationship with time.

Using the Three Test, Five Timestep linear model, Scenario 5 tests what happens to inferred values when the truth changes linearly with timesteps, and the linear relationship is not perfect.

Scenario 6: Se, Sp and P each have a different relationship with time.

Using the Three Test, Five Timestep mixed model, Scenario 6 tests what happens to inferred values when the truth for each parameter may each have a

different relationship across time. The specific situation where Se and Sp parameters have a constant relationship with time, and where P has a linear relationship with time, was investigated.

Scenario 7: Se, Sp and P each have a different relationship with time, and this relationship is imperfect.

Using the Three Test, Five Timestep mixed model, Scenario 7 tests what happens to inferred values when Se and Sp have a constant and noisy relationship with time, and where P has a linear and noisy relationship with time.

Applying time-dependent BLCMs to the real-world testing scenario.

The real-world models were supplied with test outcome data from the long-term epidemiological study of bTB infected badgers at Woodchester Park, Gloucestershire, UK, where yearly trapping and test data have been recorded since 1976 (Delahay, Brown, *et al.*, 2000). The data available to this study consisted of the test results between 2006 and 2015, and throughout this period three routinely-used diagnostic tests were consistently recorded. None of these tests are a gold standard, and they can be summarised as followed:

1. The gamma interferon release assay, which uses whole-blood samples (Dalley *et al.*, 2008).
2. The BrockTB Stat-Pak test, which uses serological samples to detect bTB antibodies (Greenwald *et al.*, 2003). Note, this test has since been replaced by the Dual-Path Platform VetTB test (Arnold *et al.*, 2021).
3. The mycobacterial culture test, in which non-blood samples—such as oesophageal aspirate, tracheal aspirate, faeces, urine, and swabs from bite wounds and abscesses—are incubated to detect growths of the bTB

bacterium (Clifton-Hadley, Wilesmith and Stuart, 1993), which when completed post-mortem is a gold standard test.

The raw data, which consisted of 3807 rows, was filtered (using the `load.data` function found on https://github.com/annabush/PhD) to ensure that the data complied with the following rules:

1. For each trapping instance recorded, results were only inputted into the real-world test array if results for all three tests were available.
2. Each result corresponds to one instance of one badger being tested, which may have been repeatedly captured, trapped, and tested.

The total size of the filtered real-world dataset was 2533 rows across three diagnostic tests, inclusive of 10 timesteps.

It is understood (Hodgson, 2022) that the ecological research community with an interest in the Woodchester Park badgers has speculated that a specific change point exists within the collected test data which marks an increase in the proportion of positive test results recorded. And specifically, that this change point is associated with the BrockTB Stat-Pak test. It is also thought that the proportion of positive test results indicate nonlinear trends in P through time (McDonald *et al.*, 2016), which may be cyclical (Rogers *et al.*, 1999). Accordingly, it is hypothesised that this change point may be identified by understanding the performance of the Woodchester testing battery at each yearly interval using time decompositions.

Accordingly, the hypothesis that motivated this chapter is that the Sp of the BrockTB Stat-Pak test changes within the 2006 to 2015 period. This hypothesis is explicitly reported on in results section 2.

Note, each of the three diagnostic tests studied is allocated the same assumed trends in diagnostic accuracy through time, as dictated by the model reference. The models used to interrogate the Woodchester Park dataset investigate 27 possible combinations of parameter relationships through time as described in Table 8-2; and for each model, it was considered that any differences among the responses of tests 1 to 3 to the assigned trends through time for P, Se or Sp may be visually determined (Figure 8-15 to Figure 8-21).

How the validation experiment was analysed.

Note, the plotting code for both the validation and real-world studies can be found on GitHub (https://github.com/annabush/PhD).

To validate the time-decomposition models, three key outputs were analysed.

1. Probability density functions that indicate the variation in the inferred errors of Sehat, Sphat and Phat (see Figure 8-2, Figure 8-4, Figure 8-6, Figure 8-8, Figure 8-10, Figure 8-12, Figure 8-14 inserted in-text in the order of Scenario 1 to Scenario 7). These figures demonstrate the certainty that can be attached to the mean error values that are reported in tables (see point 3).
2. Plots (see Figure 8-1, Figure 8-3, Figure 8-5, Figure 8-7, Figure 8-9, Figure 8-11, and Figure 8-13 inserted in-text in the order of Scenario 1 to Scenario 7) showing the true and inferred values of each of Se1, Se2, Se3, Sp1, Sp2, Sp3 and P, for each timestep, facetted by the method used to select true values and the time decomposition chosen.
3. Tables (Table 8-3 to Table 8-9) showing the error of each of Se1, Se2, Se3, Sp1, Sp2, Sp3 and P across each timestep, for each method used to select true values and the time decomposition chosen.

Note, these three key outputs are included within the text of this Chapter since the visual information that they represent is crucial to the reporting of each scenario.

How the real-world experiment was analysed.

In the real-world study, to understand how inferences of Se1, Se2, Se3, Sp1, Sp2, Sp3 and P change through time, and to ascertain whether the posited change point can be detected, the following two situations were investigated.

1. Where Se, Sp and P are assumed to each have the same relationship with time across the time series, which may be constant, independent, or linear.
2. Where Se, Sp and P are each assumed to have different relationships with time across the time series, which may be independent or linear only.

Accordingly, the Woodchester Park data was subject to the assumptions and modelling conditions described in Table 8-2, and the outputs can be found from Figure 8-15 to Figure 8-21. Moreover, it is considered that the motivating hypothesis of this chapter—that the Sp of the BrockTB Stat-Pak test changes within the 2006 to 2015 period—can be tested within this modelling setup.

A note on model comparisons.

When truths are unknown, and in the absence of a “general consensus” on which Bayesian model comparison tool is appropriate (Hooten, Hobbs and Ellison, 2015), the following workflow was used to determine the credibility of inferred trends through time.

1. The results of the validation experiment were used to understand the identifiability and accuracy of time-dependent and time-independent

BLCMs given hypothetical known trends through time. Identifiability was established by visually determining which BLCMs (described in Table 8-2) can identify known trends through time. Models that successfully identified known trends with the least error were considered the “best”.

2. Informed by the findings of the validation experiment, 27 models were specified (Table 8-2) and then provided with the Woodchester Park diagnostic test data. This selection of models allowed the hypothesised linear trend in Sp across time to be investigated in relation to the BrockTB Stat-Pak test, while also accounting for the possibility of additional unknown relationships between the values of Se, Sp and P and time.
3. Visual comparisons across the posterior inferences of these 27 models (Figure 8-15 to Figure 8-21) enabled a rapid elimination of models that did not detect the presence of assumed trends. In addition, inferences from “independent” models—which served as control scenarios since they did not emulate a time series—were compared with inferences from time-dependent BLCMs to detect inconsistencies. Seven observations are reported on.
4. The posterior inferences associated with the “best” models were compared to relevant published values (Table 8-13).

Table 8-2: The time-dependent and time-independent models applied to the Woodchester Park dataset. Assumptions are applied to the Se and Sp of all three tests.

Assumption	**How each parameter is modelled.**		
	P	Se	Sp
Se, Sp and P are independent with respect to time	Independent	Independent	Independent
Se, Sp and P have a constant relationship with time	constant	constant	constant
Se, Sp and P have a linear relationship with time	linear	linear	Linear
P is independent to time, Se and Sp can be linear or independent	Independent	Independent	Independent
	Independent	Independent	Linear
	Independent	Linear	Independent
	Independent	Linear	Linear
Se is independent to time, P and Sp can be linear or independent	Independent	Independent	Independent
	Independent	Independent	Linear
	Linear	Independent	Independent
	Linear	Independent	Linear
Sp is independent to time, Se and P can be linear or independent	Independent	Independent	Independent
	Independent	Linear	Independent
	Linear	Independent	Independent
	Linear	Linear	Independent
P has a linear relationship with time, Se and Sp can be linear or independent	Linear	Independent	Independent
	Linear	Independent	Linear
	Linear	Linear	Independent

	Linear	Linear	Linear
Se has a linear relationship with time, P and Sp can be linear or independent	Independent	Linear	Independent
	Independent	Linear	Linear
	Linear	Linear	Independent
	Linear	Linear	Linear
Sp has a linear relationship with time, Se and P can be linear or independent	Independent	Independent	Linear
	Independent	Linear	Linear
	Linear	Independent	Linear
	Linear	Linear	Linear

Results

The results of this present chapter are split into two main sections. In section 1, the results of the validation tests are presented. In Section 2, validated models from analyses in section 1 are applied to real-world test data, and new findings relating to the battery of diagnostic tests belonging to the long-term Woodchester Park study are described.

Section 1: The validation of the time-dependent BLCMs across seven progressively complex modelling scenarios

This section focuses on reporting the magnitudes of the errors of Sehat, Sphat and Phat given the test arrays and time decomposition models as described. Accordingly, a key aim of this validation exercise was to ascertain which time-dependent BLCM(s) should be used to simulate the historic values of Se, Sp and P from the Woodchester Park diagnostic test data.

Notes on interpreting the tables and plots of Section 1.

1. Within each table of Section 1 (Table 8-3 to Table 8-9), combinations of “truth” and “model” that produce the least error are emboldened. These combinations are referred to in the format of truth_model, where truth can be “independent”, “constant” and “linear” and model can be “independent”, “constant” and “linear” in accordance with the definitions provided.
2. The probability densities referred to are used to visually demonstrate the variation belonging to the mean error values reported.
3. The `ggplots`, Figure 8-1 to Figure 8-13, each show two empty plots as a consequence of faceting by parameters Se, Sp and P.

4. The purpose of the panel plots is to visually demonstrate which combinations of model and truth (as defined) are identifiable. To help the viewer pick out patterns by eye across the five simulated timesteps, the trend lines have been plotted using the function `geom_smooth` of the `ggplot2` package (Wickham, 2014).
5. Figure captions provide detailed interpretations of the panel plots.

Scenario 1: Se, Sp, and P are randomly generated.

Scenario 1 demonstrates that when there is no trend through time to detect—and Se, Sp and P are inferred for each timestep independently of time—the constant and linear models do not outperform the independent model. Scenario 1 establishes the posterior densities of Se, Sp and P that are attainable using the most data-limited model, the independent_independent model (Figure 8-2). For this model, the constraint of truth is the only source of prior information, and in comparison to Se, values of Sp and P are inferred with the least error.

Table 8-3: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 1.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
independent_ independent	**0.096**	**0.152**	**0.142**	**0.145**	**0.042**	**0.048**	**0.042**	**0.095**
independent_ constant	0.133	0.251	0.245	0.235	0.124	0.111	0.119	0.174
independent_ linear	0.144	0.252	0.220	0.238	0.107	0.101	0.105	0.167

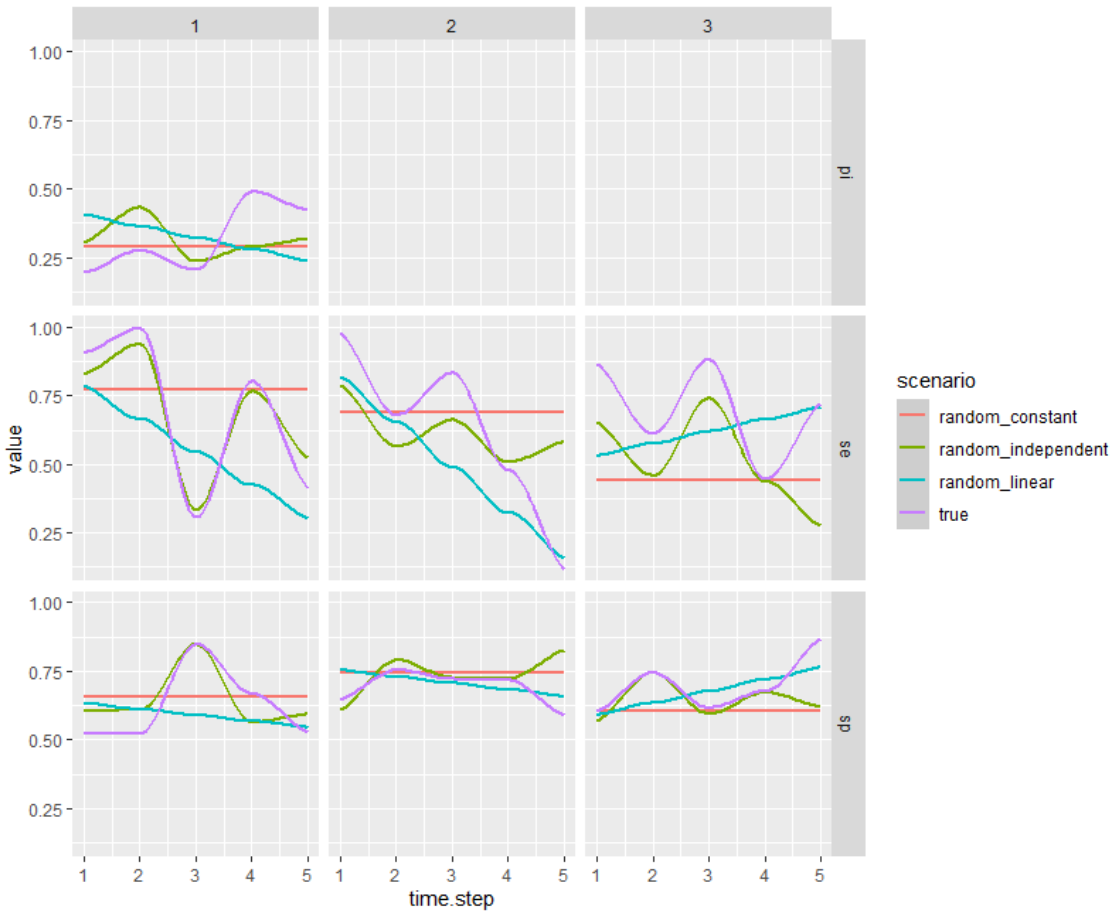

Figure 8-1: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 1, for tests 1, 2 and 3. This panel demonstrates that when true values are randomly selected, and there is no clear trend through time to detect, the constant and linear models do not correctly infer the truth; this indicates that the time decomposition models (red and blue lines) are performing as expected.

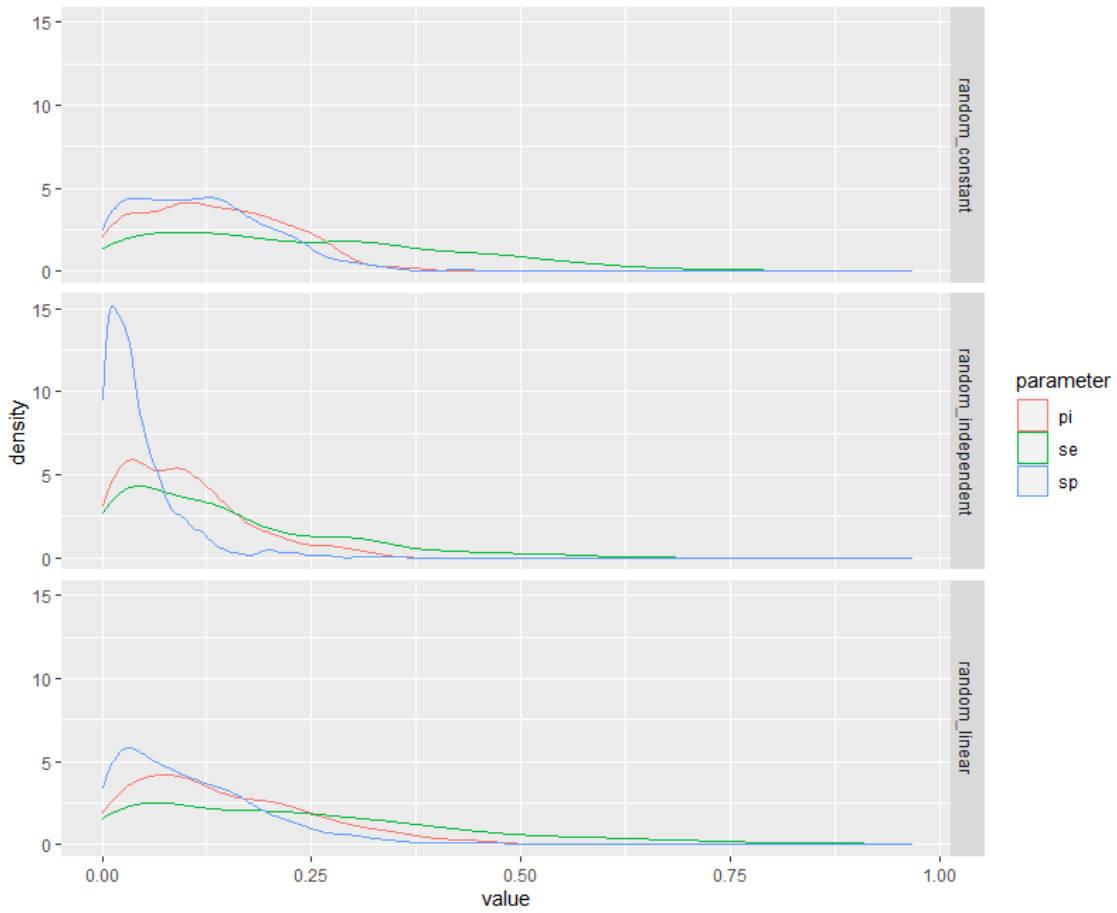

Figure 8-2: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 1. This panel shows that in comparison to the constant and linear models, the accuracy of inferences from the independent models can be associated with the most precision.

Scenario 2: Se, Sp, and P are constant across time.

When the truths are constant, the Three Test, Five Timestep constant model consistently infers each parameter with the least error in comparison to the linear or independent models, indicating that the constant trend was detected. Compared to Scenario 1, this significant reduction in error when using the most basic time decomposition model indicates that the decomposition is successfully improving inferences. Interestingly, the probability densities that inform scenario 2 (Figure 8-4) show that Sp is most accurately inferred across all models.

Table 8-4: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 2.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
constant_ linear	0.095	0.118	0.134	0.146	0.035	0.034	0.039	0.086
constant_ constant	**0.089**	**0.114**	**0.116**	**0.135**	**0.027**	**0.029**	**0.035**	**0.078**
constant_ independent	0.101	0.130	0.150	0.158	0.042	0.038	0.044	0.095

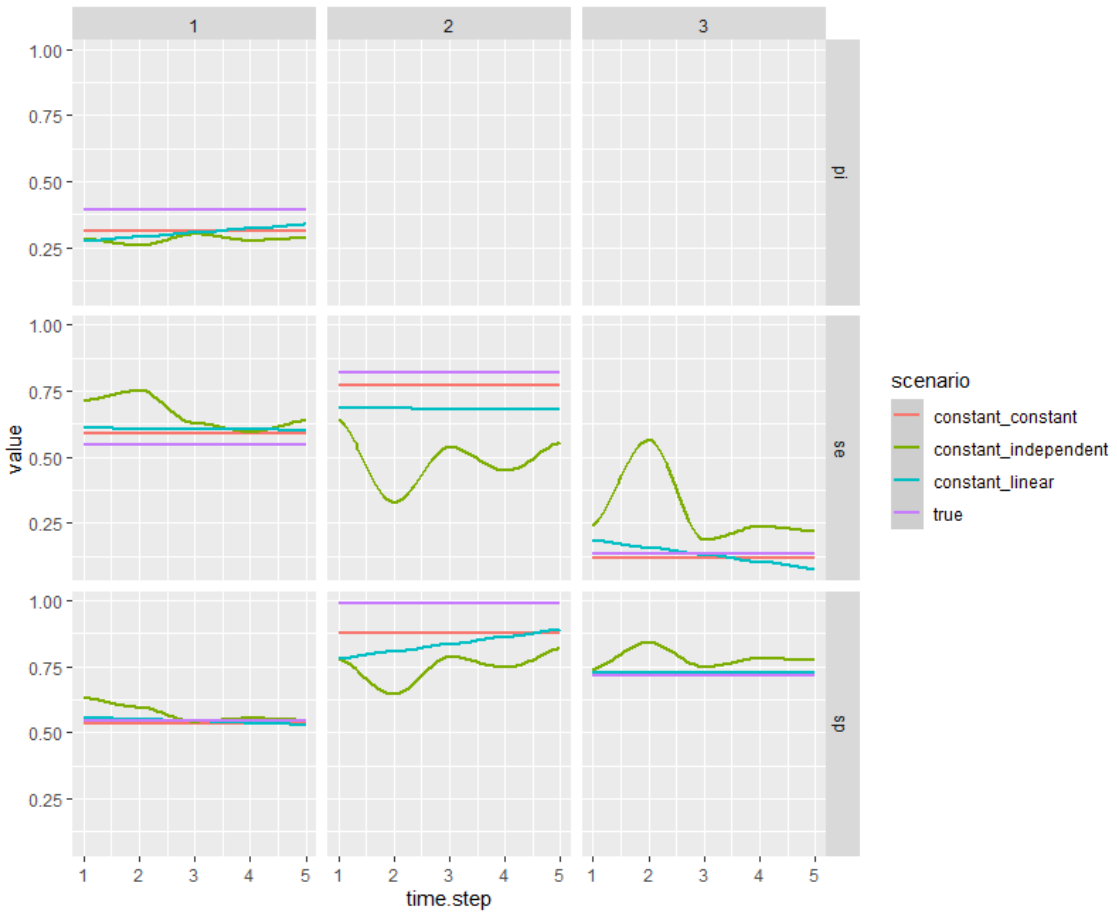

Figure 8-3: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 2, for tests 1, 2 and 3. This panel demonstrates that when there is a known constant trend through time to detect, the constant model is able to detect this trend with more accuracy than the linear or independent models.

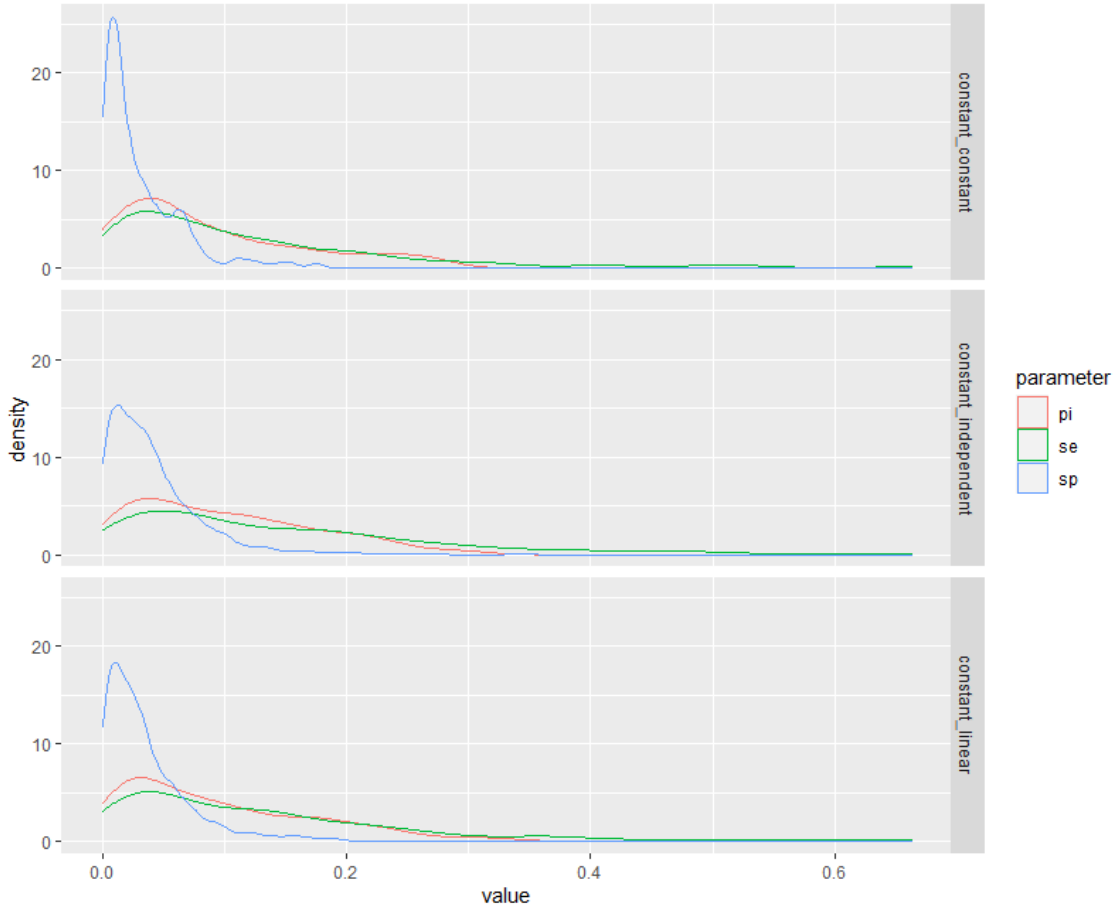

Figure 8-4: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 2. This panel shows that the regions of highest posterior density for the constant and linear models have become more obvious in comparison to Scenario 1, supporting the finding that when there is a trend through time to detect, time decomposition improves inferences.

Scenario 3: Se, Sp, and P are imperfectly constant across time.

Scenario 3 demonstrates that even if the assumption of a constant relationship through time is not precisely true, i.e. noisy, the constant model is likely to offer inferences with the least error in comparison to the independent model. Interestingly, and in contrast to Scenario 2, the probability density functions of the inferred errors are similar between models, and for Phat and Sehat show bimodality (Figure 8-6). This indicates that there are only small differences between inferences from the constant, linear and independent models, and that in comparison to Sphat, all models require more information to infer Sehat and Phat more accurately in Scenario 3.

Table 8-5: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 3.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
constant-noisy002_independent	0.093	0.149	0.169	0.153	0.053	0.046	0.043	0.101
constant-noisy002_constant	**0.080**	**0.118**	**0.136**	**0.133**	**0.042**	**0.038**	0.035	**0.083**
constant-noisy002_linear	0.085	0.127	0.140	0.140	0.043	**0.038**	**0.034**	0.087

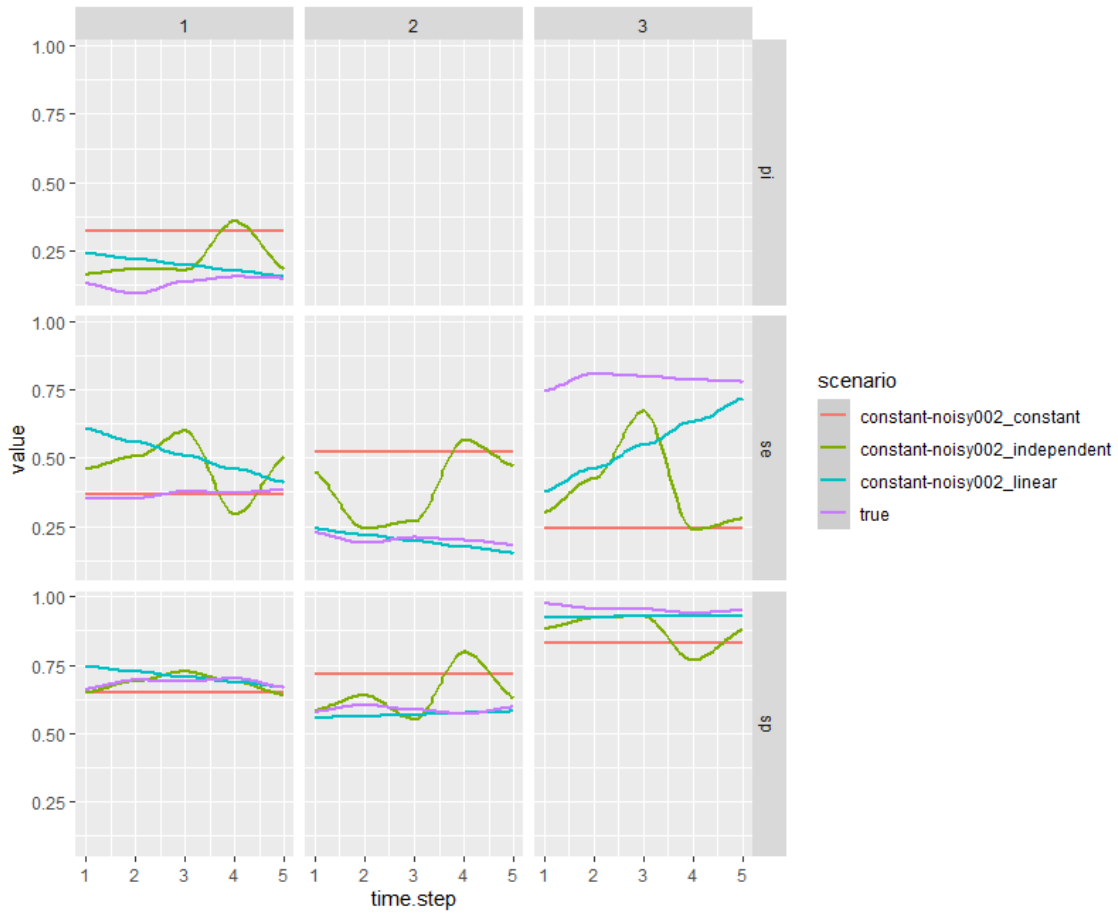

Figure 8-5: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 3, for tests 1, 2 and 3. This panel visualises the inference of parameters with a constant but noisy relationship through time with the constant, independent, and linear model. Moreover, while Table 8-5 confirms that the constant model will, on average, offer inferences with the least error in comparison to independent or linear models, this trend is not visually obvious, and there are little differences in the distributions of errors between all three models (Figure 8-6).

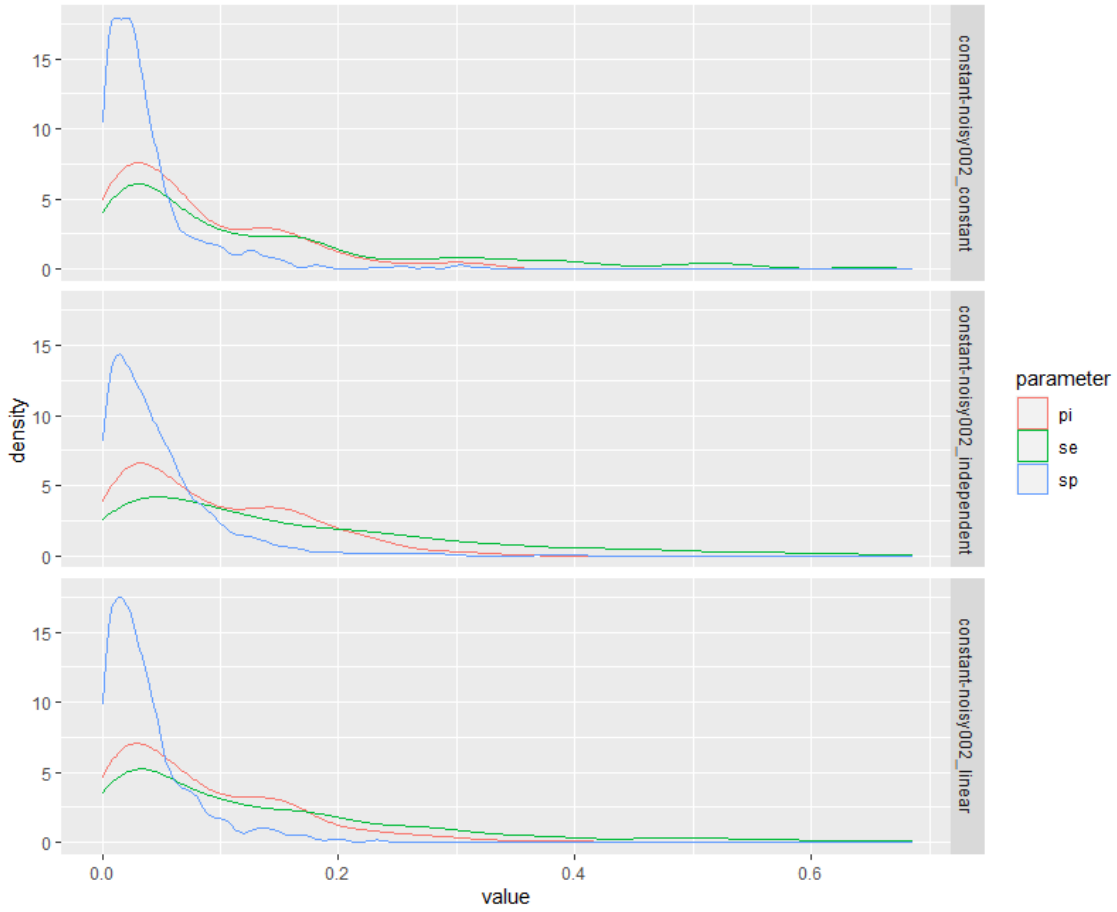

Figure 8-6: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 3. This panel shows that—similar to Figure 8-2 and Figure 8-4, and for all models—the errors of Sphat can be associated with the most certainty, and that Phat and Sehat are often associated with two regions of higher posterior density. However, for Scenario 3, the distribution of the errors of Phat, Sehat and Sphat are similar between the constant, linear and independent models, indicating that all models infer the constant but noisy relationship through time with similar precisions.

Scenario 4: Se, Sp, and P have a linear relationship across time.

In Scenario 4 where truth has a linear relationship across time, the linear model infers the values of Se, Sp and P with the least error. However, as with Scenario 3, the errors of Phat—and Sehat given the constant model—is bimodal Figure 8-8).

Table 8-6: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 4.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
linear_ independent	0.084	0.122	0.152	0.147	0.034	0.042	0.038	0.089
linear_ constant	0.083	0.117	**0.125**	0.132	0.045	0.048	0.050	0.086
linear_ linear	**0.076**	**0.095**	0.127	**0.125**	**0.024**	**0.033**	**0.030**	**0.073**

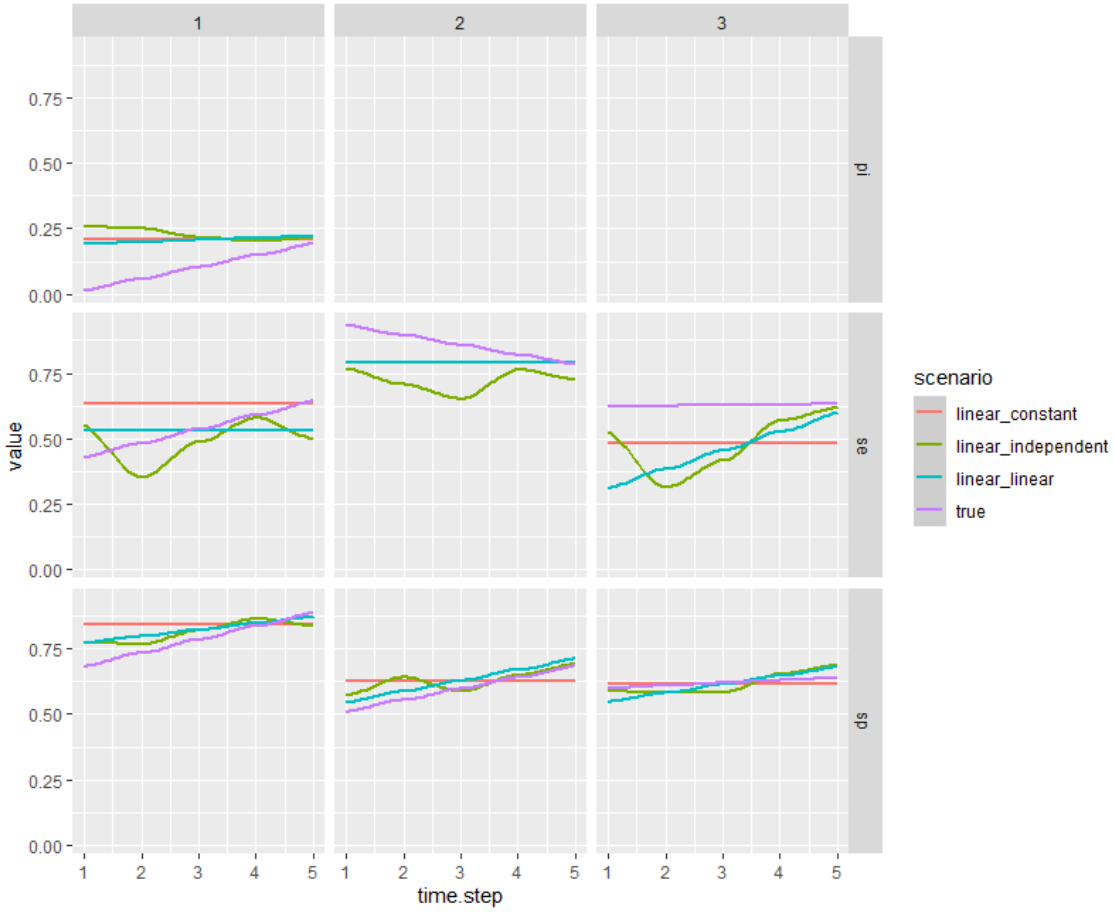

Figure 8-7: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 4, for tests 1, 2 and 3. This panel visualises the inference of parameters with a linear relationship through time with the constant, independent, and linear model. In this instance the linear model (blue lines) appears to infer the trends with the most accuracy.

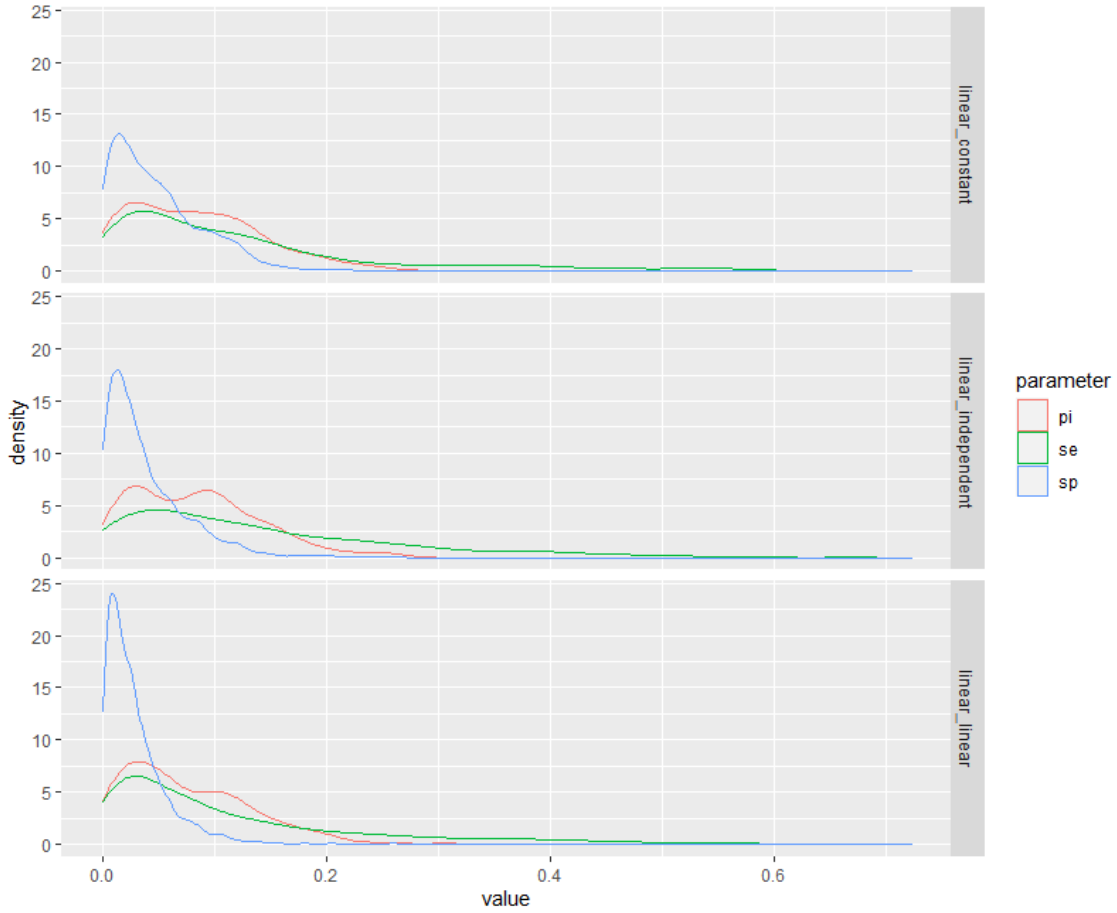

Figure 8-8: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 4. This panel shows that when there is a linear trend to detect, the linear model can infer the errors of Phat, Sehat and Sehat with the most certainty.

Scenario 5: Se, Sp, and P have a noisy linear relationship with time.

When the truth has a noisy linear relationship across time, the linear model infers Se, Sp and P with the least error. In addition, both time decomposition models offered more accurate inferences than when the independent model was used, indicating that time decompositions add model power. Despite this, for some parameters (Table 8-7) the constant model appears to offer the most accurate inferences, and in contrast to the previous scenarios, includes unimodal probability densities of the errors of Phat; it is unclear whether this inference is trustworthy (Figure 8-10).

Table 8-7 The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 5.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
linear-noisy002_independent	0.086	0.124	0.163	0.158	0.040	0.047	0.048	0.095
linear-noisy002_constant	**0.081**	0.111	**0.151**	**0.124**	0.050	0.061	0.056	0.090
linear-noisy002_linear	0.083	**0.106**	0.152	0.129	**0.034**	**0.037**	**0.043**	**0.083**

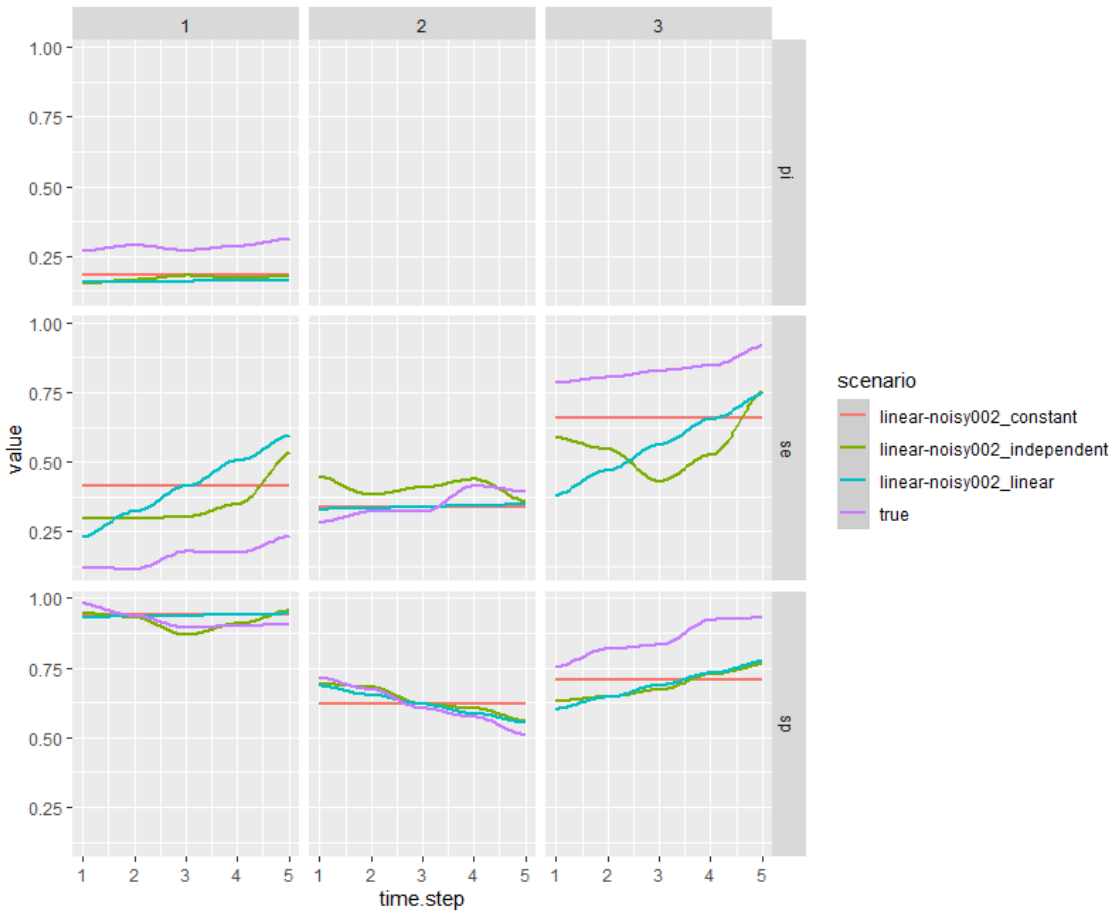

Figure 8-9: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 5, for tests 1, 2 and 3. This panel visualises the inference of parameters with a noisy linear relationship through time with the constant, independent, and linear model. In this instance the linear model (blue lines) appears to infer the noisy and linear trends with the most accuracy.

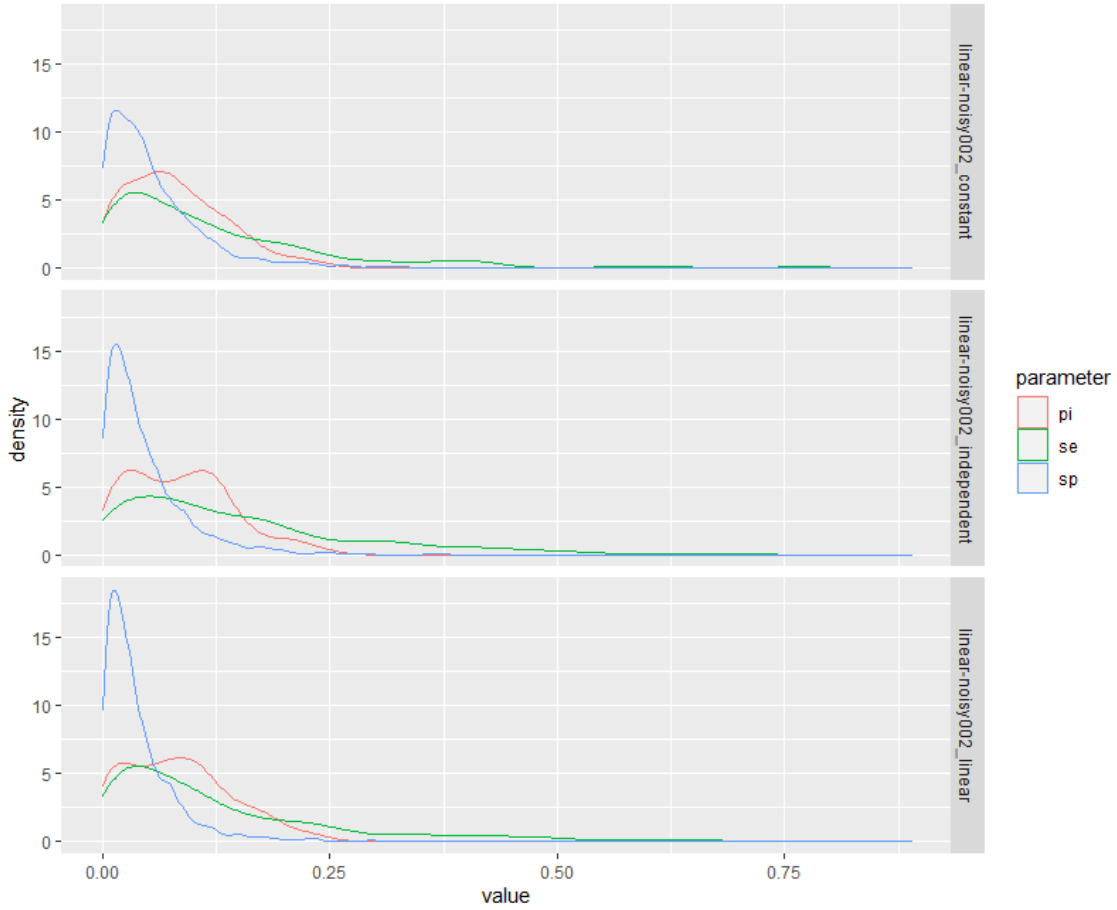

Figure 8-10 Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 5. This panel shows that for the constant model, the posterior densities of the errors of Phat are unimodal in comparison to when inferred using the independent or linear models. This panel also shows that the errors of Sphat are more certain when inferred using the linear model.

Scenario 6: Se, Sp, and P can each have a different relationship with time.

Note, in Scenario 6, Se and Sp have a constant relationship with time, and P has a linear relationship with time. Under this scenario, the average performances of each time decomposition become less varied compared to the previously reported scenarios, and the mixed model produces the most accurate inferences. In addition, the probability densities of the absolute errors of each parameter (Figure 8-12) show that Sp remains the most accurately inferred compared to P or Se, and that the bimodality issues of Phat reported in the previous Scenarios 1 to 5 are reduced.

Table 8-8: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 6.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
mixed_ independent	0.086	0.158	0.165	0.174	0.048	0.049	0.048	0.104
mixed_ constant	0.100	0.144	0.162	0.153	0.044	0.050	0.044	0.100
mixed_ linear	0.100	0.161	0.179	0.176	0.049	0.057	0.049	0.110
mixed_ mixed	**0.081**	**0.130**	**0.148**	**0.148**	**0.043**	**0.044**	**0.043**	**0.091**

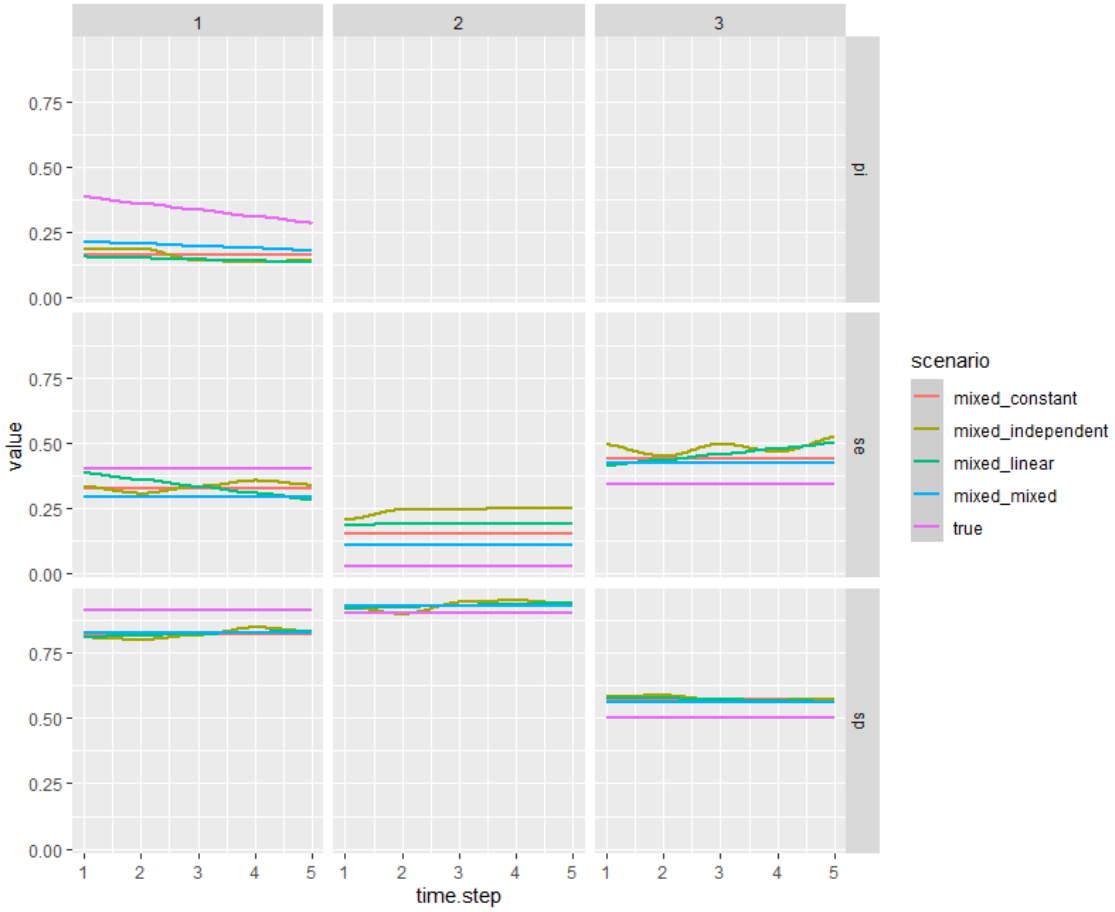

Figure 8-11: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 6, for tests 1, 2 and 3. This panel shows that when Se and Sp have a constant relationship with time, and P has a linear relationship with time, the mixed model, in this instance, identifies the linear trend in P with the most accuracy. Table 8-8 confirms that the mixed model in fact identifies every parameter with the most accuracy in this scenario.

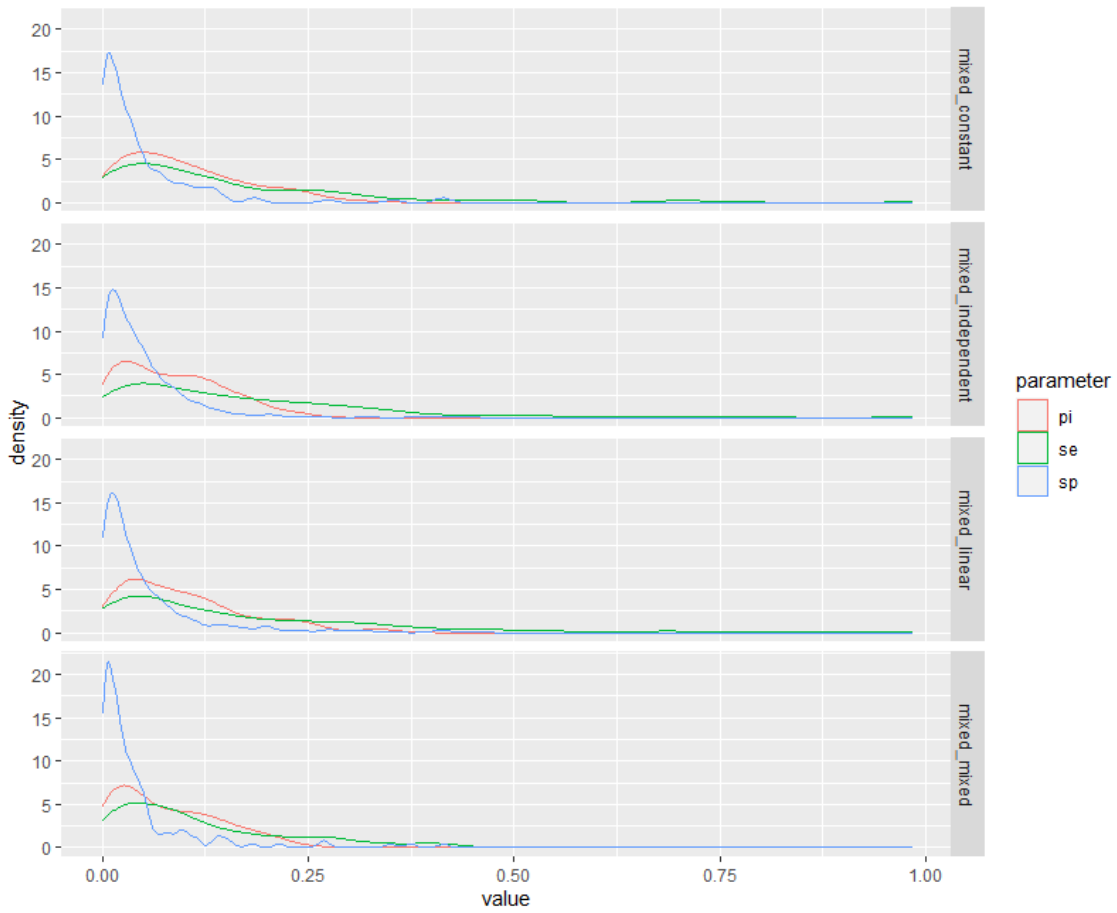

Figure 8-12: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 6. This panel suggests that the highest regions of posterior density for each parameter Phat, Sehat and Sphat are associated with the mixed model.

Scenario 7: Se, Sp, and P can each have a different and noisy relationship with time.

Note, in Scenario 7 Se and Sp have a constant noisy relationship with time, and P has a linear noisy relationship with time.

In real-world testing scenarios, Se, Sp and P may not have the same relationships with time, and these relationships might be noisy. On average, the mixed_mixed model offered the greatest accuracies, including for the inference of P, suggesting that the mixed_mixed model is the most powerful given unknown trends through time. However, as in Scenario 6, there was little difference between the power of all the time decomposition models (Figure 8-14), indicating that with more random truths, the time effect specified within the model becomes less important; and that any time decomposition model is an advantage to an independent model.

Table 8-9: The average errors across time of Se1hat, Se2hat, Se3hat, Sp1hat, Sp2hat, Sp3hat and Phat given Scenario 7.

truth_model	**Mean error across time**							
	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat	Mean
mixed-noisy002_independent	0.079	0.151	0.146	0.142	0.037	0.048	0.043	0.092
mixed-noisy002_constant	0.082	**0.113**	0.126	0.117	**0.031**	0.044	0.041	0.079
mixed-noisy002_linear	0.078	0.134	0.125	**0.108**	**0.031**	**0.039**	**0.035**	0.079
mixed-noisy002_mixed	**0.074**	0.114	**0.118**	0.114	**0.031**	**0.039**	0.039	**0.076**

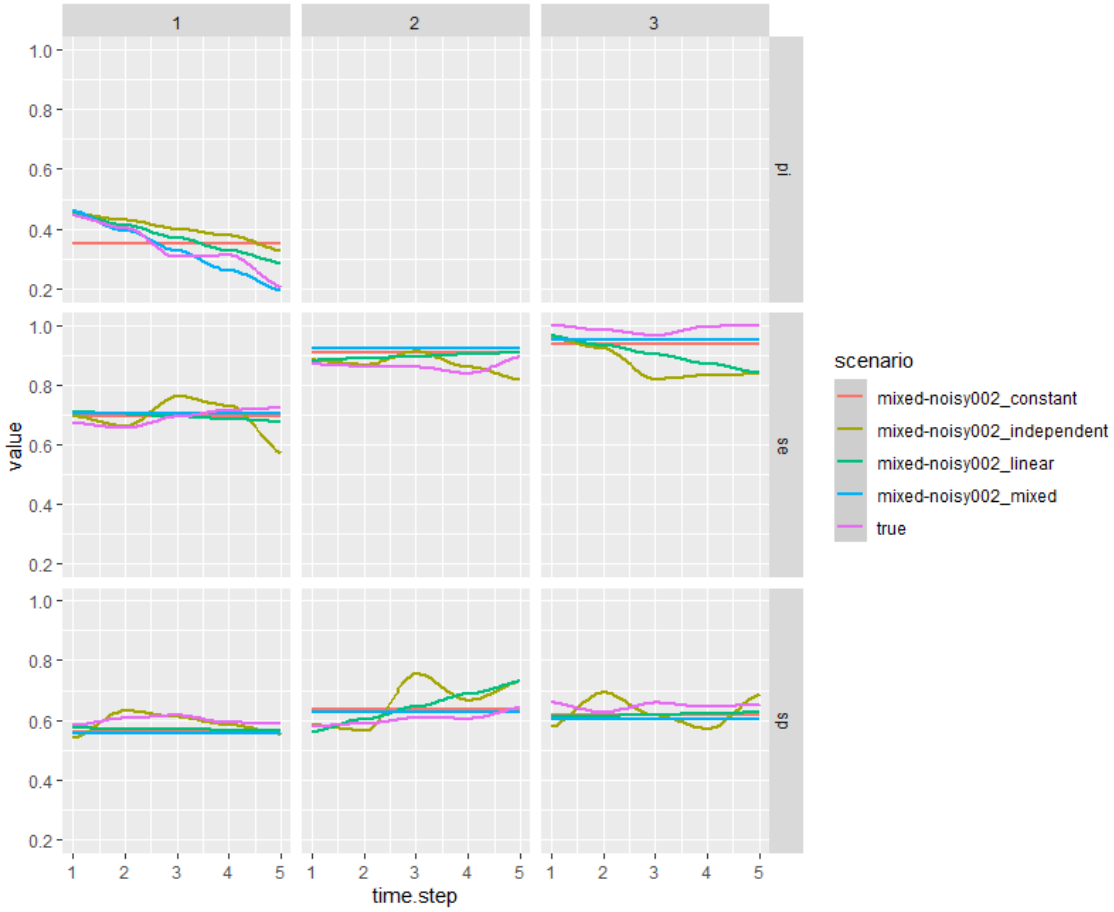

Figure 8-13: True (purple lines) and inferred values (all other lines) for each parameter and each timestep in Scenario 7, for tests 1, 2 and 3. This panel shows that when Se and Sp have a constant and noisy relationship with time, and P has a linear and noisy relationship with time, the mixed model, in this instance, generally identifies each parameter with the most accuracy. This panel also shows that the independent, linear and mixed models all identified the linear trend in P across the five timesteps that were modelled.

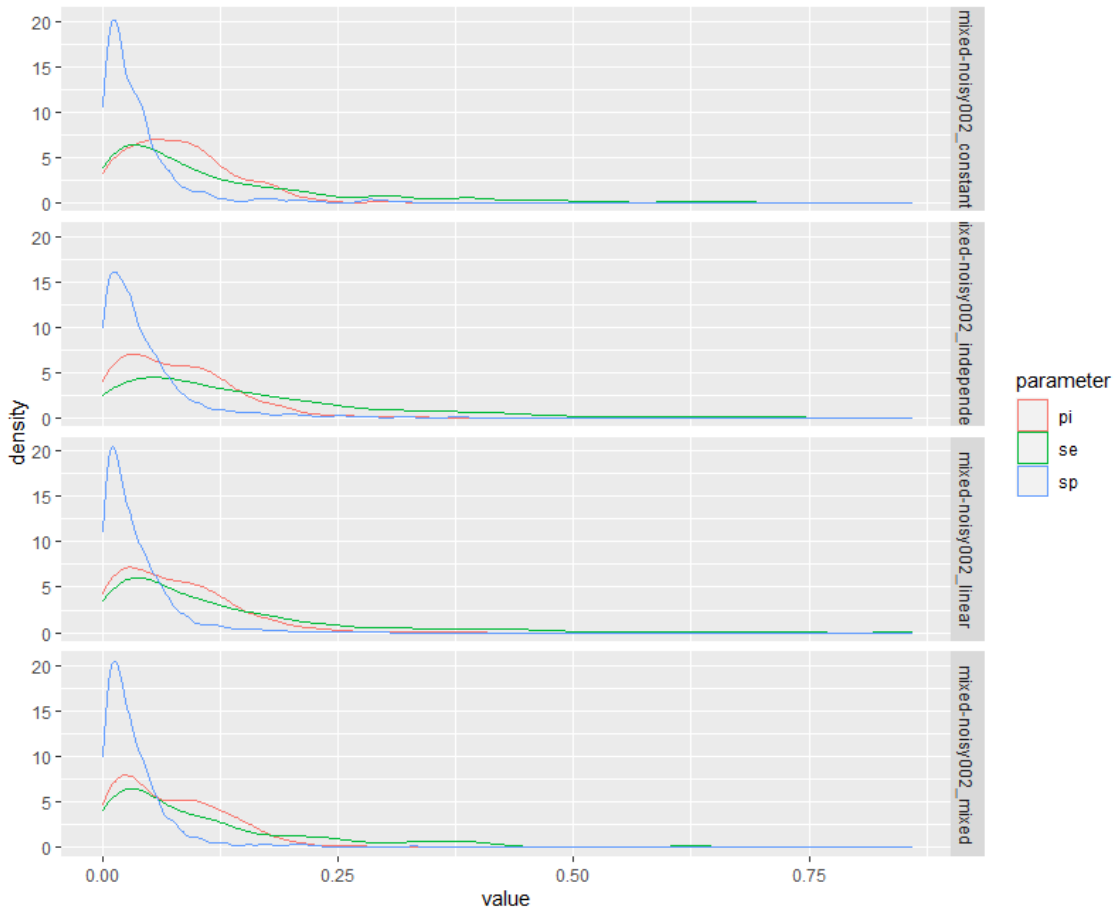

Figure 8-14: Probability densities of the errors of Phat, Sehat and Sphat for Chapter 8, Scenario 7. This panel shows that—in contrast to Scenario 6, where no noise is present—when Se and Sp have a constant and noisy relationship with time, and P has a linear and noisy relationship with time, the highest regions of posterior density for each parameter Phat, Sehat and Sphat, given any model, are more difficult to visually discern.

Section 2: How the validated BLCM infers historic values of Se, Sp and P using the Woodchester Park test array.

This section focuses on reporting observations associated with new time-dependent inferences of Se, Sp and P at Woodchester Park between 2006 and 2015 using the novel BLCM constructs which are validated in section 1.

Notes on interpreting the tables and plots of Section 2.

The truth_model format used to reference the models used in section 1 is expanded to the format woodchester_Pmodel_Semodel_Spmodel. Accordingly, while the format woodchester_independent indicates that Se, Sp and P are modelled independently of time, the format woodchester_independent_linear_independent indicates that P is modelled independently of time, Se is modelled as a linear relationship with time, and Sp is modelled independently of time.

The following list relates to the presentation of Figure 8-15 to Figure 8-21:

1. The inferred values of Se, Sp and P are plotted directly with no trend line applied.
2. The numeric facets on the x-axis of each plot are abbreviations for the following diagnostic tests: 1 = BrockTB Stat-Pak test; 2 = gamma interferon release assay; 3 = mycobacterial culture test.
3. On each x-axis, timestep 1 relates to the year 2006 and timestep 10 relates to the year 2015.
4. The y-axis shows the inferred values.
5. Figure captions provide detailed interpretations of the panel plots.

When Se, Sp and P are assumed to each have the same trend through time.

The constant model is unable to identify trends through time (Figure 8-15). However, both independent and linear trends are identified through time for all parameters (for raw results see Table 8-10). Based on this result, the constant model is omitted from the further analyses of parameter-specific changes through time.

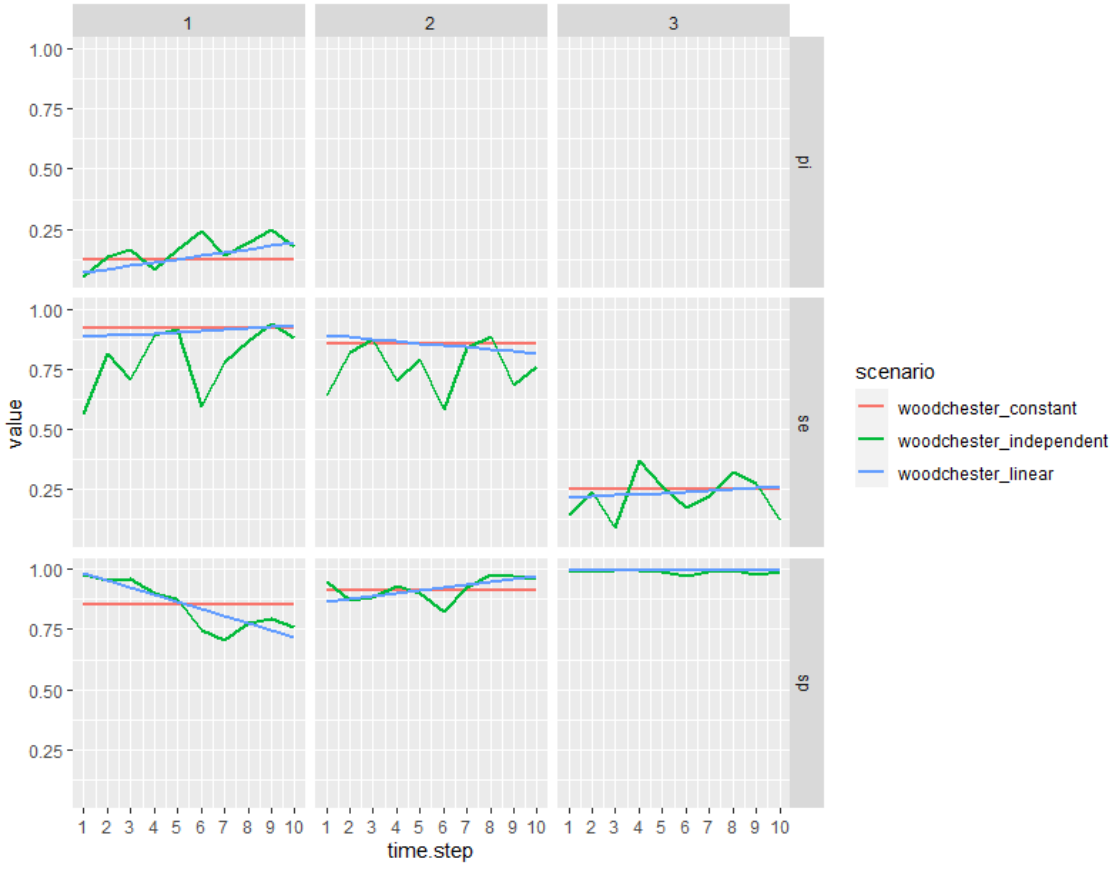

Figure 8-15: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests given the constant, independent and linear models.

Table 8-10: Raw inferred values for P, Se1, Se2, Se3, Sp1, Sp2, Sp3 at each timestep outputted from the Woodchester_independent model. In this model Se, Sp and P are modelled independently of time.

Timestep	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat
1	0.05	0.57	0.64	0.15	0.98	0.95	0.99
2	0.13	0.81	0.82	0.24	0.95	0.87	0.99
3	0.17	0.71	0.87	0.09	0.96	0.88	1.00
4	0.08	0.89	0.70	0.37	0.90	0.93	1.00
5	0.17	0.91	0.79	0.27	0.88	0.90	0.99
6	0.24	0.59	0.59	0.18	0.75	0.82	0.97
7	0.14	0.78	0.84	0.22	0.71	0.92	0.99
8	0.20	0.87	0.88	0.32	0.77	0.98	0.99
9	0.25	0.94	0.68	0.28	0.79	0.97	0.98
10	0.18	0.88	0.76	0.12	0.76	0.96	0.99

When Se, Sp and P are assumed to each have different trends through time.

The following observations were made using Figure 8-16 to Figure 8-21 listed below. Note that observations referring to “Se” and “Sp” relate to all three diagnostic tests relevant to this study.

1. Se does not have a linear relationship through time, and therefore should be modelled independently of time. When Se is assumed to have a linear trend through time, P is likely to be overestimated, and Se and Sp are likely to be overestimated.
2. On average, Sp and P have linear relationships through time, which is negative for Sp and positive for P. Further, the assumption as to whether Sp has an independent or linear relationship across time matters least when P is linear, and Se is independent.
3. Of the models investigated, the Woodchester_linear_independent_linear model is likely to be the most reliable, given that linear trends have been detected for P and Sp and non-linear trends have been identified for Se.
4. Over the decade studied, P notably and steadily increases in the Woodchester Park badger population. Using the Woodchester_linear_independent_linear model, P is observed to have increased by 10%, at around 1 percentage point per year on average (see Table 8-11 and Table 8-12).
5. The inferences of Se and Sp from linear and independent models are the least variable when a linear model is used to infer P.
6. All three tests detect a non-linear trend through time for Se when Se is modelled independently to time. All models predicted that the battery of diagnostics had the lowest values of Se in 2006 (Table 8-10, Table 8-11,

Table 8-12). In addition, the Woodchester_linear_independent_linear model suggests significant variability in the Se of all diagnostic tests.

7. Notwithstanding the above findings, the Woodchester_linear_independent_linear model does not enable the identification of any change points for Sp and P. To allow potential change points to be identified, the Woodchester_linear_independent_independent model was used to make three key observations (for raw values see Table 8-12).
    a. The Sp of the BrockTB Stat-Pak test decreased from 0.98 to 0.71 between 2006 and 2012, at which point Sp then increased.
    b. The Sp of gamma interferon release assay decreased to below 0.9 in 2007, 2008 and 2010.
    c. It is possible that P increased the most—up to 2%—between 2009 and 2010.

Does the Sp of the BrockTB Stat-Pak test change within the 2006 to 2015 period?

Yes. Linear trends in the value of Sp through time have been identified as belonging to the BrockTB Stat-Pak test. This inference was made using the Woodchester_linear_independent_linear model, which indicated that Sp generally decreased through time across the 2006 to 2015 period. Specific yearly change points were then identified using the Woodchester_linear_independent_independent model—in which Sp is inferred independently of time—and this inference suggests that the Sp of the BrockTB Stat-Pak test decreased from 0.98 to 0.71 between 2006 and 2012, at which point values of Sp then increased.

Table 8-11: Inferred values for P, Se1, Se2, Se3, Sp1, Sp2, Sp3 at each timestep outputted from the Woodchester_linear_independent_linear model. In this model P is modelled as a linear relationship with time, Se is modelled as an independent relationship with time, and Sp is modelled as a linear relationship with time.

Timestep	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat
1	0.10	0.50	0.40	0.09	0.99	0.88	1.00
2	0.11	0.83	0.84	0.26	0.96	0.89	1.00
3	0.12	0.71	0.95	0.10	0.93	0.90	1.00
4	0.13	0.87	0.59	0.30	0.90	0.91	1.00
5	0.14	0.91	0.84	0.28	0.87	0.92	1.00
6	0.16	0.81	0.86	0.23	0.84	0.93	1.00
7	0.17	0.81	0.71	0.16	0.81	0.94	1.00
8	0.18	0.89	0.89	0.34	0.78	0.96	0.99
9	0.19	0.95	0.75	0.33	0.75	0.97	0.99
10	0.20	0.83	0.78	0.11	0.72	0.98	0.99

Table 8-12: Inferred values for P, Se1, Se2, Se3, Sp1, Sp2, Sp3 at each timestep outputted from the Woodchester_linear_independent_independent model. In this model P is modelled as a linear relationship with time, Se is modelled as an independent relationship with time, and Sp is modelled as an independent relationship with time.

Timestep	Phat	Se1hat	Se2hat	Se3hat	Sp1hat	Sp2hat	Sp3hat
1	0.09	0.45	0.61	0.09	0.98	0.96	0.99
2	0.10	0.87	0.87	0.28	0.94	0.86	0.99
3	0.11	0.84	0.91	0.11	0.95	0.85	1.00
4	0.12	0.86	0.58	0.27	0.92	0.94	1.00
5	0.14	0.93	0.84	0.29	0.86	0.89	0.99
6	0.15	0.84	0.84	0.23	0.85	0.92	0.99
7	0.16	0.74	0.83	0.18	0.71	0.94	0.99
8	0.18	0.87	0.90	0.33	0.77	0.97	0.99
9	0.19	0.95	0.78	0.31	0.76	0.96	0.98
10	0.20	0.86	0.71	0.11	0.77	0.97	0.99

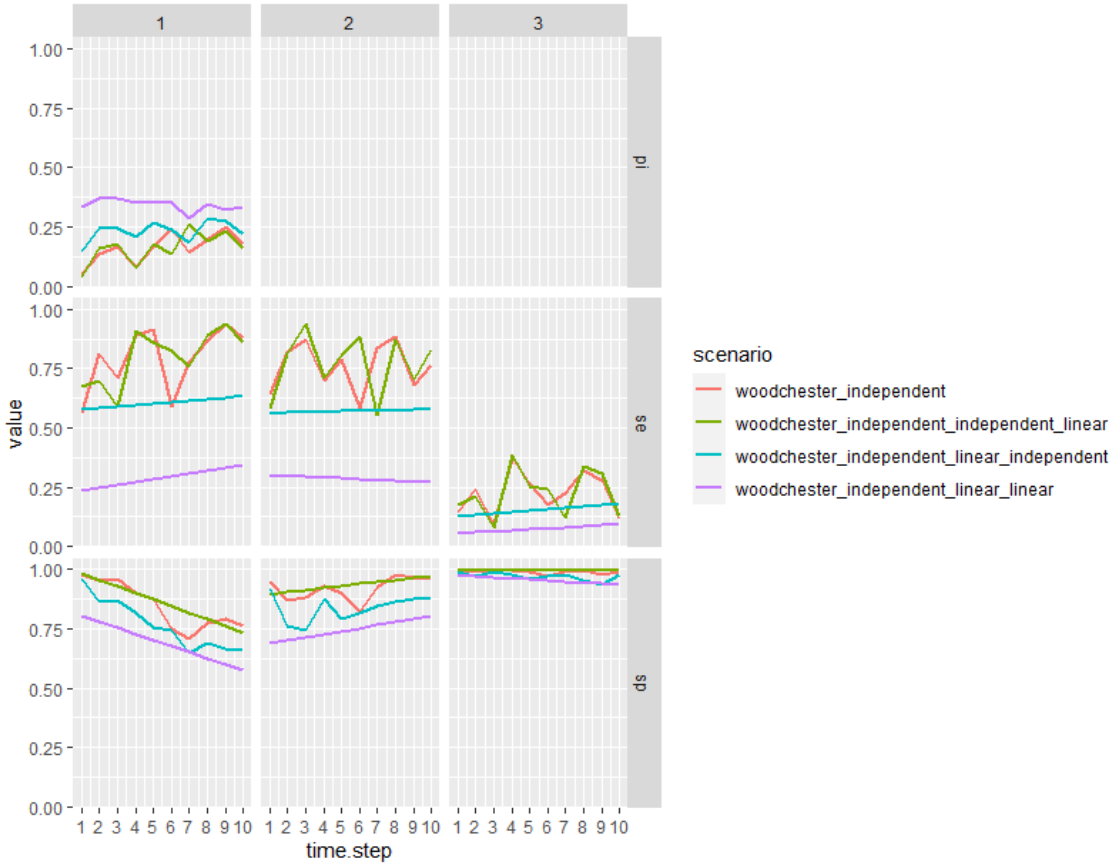

Figure 8-16: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when P is assumed to be independent of time, and Se and Sp are assumed to have either an independent or linear relationship with time. The lack of consensus across all four models could indicate that P should not be modelled as independent from time. Furthermore, the models that assume Se varies independently with time appear to agree, while the models that assume Se varies linearly with time do not. This may indicate that Se should be modelled as time independent.

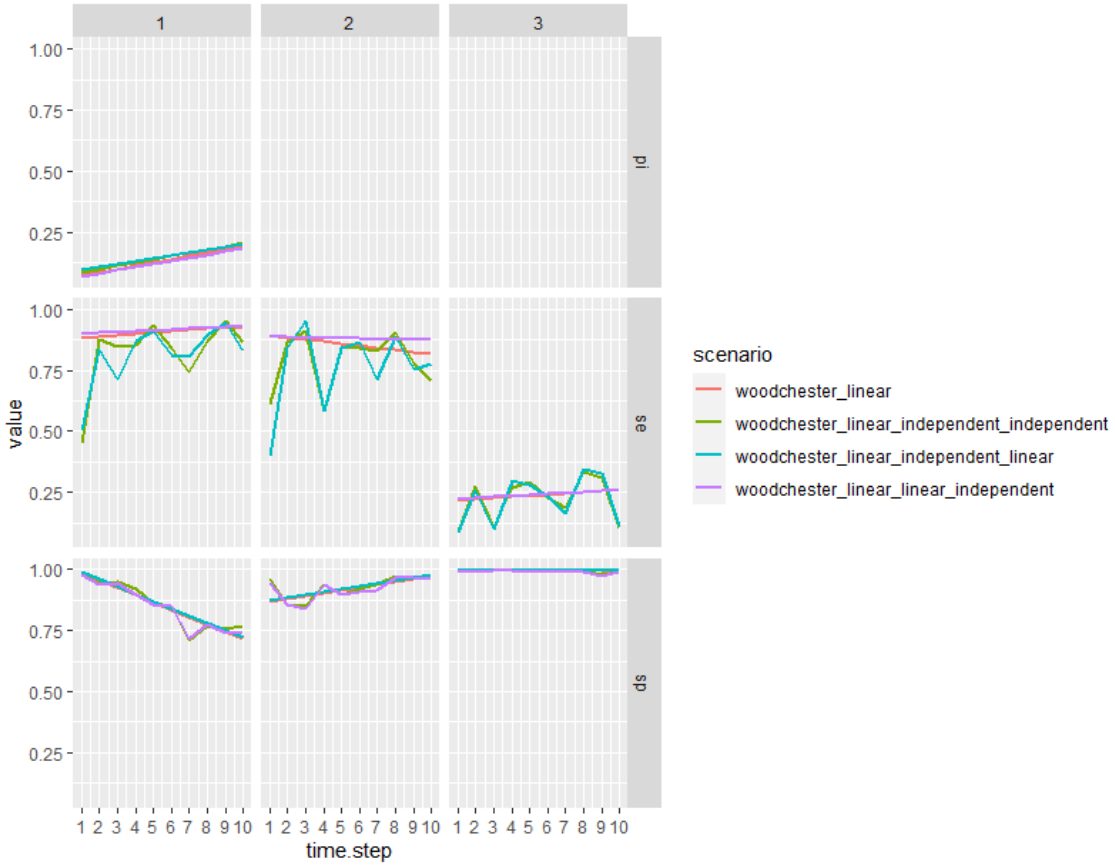

Figure 8-17: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when P is assumed to have a linear relationship with time, and Se and Sp are assumed to have either an independent or linear relationship with time. The strong agreement between the green and blue, and purple and orange inferences respectively concur with Figure 8-16, that P should be modelled as having a linear relationship with time. Furthermore, both models where P is modelled as linear with time and Se is modelled as time independent (green and blue) appear to strongly correlate, indicating that it matters little whether Sp is assumed to vary linearly with time, or be time independent.

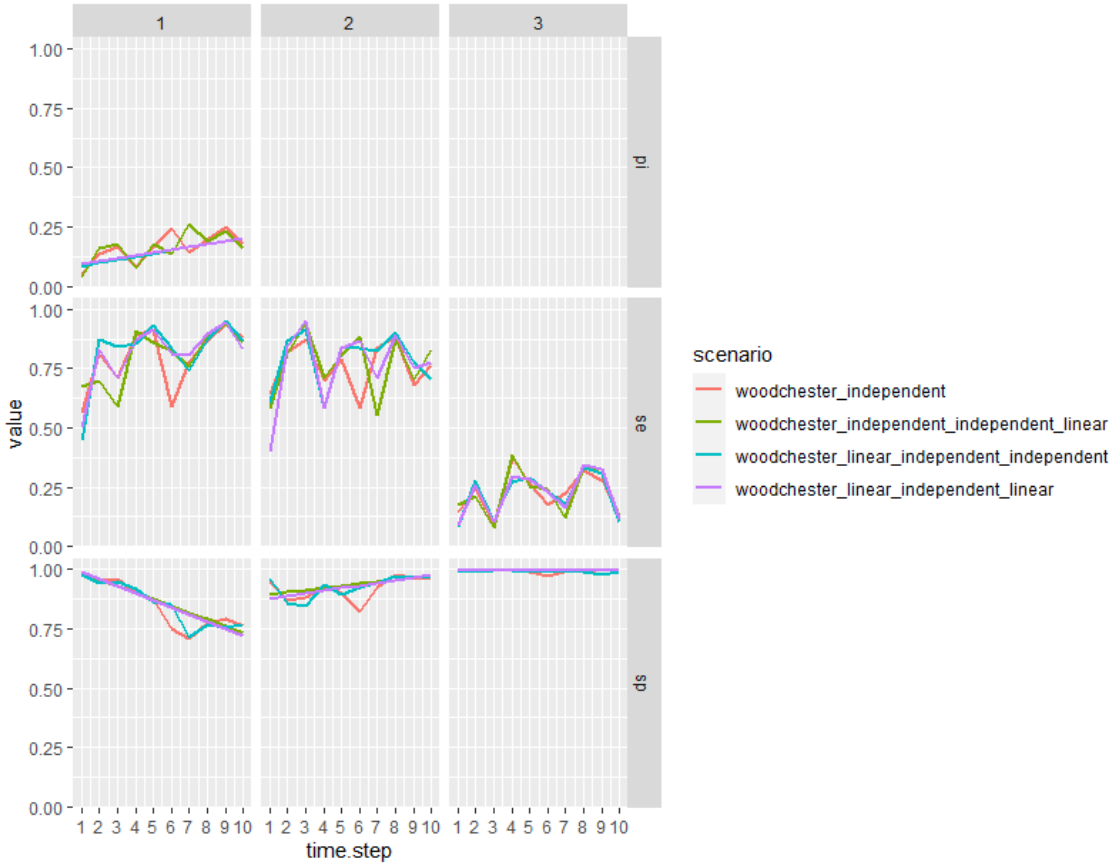

Figure 8-18: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when Se is assumed to be independent of time, and P and Sp are assumed to have either an independent or linear relationship with time. All four models appear to be in agreement, which may indicate that, in concurrence with Figure 8-16, Se should be modelled as time independent.

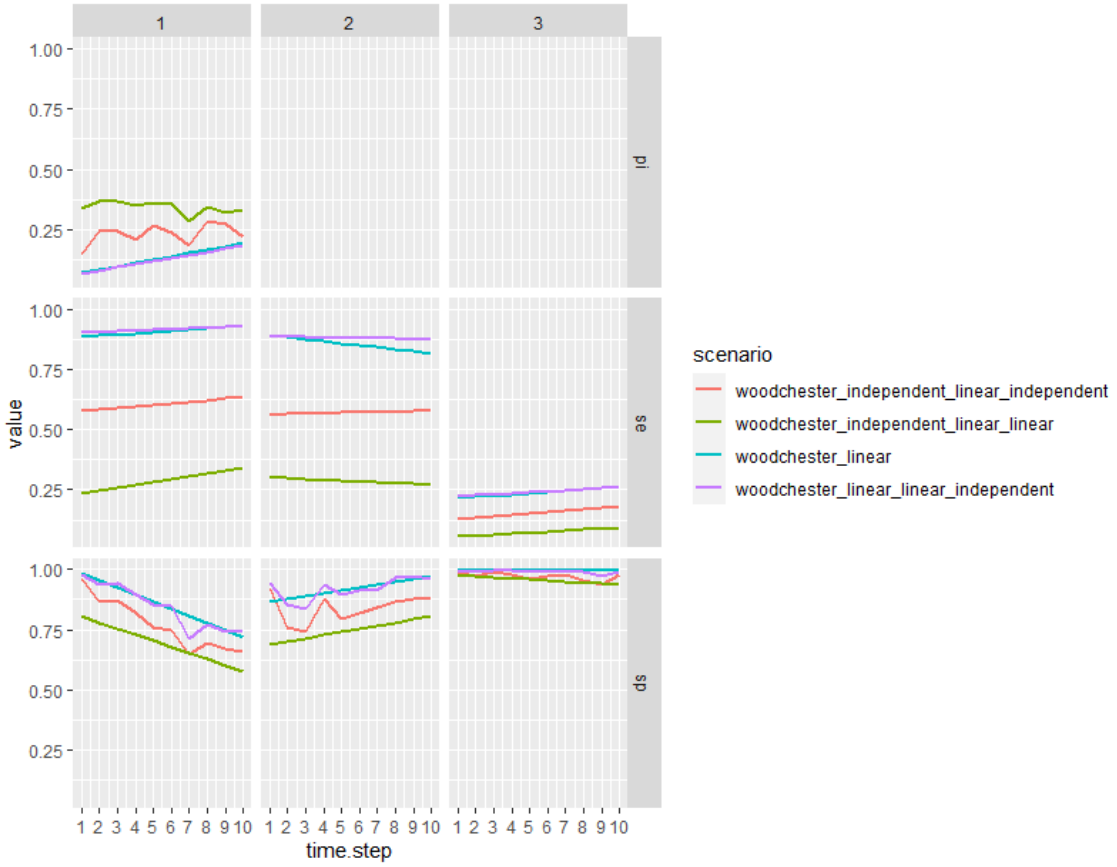

Figure 8-19: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when Se is assumed to have a linear relationship with time, and P and Sp are assumed to have either an independent or linear relationship with time. The lack of consensus across the four models indicates that Se should not be modelled as having a linear relationship with time. Furthermore, both models that assume P varies linearly with time appear to agree, while both models that assume P is time independent do not, which may indicate that, in concurrence with Figure 8-18, P should be modelled as linear with respect to time.

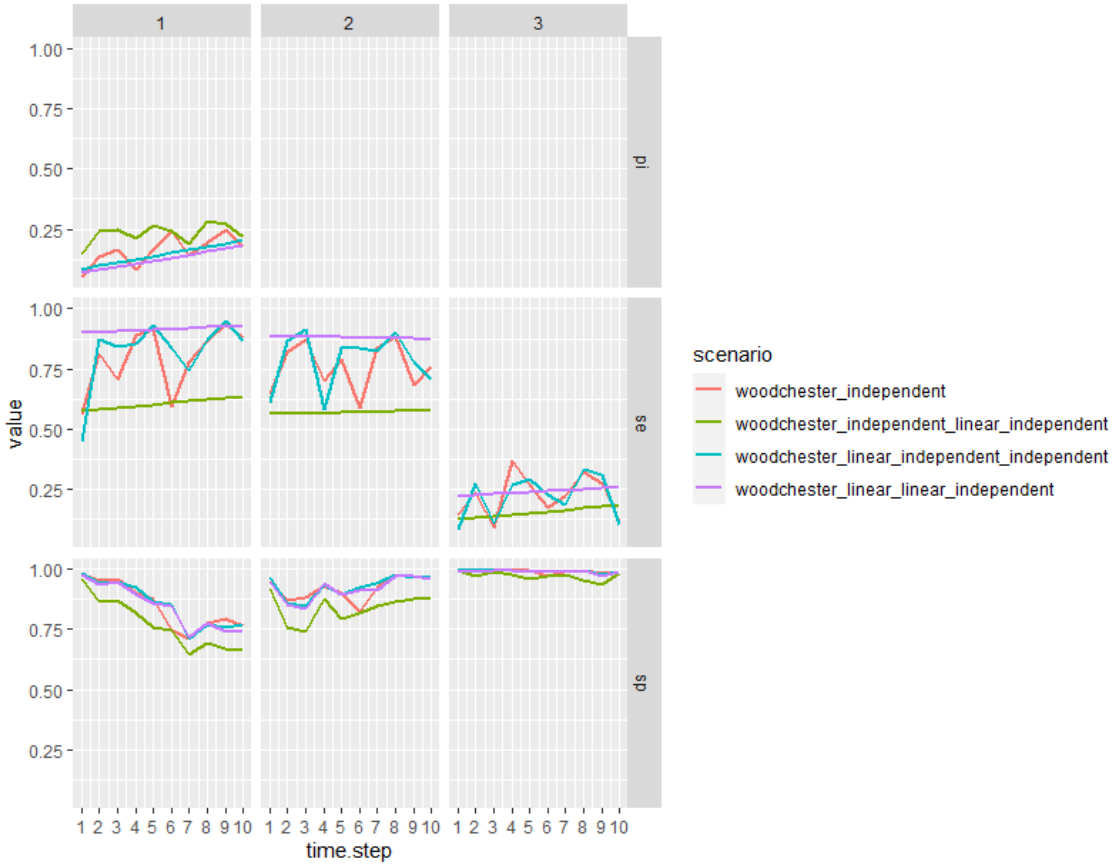

Figure 8-20: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when Sp is assumed to be independent of time, and P and Se are assumed to have either an independent or linear relationship with time. Assuming that P should be modelled as having a linear relationship with time and Se should be time independent, the lack of consensus across these four models is to be expected.

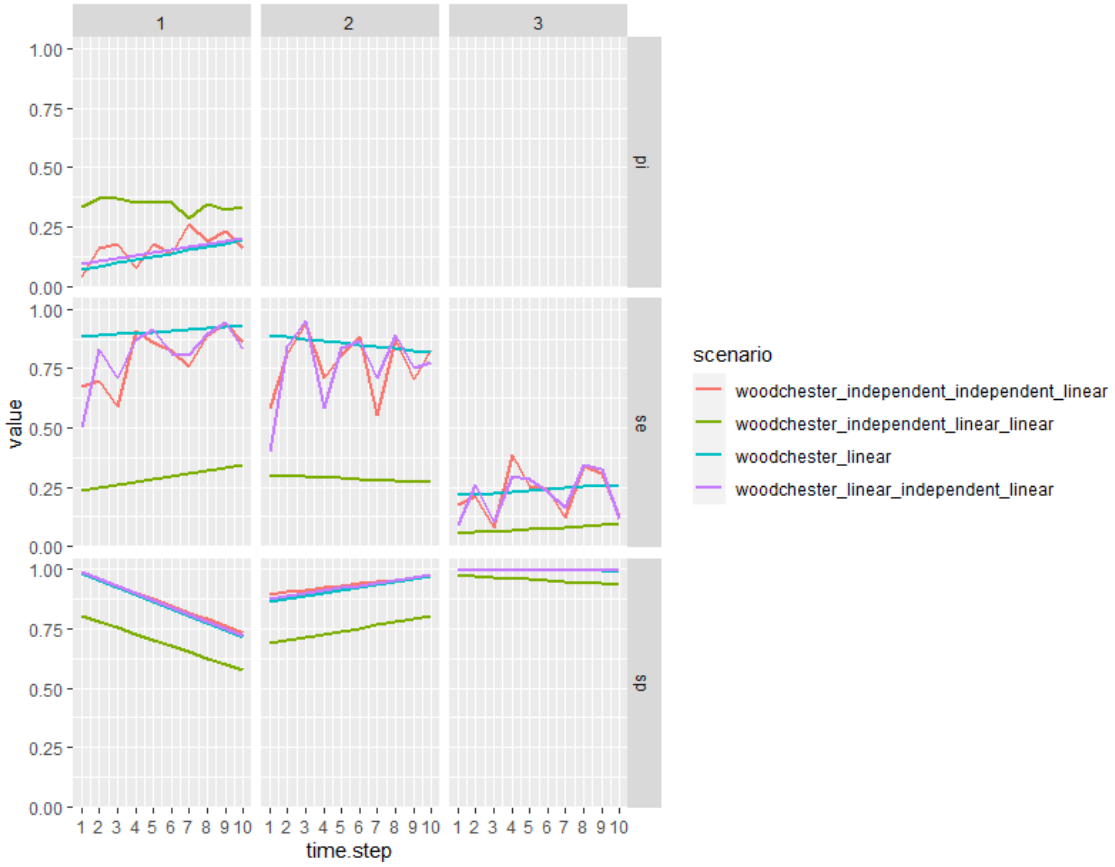

Figure 8-21: The inferred values of Se, Sp and P for the Woodchester battery of diagnostic tests when Sp is assumed to have a linear relationship with time, and P and Sp are assumed to have either an independent or linear relationship with time. Similar to Figure 8-20, assuming that P should be modelled as having a linear relationship with time and Se should be time independent, the lack of consensus across these four models is to be expected.

Table 8-13: Previous estimates (in bold) of Se, Sp and P given the Woodchester Park diagnostic test data, and the time periods that those estimates concern, in comparison to the estimates presented in Chapter 8 by the “best” models. Between 2006 and 2008, the published estimates were obtained using Latent Class Modelling techniques given data from 305 individuals tested using a battery of three diagnostics over 2.5 timesteps (Drewe *et al.*, 2010). Between 2006 and 2013 a multi-event capture-recapture approach was used given data from 541 individuals tested using a battery of three diagnostics over 8 timesteps (Buzdugan *et al.*, 2017). All estimates have been rounded to two decimal places.

Diagnostic Test	**Model**	Diagnostic accuracy across specified years					
		2006 to 2008		2006 to 2013		2006 to 2015	
		Sehat	Sphat	Sehat	Sphat	Sehat	Sphat
gamma interferon release assay	published	**0.8**	**0.95**	**0.52**	**0.97**	NA	NA
	woodchester_independent	0.78	0.9	0.76	0.91	0.76	0.92
	woodchester_linear_independent_linear	0.73	0.89	0.76	0.91	0.76	0.93
	woodchester_linear_independent_independent	0.79	0.89	0.79	0.91	0.79	0.93
BrockTB Stat-Pak test	published	**0.5**	**0.97**	**0.58**	**0.97**	NA	NA
	woodchester_independent	0.7	0.96	0.76	0.86	0.8	0.85
	woodchester_linear_independent_linear	0.68	0.96	0.68	0.89	0.81	0.85
	woodchester_linear_independent_independent	0.72	0.95	0.8	0.87	0.82	0.85

mycobacterial culture test	published	**0.08**	**0.99**	**0.08**	**1**	NA	NA
	woodchester_independent	0.16	0.99	0.23	0.99	0.22	0.99
	woodchester_linear_independent_linear	0.15	1	0.22	1	0.22	1
	woodchester_linear_independent_independent	0.16	1	0.22	0.99	0.22	0.99

Discussion

The research in this chapter demonstrates to ecologists that time-dependent BLCMs can be used to robustly infer the Se, Sp and P of a real-world dataset at specific timepoints, advancing the power and usefulness of Bayesian Latent Class Analytics to the discipline. The methodologies and models presented can be applied to any array of diagnostic test data, and any time interval of interest, and may also be readily modified to make and explore spatially-dependent inferences. It has been demonstrated that synthetic datasets—with known truths that mimic likely real-world datasets—can be used to assign confidence to the specification of time-dependent BLCMs to infer values of Se, Sp and P. And critically, the use of uninformative priors proved that the posteriors reported are driven by the Woodchester test data, rather than the set of strongly informative prior distributions used to inform the previously published studies on the same Woodchester test data using Bayesian methods (Drewe *et al.*, 2010; Buzdugan *et al.*, 2017; McDonald and Hodgson, 2018).

Overall, this chapter has contributed new model validation methodologies for time-dependent BLCMs, proof that time-dependent BLCMs increase the analytical power of diagnostic test data, and producing robust inferences of Se, Sp and P between 2006 and 2015 at Woodchester Park.

For the first time—using time-independent and time-dependent BLCMs—the parameters Se, Sp and P are inferred for each year of a decade of trap and test data, which in this case belongs to the Woodchester Park study. It is found that Se, Sp and P each change over time in the Woodchester population in different ways.

On average, between 2006 and 2015, the values of P increased across time by 10% whereas the values of Sp for the BrockTB Stat-Pak test decreased across time by 27% from 2006 to 2012. It was found that distinct change points and trends do exist among the yearly inferred values of Se, Sp and P, and these are now discussed in turn.

Sehat

The Se of the Woodchester diagnostics was found to be variable across years. For example, a 36% decrease in the Sehat of the gamma interferon release assay was identified between 2008 and 2009, following the finding of a 12% decrease in the Se of the BrockTB Stat-Pak test between 2007 and 2008, and a 16% drop in the Se of the mycobacterial culture test in this same period. Despite these substantial decreases in test performance, Phat only increased in the Woodchester badger population by ~3% between 2006 and 2009. These change points of Sehat demonstrate how time-dependent BLCMs can be used to identify setting-dependent differences in the ability of a diagnostic test to detect infection across time, which may include a diagnostic test's ability to detect latent infection.

Given that the discovered change points are not reflected in the inferences of Sp or P—and that they are present for all three diagnostic tests—these change points are most likely to be dependent on ecological factors unknown to this study that have influenced population-level disease outcomes. It is also likely that these ecological factors are system specific. Interestingly, the Government's trial of vaccinating badgers against bTB with Bacille Calmette–Guérin was conducted between 2006 and 2009 on social groups of badgers in the same geographical region as Woodchester Park (Carter *et al.*, 2012). And more generally, information from time-dependent BLCMs could be used to

answer questions such as “why is it important to vaccinate badgers?” by providing specific inferences of P.

Importantly, there are discrepancies between the published values of Se at Woodchester Park (Table 8-13) for the three diagnostic tests of interest to this study, and the values inferred within this study using versions of the Three Test Five Timestep independent and linear models (see Table 8-10 and Table 8-11). While these discrepancies are unsurprising—the published BLCMs of the Woodchester system (Drewe *et al.*, 2010; McDonald and Hodgson, 2018) have not been extensively validated against simulated and stochastic test data; they rely on the same expert-elicited prior information; and they have not been subject to time decompositions—it is possible that values of Se were particularly difficult for the BLCMs to accurately and precisely infer given the Woodchester Park test results.

Importantly, despite large discrepancies in some year-on-year inferences of Se compared with the average estimates reported by Drewe and Buzdugan *op cit.*—for example, in 2010, the woodchester_independent model inferred the Se of the BrockTB Stat-Pak test to be 33% higher than the published value for the period covering 2010—the findings reported within this chapter agree with these previous estimates on average (Table 8-13). Moreover, the time decompositions reveal a lot more variation in Se than previously reported, suggesting that time decompositions are critical for researchers wishing to optimise Se.

Sphat

The Sp of the BrockTB Stat-Pak test was observed to have decreased by 22% between 2006 and 2015 (using the woodchester_independent model), meaning

that the number of falsely identified infected individuals throughout this decade is likely to have increased, potentially explaining the high number of positive test results recorded at Woodchester Park by the BrockTB Stat-Pak test (in comparison to the number of positive test results recorded at Woodchester Park by the gamma interferon release assay). This finding supports the hypothesis that motivated this thesis chapter: that the Sp of the BrockTB Stat-Pak test changes within the 2006 to 2015 period.

Similar decreases in the Sp of the gamma interferon release assay or the mycobacterial culture test were not detected, indicating that there is a latent and test-specific dependency on the Sp of the BrockTB Stat-Pak test. Interestingly, previous research (Carter *et al.*, 2012) on Gloucestershire badgers has found that the incidence of positive BrockTB Stat-Pak test results can decrease when badgers are vaccinated with Bacille Calmette–Guérin; and some social groups of badgers at Woodchester Park are vaccinated with Bacille Calmette–Guérin. However, the decrease in the performance of the BrockTB Stat-Pak test alone may also be due to any latent process that interacts with the badger-bTB host-pathogen system, such as demographic trends, or whether there are multiple strains of bTB present in the population. Indeed, there is evidence to suggest that Stat-Pak may have become less useful at detecting badgers with the greatest transmission risk (Chambers *et al.*, 2008).

Phat

In contrast to the predictions about P made by Rogers *et al.*, 1999 and Mcdonald *et al.*, 2016—as stated in the methodology of this chapter—this study finds that P assumes a linear and non-cyclical trend between 2006 and 2015 in the Woodchester Park badger population. A similar trend has been reported in terms of the annual apparent prevalence of bTB in cattle in the High Risk Area

that Woodchester Park was located within, in the period of this present study (More *et al.*, 2018). It has been reported that prior to 2010, the P value of the Woodchester Park population had a 95% chance of falling within the range of 16–35% (Drewe *et al.*, 2010)—with true P likely being slightly higher than the estimated 2011 UK national average of ~16.6% (Allen, Skuce and McDowell, 2011). However, this study finds that P could have been as low as 0.05 in 2006 and as high as 0.25 in 2014 (using the woodchester_independent model); and when P is assumed to have a linear trend across time (between 2006 and 2015), models indicate that P increased from 0.1 to 0.2 at a steady rate throughout this period. The values of Phat reported in this chapter are in closer agreement with the predicted annual estimates of P at Woodchester Park between the years 1982 and 1996, which are between 10.3% and 17.7% (Delahay, Langton, *et al.*, 2000). Considering the potential number of false positive records in the Woodchester Park dataset attributable to the decreasing performance of the bTB BrockTB Stat-Pak test between 2006 and 2015, it is possible that there are significantly fewer bTB infected badgers at Woodchester Park than previously assumed.

Investigating spatial dynamics next?

This chapter has not investigated the possibility that in the real-world, Sehat, Sphat and Phat change as a function of the *space* across which they are being inferred. If Sehat, Sphat and Phat have latent dependencies on latent variables that are known to be heterogeneous across spaces—such as the genetic strain of *Mycobacterium bovis* as identified by spoligotyping (Swift *et al.*, 2021)—then, in theory, the power of time-dependent inferences may be further improved. That is, if the relationship between diagnostic accuracy and space is understood, and diagnostic test data grouped by location is available, it is

hypothesised that the power of inferences of diagnostic accuracy will increase, and in turn, allow spatially-dependent inferences of P.

Conclusion

This chapter has demonstrated how time-dependent BLCMs can be specified, validated, and used to infer Se, Sp and P through time from historic datasets. The findings provide evidence to support the assumed existence of trends and change points in the year-on-year performance of the Woodchester Park diagnostics, as well as in the resulting P values. Importantly, since no prior information was provided to the BLCMs, all of the results were driven solely by the raw diagnostic test data. It was observed that the values of P increased linearly across time at a rate of 1 percentage point per year on average, while the values of Sp for the BrockTB Stat-Pak test significantly decreased across time. Importantly, on average, the inferred values of Se and Sp agree with the existing literature on the Woodchester Park bTB study, although evidence has been presented to suggest that values of P have been overestimated within these studies. Moreover, given the high year-on-year variability in the performance of the Woodchester Park diagnostic battery, BLCMs in general must be able to account for changes in a diagnostic test's ability to detect infection as disease progresses. There is therefore a need for ecologists to model Se, Sp and P across independent time points to ascertain specific historic change points, and also across suspected trends in Se, Sp and P across time, to understand average changes in these parameters. In addition, this chapter illustrates that there is a demand for robust and standardised Bayesian model comparison tools. Finally, based on the high year-on-year variability in the performance of the Woodchester Park diagnostic battery, it is

recommended that ecologists should relax the assumption of the Hui-Walter theorem that Se and Sp are independent of diagnostic testing scenarios.

CHAPTER 9

9. A summary of the contributions of this thesis and their impacts.

Overview

BLCMs are recognised as belonging within a class of models representing the state of the art for diagnosing infection in the absence of a gold standard test, and their application to the problem of diagnosing infection in wild animals is growing in frequency and importance. Yet evaluations of the usefulness of BLCMs in influencing wildlife disease management decisions are scarce, with a recent study on Leptospira infection in California sea lions (Helman *et al.*, 2020) being the only obvious publication to cite at the time of writing this thesis.

Across the five empirical chapters presented in this thesis (Chapter 4 to Chapter 8), BLCMs have been applied to simulated data that was representative of diagnostic testing scenarios in diseased wildlife populations, and advances are identified in the application of BLCMs to the problem of how to accurately diagnose infection in wild animals.

The paucity of information on ante-mortem disease states combined with a lack of standardised methods to evaluate the power of BLCMs were key motivators of this thesis. And in response, the studies presented demonstrate how simulated data and the Bayesian approach to LCM—as opposed to numerical estimation—enables information on disease states to be utilised flexibly.

Researchers generally agree that Sehat, Sphat and Phat provide useful information when inferred using a BLCM specified to represent a Three-Test,

One-Population scenario. However, given that the Three-Test, One-Population scenario is associated with enough degrees of freedom to theoretically infer all required parameters—in most instances—it would be difficult not to achieve useful values of Sehat, Sphat and Phat with a deterministic modelling setup where an entire population is selected and tested. In response, while the studies presented in this thesis concern diagnostic scenarios where gold standard tests are unavailable, the data that informs these studies also accounts for the fact that it is impossible to achieve a perfect trapping effort in a wild animal population.

So, what does this thesis contribute?

This thesis provides original contributions to advance the discipline of disease ecology through the creation and interpretation of a substantial body of new knowledge regarding the inference of Se, Sp and P using BLCMs, which has resulted from testing and uncovering new theories and hypotheses.

The contributions of this thesis can be considered in terms of its methodological contributions—inclusive of the library of bespoke functions, and the relevant R code used to formulate them (see https://github.com/annabush/PhD)—as well as the findings that these methodologies enabled, inclusive of the new theories and hypotheses that are put forward, and the extensive directory of code used to manipulate thousands of posterior inferences (Table 10-1) into formats suitable for plotting graphs and inputting into regression models.

Five key contributions are made:

1. A framework for the inference of P that (i) generalises the classic Hui-Walter model for the handling of any number of diagnostic tests and populations that may describe a wildlife disease study; and (ii) describes

how diagnostic test data can be generated to account for the noise of trapping and testing live animals. This is the first known instance of an Any-Test, Any-Population BLCM being openly specified using BUGS-type code, and only one other relevant study (Helman *et al.*, 2020) has been found to explore the inferences of Se, Sp and P using stochastic diagnostic test data. The modelling framework presented is crucial for understanding the power of user-specified BLCMs using simulation analyses, prior to their application to real-world testing scenarios.

2. Methodologies and hypotheses that contribute to an improved validation of BLCMs. These contributions provide a template to guide the validation of BLCMs in a field where no relevant guidance exists, but where the inferences of Se, Sp and P depend upon applying a model, with credibility, to the specific diagnostic testing scenario in which a researcher is interested.
3. The identification of two statistical artefacts important to reporting credible inferences from BLCMs: (i) the reciprocal relationship between Sehat and Sphat and (ii) mean-variance relationships across parameter space. The existence of these artefacts is tested, and advances in the understanding of these artefacts are made. Understanding the statistical artefacts that may apply to any specific diagnostic testing scenario is prerequisite to understanding the credibility of any inferences of Se, Sp and P relating to that scenario.
4. Methodologies to understand how generalisations of the Hui-Walter model are sensitive to changes in model assumptions and new information. These methodologies are critical to the credibility of any real-

world study, since the range of possible truths that may be encountered in nature is always unknown, and may be unexpected.

5. Methodologies enabling BLCMs to infer the Se, Sp and P of real-world data through time. These methodologies are critical to the discovery of trends and change points in diagnostic test data, which can provide valuable data for predictive models of P, and subsequent disease management decisions.

The application of these five core methodologies to the experimental designs—inclusive of data, assumptions and modelling conditions—that are analysed within this thesis resulted in a substantial body of new findings, which can be summarised on a chapter-by-chapter basis as follows:

- **Chapter 4** establishes the key relationships between BLCM model specifications and the errors of the resulting inferences for further investigation.
- **Chapter 5** discovers statistical artefacts key to the interpretation of BLCM posterior inferences, inclusive of relationships between error and position in parameter space, and the existence of edge effects.
- **Chapter 6** reports on the mean-variance relationships that exist across the constrained parameter spaces available to the experimental BLCMs—as identified in Chapter 5—and contributes further findings on the fundamental relationship between error and the number of diagnostic tests available as first reported in Chapter 4.
- **Chapter 7** provides evidence to demonstrate that the findings made in Chapters 4 to 6 are robust across most diagnostic testing scenarios, and further highlights where inferences should be interpreted with caution.

- **Chapter 8** demonstrates that it is possible to infer changes in Se, Sp and P across time using real-world data, opening up a range of questions behind the efficacy of diagnostic testing in animal populations, and certainly at Woodchester Park.

The following text highlights the core challenges behind, and contributions and impacts of, the five key methodologies presented within this thesis.

Contribution 1: Developing a framework for the inference of P that (i) generalises the classic Hui-Walter model for the handling of any number of diagnostic tests and populations and (ii) describes how diagnostic test data can be generated to account for the noise of trapping and testing live animals.

The challenge

Although simulation analyses are commonly used to research data-poor problems, for the problem of diagnosing infection in wild animals, the data needs to be representative of imperfect trapping and testing. Therefore, the first challenge addressed by the body of work in this thesis was the development of a framework to generate noisy data, handle a variety of user-changeable modelling conditions, run over the required number of simulations automatically, and process the large amount of data efficiently. Resolving this challenge was the purpose of Chapter 3.

The contribution

This challenge was overcome with the creation of a modelling architecture capable of both generalising the original Hui-Walter construct to handle any number of tests and populations—by creating an Any-Test, Any-Population

construct—and allowing the stochasticity of synthetic diagnostic test arrays to represent imperfect trapping and testing.

This modelling architecture is described in Chapter 3, and delivers three key advances:

1. An environment where the practical advantages of Bayesian inference—in terms of the how prior information is specified—can be combined with the relaxation of the common assumptions that Sp should be fixed or close to 100%, which is often not the case for real-world tests; and that perfect trapping and testing efficiencies should be modelled within simulation analyses.

2. An environment where the focus of the user shifts towards evaluating the reasonableness of assumptions and provision of informative prior information, ultimately ensuring that BLCM inferences, when given real-world data, are scientifically reasonable.

3. Guidance on how to calibrate three key performance indicators of BLCMs: specifying prior information, issues of non-convergence, and understanding the metrics accuracy and precision.

Impacts

The modelling architecture developed allows flexible study designs that have been used in this thesis to:

1. Simultaneously evaluate improvements in the accuracy and precision of Sehat, Sphat and Phat given a 2nd, 3rd, 4th, and 5th, diagnostic when sampling and testing efforts are imperfect.
2. Evaluate the credibility of Sehat, Sphat and Phat across many modelling conditions, and diagnostic testing scenarios, simultaneously.

3. Validate the credibility with which a BLCM can infer values of Sehat, Sphat and Phat from simulated test results before it is applied to real-world data.
4. Test the assumptions of BLCMs readily, for example, by removing constraints on parameter space, or by changing prior precision.
5. Parameterise complex BLCMs, for example, Chapter 8 demonstrates how Sehat, Sphat and Phat can be inferred as a time series to test the assumption of non-constant diagnostic accuracy over years.

Contribution 2: Developing methodologies and hypotheses to validate BLCMs.

The challenge

Currently, even with the OIE's recommendation (Gardner *et al.*, 2021) for the use of LCMs for diagnosing animal disease, and guidelines (Kostoulas *et al.*, 2017) for presenting research using BLCMs, ecologists still lack a standard protocol for describing how to validate their custom built BCLM algorithms—a procedure that should occur before any model selection (Hooten, Hobbs and Ellison, 2015) takes place, before any diagnostic test performances are validated, and certainly before any research is presented.

The application of BLCMs for disease management seems conflicted by a confusion over how best to validate models (Augusiak, Van den Brink and Grimm, 2014), and also by the complexities of modelling uncertainty in natural systems themselves (Dietze, 2017). This has led to uncertainty analyses in ecology being sparsely applied (Hines, Ray and Borrett, 2018; Yanai, See and Campbell, 2018), complex (Milner-Gulland and Shea, 2017; Lachish and Murray, 2018), and difficult to quantify (Wu and Li, 2006), as well as to different

sources of uncertainty being poorly defined (Regan, Colyvan and Burgman, 2002).

The use and interpretation of BLCMs therefore demands care (Schofield *et al.*, 2021) and is dependent upon understanding how accurately BLCMs infer Se, Sp and P in the relevant parameter space. To do this, it is important to validate the power of a BLCM before inputting real diagnostic test data, in order to ensure that a BLCM can perform as expected. Resolving this challenge was the purpose of Chapter 4.

The contribution

Accordingly, Chapter 4 makes three core contributions to the validation of BLCMs:

1. The development of methods to validate BLCMs, which may be replicated by ecologists wishing to evaluate their own BLCMs.
2. The specification of LMM's, which are shown to be a useful means to interrogate the accuracy of Bayesian inferences, revealing the structure of the random effects that influence accuracy, and going some way towards explaining parameter-specific errors.
3. Seven general trends, termed "stylised facts" relating to evaluating the accuracy of BLCM inferences to be identified across the parameter spaces studied, demonstrating the specific type of information that can be gathered from validating BLCMs. These stylised facts also held true when simulation studies became more complex in later chapters.

Four notable findings were made:

1. Practical identifiability is influenced by the number of tests available, model constraints, and prior precision.

2. For any modelling scenario, the parameters Sphat and Phat are generally less accurately inferred than Sehat.
3. A trade-off between the accuracies of Sehat and Sphat appears to exist.
4. A positive relationship between diagnostic accuracy and the number of diagnostic tests available was demonstrated graphically, and produced for the first time using stochastically generated data. As a result, it was hypothesised that where the n.tests trend exists, practical identifiability of a BLCM is possible; a hypothesis supported by the findings presented in Chapters 4 to 8.

Importantly, while the degrees of freedom rule may explain where practical identifiability can be found, the accuracies, and precisions of Sehat, Sphat and Phat are found to be dependent on a range of prior information being available, with differing numbers of diagnostic tests available being only one type of prior information with which a model could be provided. This finding sits in agreement with research such as Jones *et al.*, 2010 and Goodman, 1974, which has already proved that the degrees of freedom rule alone cannot determine whether parameters of a Latent Class Model are identifiable.

Impacts

In ecology, mechanistic models are often used to justify ecological findings, and ecological justifications (Lindén and Mäntyniemi, 2011) are often used to determine model specifications. In contrast, the studies presented utilise LMM's as a mechanistic way to understand datasets of the accuracies and precisions of Sehat, Sphat and Phat given truth. Yet the justifications for the findings that are reported in this thesis, as well as the models that are specified, are statistical in nature; this illustrates a critical step-change in how ecologists may wish to think about evaluating their BLCMs and sits in agreement with

(DiRenzo, Hanks and Miller, 2023) who emphasise that determining the statistical properties of an “estimation approach” is a critical step to model validation.

The model validation approaches described in Chapter 4 enabled the identification of seven stylised facts, which have implications for researchers who:

1. Consult the medical and the wildlife literature for what to consider when validating a BLCM.
2. Are considering the intended use of a specific diagnostic test, which may be a proxy test.
3. Have the freedom to determine diagnostic thresholds.
4. Aim to understand the limitations of their BLCMs.
5. Wish to understand how information should be added into BLCMs to improve the inference of Se, Sp and P.

Contribution 3: Advancing understanding of two statistical artefacts important to understanding the inference from BLCMs: (i) the reciprocal relationship between Sehat and Sphat and (ii) mean-variance relationships across parameter space.

Note, contribution 3(i) relates to the findings of Chapter 5, and contribution 3(ii) relates to the work presented in Chapter 6.

For real-world studies, if unaccounted for, the inaccuracies associated with the presence of artefacts—trends explainable by statistics rather than ecology—could have direct disease management implications. Key to avoiding these inaccuracies is understanding how BLCM identifiability changes across regions

of parameter space, which may be attributed to artefacts. And key to monitoring changes across regions of parameter space is being able to represent high-dimensional parameter space on an easily interpretable scale, accounting for key sources of bias. Methodologies that can represent high-dimensional parameter space are therefore critical for understanding where BLCM inferences lack identifiability across a wide range of possible infection scenarios.

The premise of this challenge is usefully explained in the following quote: "*Bayesian inference is conditional on the space of models assumed by the analyst. Within that assumed space, the posterior distribution only tells us which parameter values are relatively less bad than the others. The posterior does not tell us whether the least bad parameter values are actually any good*." (Kruschke, 2013). However little research has previously been conducted to evaluate parameter values across space.

The questions on statistical artefacts that this thesis explores are useful to ecologists as they concern the "simplest" problem that a researcher may wish to ask about parameter space—*"is there a region of my parameter space where condition X holds?*" (Chalom and de Prado, 2012)—where condition X is practical identifiability.

Contribution 3i: Advancing understanding of the reciprocal relationship between Se and Sp

The challenge

In non-gold experiments ecologists must invariably accept a trade-off between the Se and Sp available (Lütkenhöner and Basel, 2013), yet this dynamic is rarely quantified. For studies adopting BLCMs researchers should understand

the error, bias and precision associated with Sehat and Sphat. Chapter 5 addresses this challenge.

The contribution

Two methods were developed and explored in Chapter 5:

1. A method to sample across high-dimensional parameter space, allowing the dynamics between Sehat and Sphat to be quantified across parameter space of up to 11 dimensions.
2. A method to map the error, bias, and precision of BLCM inferences, allowing the production of a series of heatmaps and regression analyses that represent uncertainty across parameter space as modelling conditions are varied.

In combination, the two methods presented offer an alternative approach to the classical ROC approach for assessing diagnostic accuracy for batteries of diagnostic tests, specifically as a contingent of the relationship between the accuracies of Sehat and Sphat values. No known study has provided a methodology to evaluate how the relationship between Sehat and Sphat across hyper-dimensional parameter space can be optimised.

In addition, advances are also made in uncovering structured patterns in the variance of error across parameter space, allowing hypotheses of edge effects to be made.

These advances led to the following key findings:

1. Phat has an intricate relationship with Sehat and Sphat which is not the same as between Phat and Sehat, and as between Phat and Sphat.

2. There is a strong relationship between error, bias and precision, and position in parameter space.
3. Edge effects exist in “extreme” parameter space, and these effects interact with the provision of prior precision and constraint.

Contribution 3(i) will be of interest to researchers who wish to understand:

1. How the accuracy and precision of—and the relationship between—Sehat and Sphat can be dependent on how the BLCM is specified.
2. Combinations of parameter values that lack practical identifiability.
3. The unstable properties of the relationship between Sehat and Sphat.
4. When to use global error metrics.
5. How to map the variance of a parameter across parameter space.

Impacts

Understanding the bias associated with diagnostic outcomes in specific volumes of parameter space is useful to any researcher with population-level diagnostic test outcomes which are a mixture of positive and negative test results. For example, it is found that when infection rates are low at ~0.15, including more diagnostic tests would be the best way to improve BLCM inferences of Sp, compared to the addition of information via other means, such as more informative priors.

Fundamentally, methods to evaluate whether Sehat, Sphat and Phat are “*actually any good*” (Kruschke, 2013) are needed so that ecologists can be informed about when it is useful to use BLCMs to support a diagnostic testing regime.

Contribution 3ii: Advancing understanding of the mean-variance relationships across parameter space.

The challenge

When analysing a multivariate problem across space, changes in the variance of BLCM inferences across space are to be expected, as well as a variance in the prior information available.

Volumes of extreme parameter space have been identified by this thesis as dependence structures—dependencies between two or more variables of interest—critical to BLCM identifiability. A relationship between mean error and its variance across space was uncovered in Chapter 6. Consequently, a need arose to understand the influence of this mean-variance relationship on the ability of the BLCM to infer parameters, as well as uncover any identifiability issues present.

The contribution

Methods to produce heatmaps were developed, which for the first time demonstrate how the errors of Sehat, Sphat and Phat change across parameter space given different modelling conditions. In addition, regressions were specified to demonstrate the dependencies that occur between error, bias and precision of Sehat, Sphat and Phat, and model conditions. Overall, the heatmaps and regression analyses presented in Chapter 6 allowed recommendations on how heteroscedasticity across parameter space should be interpreted.

The key findings were:

1. At edges, ecologists should consider the error Phat, Sehat and Sphat separately, rather than rely on a global metric.

2. The *shape* of the mean-variance relationships belonging to the errors of Sehat, Sphat and Phat all exhibit heteroscedasticity, and were found to be highly distinctive given any form of prior information.
3. The *size* of the edge effects on the errors of Sehat, Sphat and Phat is dependent on the amount of prior information provided.
4. The n.tests trend is a relevant consideration for researchers seeking identifiability in extreme parameter space. Notably, Chapter 6 confirms that the breakdown of the n.tests trend does indicate identifiability issues, and that this trend breaks down at edges.

Impacts

Contribution 3(ii) should interest ecologists seeking to understand:

1. How BLCM inferences may vary as the information provided to the model changes.
2. Variance in BLCM inferences.
3. How to interpret inferences at the edges of parameter space.

The heatmaps presented within this thesis were developed in order to respond to the need to reduce 11-dimensional data into information that can be readily interpreted. In combination with the use of LMM's, the heatmaps proved essential to understanding the identifiability of BLCMs. More generally, this approach supports the notion put forward (Heisey *et al.*, 2010) that to advance the field, data needs to be looked at "*in as many ways as possible, Bayesian and otherwise, to ensure consistency and reasonableness. There is a paucity of useful diagnostic tools at present and this is an area that needs a lot of work.*" In addition, the heatmaps presented also contribute more widely to the use and development of "global models"—an area of ecology focussed on mapping

global (though often geographic) data describing ecological parameters (Meyer and Pebesma, 2022).

Contribution 4: Developing methodologies to understand how generalisations of Hui-Walter model are sensitive to changes in model assumptions and new information.

The challenge

In ecology, model identifiability is usually considered in terms of unique sets of parameter values that have been calibrated to maximise the likelihood of a model under certain assumptions. For stochastic models, however, not defining the "generality" of the findings (Spake *et al.*, 2022) is known to mislead likelihoods (Stocks, Britton and Höhle, 2021).

Local sensitivity analyses are the most common type of sensitivity analysis found in the field of ecology, where one parameter and or value is varied at a time (Naujokaitis-Lewis *et al.*, 2009; Olsen *et al.*, 2022) while others are fixed (Xu *et al.*, 2004), as demonstrated with "Example A" in Chapter 4 of this thesis. However, in studies across high-dimensional space, confidence in BLCM specifications and assumptions should be generalisable across the possible truths that may be encountered in nature. Global Sensitivity Analyses of BLCMs are limited by the availability of methods, including how to manage "big data". Addressing this challenge was the aim of Chapter 7.

Contribution

Methodologies are presented in Chapter 7 highlighting ways of conducting a Global Sensitivity Analysis across parameter space. The findings in question should be of interest to ecologists concerned with understanding identifiability

issues across parameter space in terms of inferential bias; edge effects; issues identifying P values around 0.5; and the dependencies between error, bias and precision on the “n.tests” trend.

Notable findings include:

1. Evidence to suggest that the findings of the preceding empirical chapters are generally robust to changes in the values of truths and the size of parameter space.
2. Evidence to demonstrate that parameter space is useful for ecologists outside of the constraints of P and Sp enforced in previous chapters.
3. A basic set of “rules” for interpreting the directionality of error in global parameter space.
4. The finding that the n.tests trend is a useful proxy for unidentifiability in unconstrained parameter space.
5. Evidence to suggest that Phat values of ~0.5 may suffer from the label switching problem.
6. The finding that conclusions regarding edge effects may not apply to unconstrained parameter space.

Impacts

The findings of Chapter 7 should be of value to ecologists concerned with determining whether their simulation models are sufficiently robust to new information, or changes in model assumptions. Evidence is also provided to support the assumption that BLCMs are not automatically identifiable simply because their degrees of freedom are as large as the number of parameters (Jones *et al.*, 2010). For example, Chapter 7 demonstrates that when using batteries of diagnostic tests, the use of tests that a typical ROC analysis would

consider as no better than chance alone does not automatically prevent identifiability.

Given that test data, model constraints, and priors interact via a complex function to enable identifiability (Joseph, Gyorkos and Coupal, 1995)—and that model validation and sensitivity analyses do not guarantee good models (Gustafson *et al.*, 2005)—this chapter's findings will not generalise to all testing scenarios. However, for ecologists wishing to conduct a sensitivity analysis of their inferences, there remain good reasons to work in a constrained parameter space where the information exists to make assumptions about truths.

Contribution 5: Developing a statistical procedure enabling the BLCM to infer the Se, Sp and P of real-world data through time.

Challenge

The values of parameters Se, Sp and P are known to be dependent on ecological variables such as the strain of pathogen, or variables that describe population demographics. Moreover, research such as Helman *et al.*, 2020 and Patel *et al.*, 2022 provide evidence to support this widely-held view in the context of ante-mortem wildlife disease studies. Yet time-dependent BLCMs for the inference of Se, Sp and P across relevant time intervals are not available to ecologists, and so trends and change points in the values for these parameters are not understood. The ability to detect trends and change points in the values of Se, Sp and P through time is powerful; the percentage of infected individuals can be understood as a function of time. This is particularly important for understanding diseases like bTB, where pathogens can exhibit latency in individuals. The information gained from applying time decompositions on historic datasets can therefore inform present disease monitoring decisions, and

ultimately, inform better future inferences of P. Accordingly, this was the challenge addressed in Chapter 8.

Contribution

Chapter 8 offers three key contributions: time-dependent BLCMs; evidence that synthetic datasets can be used to assign confidence to the specification of time-dependent BLCMs; and the first historic inferences of how the values of Se, Sp and P have changed across a decade of real-world diagnostic test data at the well-studied Woodchester Park badger population.

The time-decomposition methodologies presented in Chapter 8 advance the power and usefulness of Bayesian Latent Class Analytics to real-world wildlife disease studies, since Sehat, Sphat and Phat may be modelled as a function of the period across which they are being inferred. The time-dependent BLCMs can respond to different patterns of change through time, as well as handle the most efficient way of using degrees of freedom and prior information, including the number of tests available.

Key trends and change points in the Woodchester Park data were uncovered, namely:

1. On average, between 2006 and 2015, the values of P were observed to increase across time by 10%.
2. The values of Sp for the BrockTB Stat-Pak test were observed to decrease across time by 27% from 2006 to 2012, supporting speculations to this effect from among the research community.
3. Despite large differences between some year-on-year inferences of Se compared to average estimates for the same time interval reported within

the literature, the findings of Chapter 8 sit broadly in agreement with previously published estimates.

4. Sehat is a highly variable parameter across all tests in the diagnostic battery.
5. It is possible that there are significantly fewer bTB infected badgers in Woodchester Park than previously assumed.

Impacts

This thesis presents the first known time-dependent BLCMs, showcasing their ability to detect trends and change points in the values of Sehat, Sphat and Phat through time. This new capability significantly develops the power of the LCMs used in previous Woodchester Park badger studies, such as by Branscum, Gardner and Johnson, 2005; Drewe *et al.*, 2010a; and McDonald and Hodgson, 2018.

Critics of this finding may quickly point out that test outcomes depend heavily on the progression of disease in individual badgers, and cite work such as Buzdugan *et al.*, 2017. This thesis considers that BLCMs are key to unlocking better inferences of Se, Sp and P at Woodchester Park—particularly since previous research (Buzdugan *et al.*, 2017) indicates that new diagnostic tests need to have Se values of over 80% and Sp values of 94% or above—which in turn may unlock capabilities to form conclusions on disease progression at an individual level.

Latent infections are not explicitly considered in this thesis, and so pre- and post-infection periods are not considered at an individual level, or the possible consequential lag periods post infectivity; these dependencies are useful directions for further research. However, this thesis provides the modelling

infrastructure needed to account for changes in a diagnostic test's ability to detect infection as disease progresses, showing that the year-on-year variability in the performance of the Woodchester Park diagnostic battery is high.

Critics may also point out that the findings presented do not suggest new cut-offs for the Woodchester tests, however since these thresholds are ultimately a policy decision they are not debated in this thesis. Rather, this thesis presents a large amount of information that could help inform decisions on test performance. Critically, it is demonstrated that when using batteries of diagnostic tests, the use of tests that a typical ROC analysis would consider as no better than chance alone does not automatically prevent practical identifiability. A significant number of recommendations are also made in this thesis with respect to the reciprocal relationship between the accuracies of Sehat and Sphat, which support the argument that diagnostic uncertainty should be a key component of how to classify test results (Shinkins and Perera, 2013).

What do the contributions of this thesis mean for ecologists wishing to use BLCMs for their own research?

The methods and findings presented in this thesis offer a wealth of information to ecologists wishing to specify their own BLCMs for use in both simulation experiments and real-world disease studies. A specific set of definitions for the communication of research on BLCMs is provided at the start of this thesis, and the repository of annotated code (provided on GitHub https://github.com/annabush/PhD) is already highly generalised, and written using a package of bespoke functions, which can be easily adapted to allow the inference of Se, Sp and P given a large range of modelling conditions.

In particular, the contributions presented within this thesis cater for those who wish to account for the errors of trapping and testing infected animals in their BLCMs, as well as understand the errors associated with making their inferences of Se, Sp and P. To do this, **Chapter 3** provides methods for stochastic data generation, and a verified means of generalising BLCMs across tests and populations is provided. And based on this modelling framework, the subsequent chapters provide methods for validating BLCMs, interrogating statistical artefacts, and conducting Global Sensitivity Analyses—tasks which are essential to the production of credible inferences in wildlife disease studies—as well as the specification of novel time-dependent BLCMs.

To summarise, each of Chapters 4 to 8 offers ecologists insights into the development of robust BLCMs as follows:

Chapter 4 presents two model validation examples, which serve as a foundational template for model validation exercises. Insights into the accuracies of Sehat, Sphat and Phat across parameter space—summarised as seven stylised facts—suggest how ecologists may optimise their own BLCMs.

Chapter 5 demonstrates the critical need for ecologists to understand the relationship between Se and Sp when using batteries of diagnostic tests, expanding on the information that this thesis provides on the optimisation of BLCMs.

Chapter 6 expands on prerequisite awareness that ecologists developing BLCMs must have on the non-constant variance of error metrics across parameter space. Understanding this variance is key to correctly interpreting model uncertainty.

Chapter 7 is highly relevant for ecologists wanting to determine whether their simulation models are sufficiently robust to new information, or to changes in model assumptions.

Chapter 8 is key for ecologists who wish to understand artefacts within a time series of diagnostic test data. Capabilities are provided for ecologists to model Se, Sp and P across independent time points to detect specific historic change points, and also across suspected trends in Se, Sp and P across time, to understand average changes in these parameters.

How could the contributions of this thesis inform future wildlife disease management and or conservation policy?

Finally, the research described in this thesis may be taken forward to benefit conservation practitioners in the following ways.

1. This thesis shows how BLCMs can be used to understand how P changes given time using real-world historic diagnostic test data. This capability can be used to provide evidence on the efficacy and effects of policy decisions such as culling livestock in response to infection.
2. This thesis describes methods to achieve robust inferences of P, which are necessary to support the future decision-making and risk assessments of epidemiologists and policy makers, for instance in response to new epidemics, or when novel pathogens emerge.
3. This thesis demonstrates the importance of analysing the power of BLCMs using simulated diagnostic test data before applying BLCMs to real-world diagnostic test data. Using validated BLCMs, it is possible that more information about P can be inferred from data belonging to existing wildlife disease studies.

Concluding remarks

The research put forward in this thesis has shown that simulation studies can be used to test the assumptions and fit of BLCMs to data representing diagnostic testing scenarios in wildlife populations. In achieving this, a number of advances have been made to the methodologies used to specify and check BLCMs including: a library of new modelling architecture, including functions required for manipulating the “big data” involved; methods for BLCM model validation and sensitivity analysis; methodologies to explore the error structures of Sehat, Sphat and Phat—and for the first time—time-dependent BLCMs have been developed.

Equally as important, the directory of new findings—spread across empirical Chapters 4 to 8—resulting from these methodologies also forms a key contribution of this thesis. These core findings are communicated using a new *lingua franca* set out at the start of this thesis, and they relate to the use and specification of BLCMs in wildlife disease ecology, including the identifiability of BLCMs, the validation of BLCMs, the sensitivity of BLCMs to new data and assumptions, and the extension of the Hui-Walter model to allow the detection of change through time. These core findings include numerous new theories and hypotheses.

So, in a sentence, what should ecologists take away from this thesis? Simply this: with BLCMs and simulated diagnostic test data now established as essential research tools for the estimation of P in infected wildlife populations, a significant amount of additional information relating to the trends of Se, Sp and P can be gained from the methodologies presented, furthering the potential of BLCMs in informing and influencing wildlife disease management decisions.

10. Appendices

Appendix 1: Simulated datasets

Table 10-1: The simulated datasets used by the experiments presented within this thesis, including the dimensions of those datasets and the total number of simulations that they represent.

Simulated Dataset	Name of dataset	Method for simulated data generation	Dimensions of the simulation problem analysed	Total number of simulations
1	Validation Example 1	Chapter 4 methods	10 replicas 4 batteries of diagnostic tests 7 sets of true values	280
2	Validation Example 2	Chapter 4 methods	25 sets of true values 4 batteries of diagnostic tests 2 levels of prior precision 4 levels of constraint	800

3	Unconstrained and constrained priors, original truths, constrained truths	Chapter 5 methods	10 replicas per voxel of parameter space 5 voxels in P-direction 10 voxels in Se1-direction 5 voxels in Sp1-direction 4 batteries of diagnostic tests 3 sample sizes (500, 1000, 1500) 3 levels of prior precision 2 levels of prior constraint	180,000
4	Unconstrained and constrained priors, new truths, constrained truths	Chapter 7 methods	10 replicas per voxel of parameter space 5 voxels in P-direction 10 voxels in Se1-direction 5 voxels in Sp1-direction 4 batteries of diagnostic tests 3 sample sizes (500, 1000, 1500) 3 levels of prior precision	180,000

			2 levels of prior constraint	
5	Constrained and unconstrained priors, constrained and unconstrained truths, original truths	Chapter 7 methods	10 replicas per voxel of parameter space 10 voxels in P-direction 10 voxels in Se1-direction 10 voxels in Sp1-direction 4 batteries of diagnostic tests 3 sample sizes (500, 1000, 1500) 3 levels of prior precision 2 levels of prior constraint	720,000
6	Time decomposition validation	Chapter 8 methods	50 sets of true values 7 true parameter-time relationships 4 assumed parameter-time relationships 1 battery of diagnostic tests 1 sample size (300 individuals) 1 level of prior precision (uniform)	1,400

1 level of prior constraint

Appendix 2: Key parameters, hyperparameters and functions

Table 10-2: The standard user-changeable parameters provided to the BLCM, the abbreviations of those parameters in the format used within the supporting R code, their standardised input values if applicable, and their corresponding justifications and assumptions.

Parameter	Abbreviation	Input	Justifications and assumptions
The number of diagnostic tests	`n.tests`	2:5	In general, it was assumed that the accuracy and precision of Sehat, Sphat and Phat are positively influenced by the number of diagnostic tests available via a linear trend. However, underneath this assumption it was speculated that a step change (i.e., non-linear trend) exists in the ability of a BLCM to infer the accuracy and precision of Sehat, Sphat and Phat between models using a battery of two diagnostic tests and models using a battery of three diagnostic tests. However, the circumstances of this "step change" were uncertain—given that it was initially unclear how prior information could influence the accuracy and precision of Sehat, Sphat and Phat—and so the simplest assumption was used: that the power of BLCMs increased linearly with the number of diagnostic tests available. Based on this logic, the parameter `n.tests` was modelled as a continuous

			variable within LMM's to facilitate an understanding of when the number of diagnostic tests available is related to robust inferences of Se, Sp and P given a range of testing scenarios (i.e. batteries of 2, 3, 4 or 5 diagnostic tests). For the avoidance of doubt, the parameter `n.tests` was not modelled as a categorical variable since evaluating the large number of combinations of dependencies between the accuracies and precisions of Sehat, Sphat and Phat, and batteries of 2, 3, 4 and 5 diagnostic tests, given any other modelling conditions, was not a research aim.
The sample size of the population	`n.samples`	500, 1000, 1500	It was assumed that increases in sample size will decrease the variance of errors at any position in parameter space, and that it is necessary to investigate the effect of sample size in situations where gold standards are not available based on previous research such as (Rydevik, Innocent and McKendrick, 2018). Note, the chosen values (500, 1000, 1500) loosely reflect the suggestion from the Bacille Calmette–Guérin badger vaccination study—namely that a representative mean badger population size is ~671 (Byrne *et al.*, 2012), and that ~300 badgers are trapped and tested in the Woodchester Park study per year. Based on these guidelines, the sample sizes were considered to reflect plausible

			sizes of longitudinal test data collected in the field. Also based on these guidelines, in Chapter 8, sample sizes of 300 are used when validating the time decomposition BLCM prior to inputting the Woodchester Park test data.
Prior precision	`prior.sd`	0.05, 0.15	In ecology, the use of uniform priors appears the default choice (Banner, Irvine and Rodhouse, 2020), yet the consequences of this choice are rarely evaluated; in response, the power of BLCMs given two types of informative priors were considered and compared to the power of BLCMs given uniform priors. In models where priors are inputted as normal distributions, prior precision was specified as either being imprecise ($\sigma = 0.15$) or precise ($\sigma = 0.05$), where imprecise priors are weakly informative in comparison to precise priors, and precise priors do not fully inform the model, particularly given the size of each dataset (see Appendix 1: Simulated datasets Table 10-1). The values of sigma were chosen using a prior sensitivity analysis (see Figure 3-2).

Draw standard deviation	`Draw.sd`	0.05	The draw standard deviation defines the proximity of the mean of the prior to the true value. This value is kept constant throughout each experiment to enable comparisons to be made.
Number of simulations	`n.sim`	NA	The number of simulations defines the number times the model is validated with different true values. The value is design-dependent to provide the maximum number of simulations given the RAM and time available.
Number of cores	`n.cores`	NA	The number of cores defines the number of parallel computer cores over which a model is run, and its purpose is to maximise computational speed.
Seed	`seed`	NA	The value of the seed defines the initial state of any random number generation processes within the R script to ensure that any experiment can be replicated. The seeds used can be found within scripts on GitHub (https://github.com/annabush/PhD).
Limits of P	`pi.limit`	0 – 0.5	The upper and lower limits of values of P. The limits of P were considered to represent most possible real-world testing scenarios, under the assumption that most individuals, even within diseased populations, are healthy.

Limits of Se	`se.limit`	0 – 1	The upper and lower limits of values of Se. The decision to never constrain the limits of Se was made for two reasons. First, if both Se and Sp are constrained, there is a risk of identifiability issues due to the lack of reasonable solutions available to the MCMC algorithm. And constraining Sp was considered most important given the limits of P. Second, the battery of tests for bTB in badgers include Se values that range below 0.5 therefore to retain realistic modelling conditions, the limits of Se were never constrained.
Limits of Sp	`sp.limit`	0.5 – 1	The upper and lower limits of values of Sp. The limits of Sp were considered to represent tests that are better than chance alone—i.e., epidemiologically “useful” parameter space—based on the theory of an ROC curve. An ROC curve for a single diagnostic test is a plot of true positives versus false positives i.e., sensitivity versus 1 – specificity. The Area Under the Curve (AUC) of an ROC plot represents an uninformative test of no better than chance alone. ROC curves are typically used to explain the trade-off between Se and Sp (Hanley and McNeil, 1982). On an ROC curve, a coordinate of (0, 1) describes a gold standard test, and a coordinate of (0.5, 0.5) a random test of no discriminating ability above chance. Based on this, a perfect diagnostic classifier (such as a BLCM) would

		have Se and Sp values of 1, a random classifier would have Se and Sp values of 0.5 (Swets, 1988), and useful tests would generally have a Se and Sp of above 0.5.
Prior limits of P	`NA`	The prior limits of P are set by the lower and upper limit for true prevalence values.
Prior limits of Se	`prior.se`	The maximum and minimum values for the distribution describing Se, assigned using `se.limit`.
Prior limits of Sp	`prior.sp`	The maximum and minimum values for the distribution describing Sp, assigned using `sp.limit`.

Table 10-3: The MCMC hyperparameters used to define the JAGS models written using the jagsUI package (Kellner, 2015), their values, and why those values were chosen. These hyperparameters are relevant to the simulation analyses conducted between Chapters 5 to 7.

Hyperparameter	Definition	Values	Why chosen
`ni`	Number of iterations	1,000,000	A length of one million was chosen to ensure the required Effective Sample Size was met. The gelman.plot function of the coda package (Plummer et al., 2006) was used to visualise how the potential scale reduction factor changed throughout the MCMC chain to achieve convergence.
`nt`	Thinning interval	1	Thinning was not required to address autocorrelation. Instead, chain length was maximised. The autocorr.plot function of the coda package op cit. was used to visualise the effect of chain length maximisation.
`nb`	Burn-in interval	100,000	The burn-in interval was set to one-tenth of the number of iterations to ensure that the MCMC reached a reasonable posterior probability. The functions traceplot and gelman.plot of the coda package op cit. was used to visually decide the discard ratio.

nc	Number of chains	3	To increase reliability of any convergence achieved the default number of chains were used (Muma et al., 2007).
na	Adaptation period	NULL	The period before burn-in and sampling was not altered since the impact of na on MCMC is complex (Monnahan, Thorson and Branch, 2017) as it affects the proposal distribution. This parameter was set to NULL and JAGS was relied upon to tune the model automatically.

Table 10-4: A select list of key R functions created, their purposes, and how they are specified.

`get.outcome.matrix(tests)`

Parameters:

- `tests` (`integer`), the number of diagnostic tests

Returns:

A matrix denoting all possible combinations of diagnostic test outcomes given the number of diagnostic tests, where 0=negative and 1=positive.

```
get.outcome.matrix <- function(tests) {

  # Initialise the (2^tests x tests) outcomes matrix with NA
  outcomes <- matrix(data=NA, nrow=2^tests, ncol=tests)

  # Populate all possible outcomes as a factorial array
  # 0=negative, 1=positive
  For (i in 1:tests) {
    outcomes[, i] <- rep(
      rep(c(0, 1),
      each=nrow(outcomes) / (2^i)), 2^(i - 1)
    )
  }

  return (outcomes)
}
```

`get.test.results(pi, se, sp, n.tests seed=NULL)`

Parameters:

- `pi` (`float`), disease prevalence
- `se` (`vector of floats`), diagnostic test sensitivities
- sp (`vector of floats`), diagnostic tests specificities
- `n.tests` (`vector of integers`), the number of diagnostic tests to consider
- `seed` (`integer`), seed for random number generation

Returns:

A list of diagnostic test outcome summaries for each number of diagnostic tests, where each summary is a matrix with of length 2^{tests}.

```
get.test.results <- function(pi, se, sp, n.tests, seed=NULL){

  # Set seed for random number generation
  if (!is.null(seed)) {
    set.seed(seed)
  }

  # Set array to store resuts for each badger
  raw.data <- array(
    data=NA,
    dim=c(max(n.badgers), max(n.tests)),
    dimnames=list(badger=1:n.badgers, test=1:max(n.tests))
  )

  # Simulate status of all badgers
  status <- rbinom(n=max(n.badgers), size=1, p=pi)
```

```r
  # For each test
  for (test in 1:max(n.tests)){

    # Probability that each badger tests positive
    p <- status * se[test] + (1 - status) * (1 - sp[test])

    # Store test results in raw.data
    raw.data[, test] <- rbinom(max(n.badgers), 1, p)
  }

  # This will store a list of the test arrays
  results <- list()

  # For each number of tests
  for (tests in n.tests){

    # Make a test array: i.e. (000, 001, 010, 011, ..., 111)
    test.array <- array(
      0,
      dim=c(2 ** tests),
      dimnames=list(result=get.result.names(tests))
    )

    for (badger in 1:n.badgers){
      out <- paste(raw.data[badger, 1:tests], collapse="")
      test.array[out] <- test.array[out] + 1
    }

    results[[paste(tests, "tests")]] <- test.array
  }

  return(results)
}
```

`set.model(filepath)`

Writes the BLCM definition to file, note that this is an example, actual model definitions vary depending on experiment design.

Parameters:

- `file.path` (`string`), path to write the BLCM JAGS text file to

Returns: `NULL`

```
set.model <- function(filepath) {

  writeLines("model{
  for (i in 1:n.tests) {
    se[i] <- mu.se[i]
    sp[i] <- mu.sp[i]
  }
  pi <- mu.pi
  for (i in 1:n.outcomes) {
    for (j in 1:n.tests) {
      # A = se if badger is positive, 1 - se otherwise
      # B = 1 - sp if badger is positive, sp otherwise
      A[i, j] <- outcomes[i, j] * se[j] + (1 - outcomes[i, j]) * (1 - se[j])
      B[i, j] <- outcomes[i, j] * (1 - sp[j]) + (1 - outcomes[i, j]) * sp[j]
    }
    p[i] <- pi * prod(A[i, 1:n.tests]) + (1 - pi) * prod(B[i, 1:n.tests])
  }
  y[1:n.outcomes] ~ dmulti(p[1:n.outcomes], n)
  for (i in 1:n.tests) {
     mu.se[i] ~ dnorm(prior.se[i], precision) T(se.limit[1], se.limit[2])
     mu.sp[i] ~ dnorm(prior.sp[i], precision) T(se.limit[1], sp.limit[2])
  }
  mu.pi ~ dunif(pi.limit[1], pi.limit[2])
```

```
}", con=filepath)
}
```

run(sim, model.file)

Writes the BLCM definition to file, note that this is an example, actual model definitions vary depending on experiment design.

Parameters:

- `data` (`list`), true P, SE, and SP values to use to in the simulation
- `model.file` (`string`), path to JAGS BLCM text file

Returns: `NULL`

```
run <- function(model.file, data){

  # Get simulated badger test results
  test.results <- get.test.results(
    pi=data$pi,
    se=data$se,
    sp=data$sp,
    seed=data$seed
  )

  # Make array to store results from a single simulation
  sim.results <- array(
    data=NA,
    dim=c(
      length(prior.sd),
      length(n.badgers),
      length(n.tests),
      4,
```

```
    2 * max(n.tests) + 1
  ),
  dimnames=list(
    prior.sd=prior.sd,
    n.badgers=n.badgers,
    n.tests=n.tests,
    statistic=c("true", "mean", "sd", "error"),
    param=get.names(max(n.tests))$all
  )
)

# Run for each prior sd, number of tests and number of badgers
for (p in 1:length(prior.sd)){
  for (b in 1:length(n.badgers)){
    for (t in 1:length(n.tests)){

      names <- get.names(n.tests[t])

      bugs.data <- list(
        y=test.results[[t]][b, ],
        n=n.badgers[b],
        n.tests=n.tests[t],
        outcomes=get.outcome.matrix(n.tests[t]),
        n.outcomes=2**n.tests[t],
        se.limit=se.prior.limit,
        sp.limit=sp.prior.limit,
        pi.limit=pi.prior.limit,
        precision=1 / prior.sd[p] ^ 2,
        prior.se=values$mean[data$sim, names$se],
        prior.sp=values$mean[data$sim, names$sp]
      )

      output <- jags(
        data=bugs.data,
        inits=NULL,
```

```
      model.file=model.file,
      parameters.to.save=c("pi", "se", "sp"),
      n.adapt=na,
      n.chains=nc,
      n.thin=nt,
      n.iter=ni,
      n.burnin=nb,
      store.data=TRUE
    )

    # Store some of the outputs
    sim.results[p, b, t, "true", "pi"] <- values$true[data$sim,"pi"]
    sim.results[p, b, t, "true", names$se] <- values$true[data$sim, names$se]
    sim.results[p, b, t, "true", names$sp] <- values$true[data$sim, names$sp]
    sim.results[p, b, t, "mean", "pi"] <- output$mean$pi
    sim.results[p, b, t, "mean", names$se] <- output$mean$se
    sim.results[p, b, t, "mean", names$sp] <- output$mean$sp
    sim.results[p, b, t, "sd", "pi"] <- output$sd$pi
    sim.results[p, b, t, "sd", names$se] <- output$sd$se
    sim.results[p, b, t, "sd", namessp] <- output$sd$sp

    true.pi <- sim.results[p, b, t, "true", "pi"]
    true.se <- sim.results[p, b, t, "true", names$se]
    true.sp <- sim.results[p, b, t, "true", names$sp]

    pred.pi <- sim.results[p, b, t, "mean", "pi"]
    pred.se <- sim.results[p, b, t, "mean", names$se]
    pred.sp <- sim.results[p, b, t, "mean", names$sp]

    sim.results[p, b, t, "error", "pi"] <- true.pi - pred.pi
    sim.results[p, b, t, "error", names$se] <- true.se - pred.se
    sim.results[p, b, t, "error", names$sp] <- true.sp - pred.sp
  }
 }
}
```

```
  return (sim.results)
}
```

Appendix 3: Directory of Linear Mixed Effects Models

Table 10-5 to Table 10-12 summarise the LMMs used in Chapter 4, and are formatted as follows:

1. SE = standard error
2. SD = standard deviation
3. DOF = Degrees of Freedom
4. NEGL indicates where the percentage variance is negligible (i.e., lower than 0.1%).
5. NA indicates where between-group variance is insufficient i.e., very close to 0.
6. Asterisks denotes statistically significant effects in terms of p-values generated from a t-test using Satterthwaite's method where * = $p < 0.01$, ** = $p < 0.001$, and *** = $p < 0.0001$
7. Within the AONVA analyses, models were fitted using maximum likelihood methods and the Chi-squared approximation.
8. The variances reported are rounded to one decimal place, and therefore may suffer from rounding error.

Table 10-5: Outputs for Chapter 4 LMM's 1 to 3.

	LMM 1 **(errors of Phat)**	**LMM 2** **(errors of Sehat)**	**LMM 3** **(errors of Sphat)**
Fixed regression coefficients			
(SE)	1.27e-01 (2.36e-02)*	6.25e-02 (4.55e-03)***	1.57e-02 (2.89e-03)***
Intercept	-2.05e-02 (1.69e-03)***	-9.44e-03 (7.31e-	1.69e-04 (6.43e-04)
n.tests		04)***	
Random effects variances			
(SD)	8.35e-13 (9.14e-07)	2.07e-05 (4.55e-03)	NA
pi.rel intercept	3.11e-05 (5.57e-03)	1.51e-05 (3.89e-03)	6.81e-06 (2.61e-03)
se.rel intercept	1.51e-03 (3.89e-02)	1.98e-07 (4.45e-03)	3.52e-07 (5.94e-04)
sp.rel intercept	1.00e-03 (3.17e-02)	1.87e-04 (1.37e-02)	1.45e-04 (1.20e-02)
Residual			
Number of data points †	280	280	280

† a single data point is equal to one error value

Table 10-6: ANOVA outputs showing the relationship between the number of diagnostic tests available and error given Chapter 4 LMM's 1 to 3. Full models contain "n.tests" as a fixed parameter, and null models only have a fixed intercept.

	Nested Model	DOF	AIC	Log-likelihood	deviance	χ2 value	p-value
Pi errors	Full	6	-1112.8	562.39	-1124.8	117.58	<0.0001
	Null	5	-997.2	503.6	-1007.2		
Se1 errors	Full	6	-1588.9	800.46	-1600.9	130.69	<0.0001
	Null	5	-1460.2	735.11	-1470.2		
Sp1 errors	Full	6	-1666.5	839.26	-1678.5	0.0692	0.7924
	Null	5	-1668.5	839.23	-1678.5		

Table 10-7: The proportion of total variance explained by each random effect, and the residual effects, expressed as a percentage, for LMM's 1 to 5.

Random effects	**% variance of total random effects**				
	LMM 1 (errors of Phat)	LMM 2 (errors of Se1hat)	LMM 3 (errors of Sp1hat)	LMM 4 (errors of Se2hat)	LMM 5 (errors of Sp2hat)
pi.rel intercept	NEGL	9.3	NA	10.3	3.2
se.rel intercept	1.2	6.8	4.5	12.0	NEGL
sp.rel intercept	59.3	0.1	0.2	11.7	3.5
Residual	39.4	83.8	95.3	66.1	93.3

Table 10-8: Outputs for Chapter 4 LMM's 6 to 8 which investigate the influence of prior precision on error.

	LMM 6 **(errors of Phat)**	**LMM 7** **(errors of Sehat)**	**LMM 8** **(errors of Sphat)**
Fixed regression coefficients			
(SE)	1.03e-01 (1.08e-02)***	4.76e-02 (6.63e-03)***	5.52e-02 (7.49e-03)***
Intercept	-1.16e-02 (2.54e-03)***	-3.36e-03 (1.40e-03)*	-3.18e-03 (1.41e-03)*
n.tests	-3.27e-02 (5.68e-03)***	-4.90e-03 (3.13e-03)	-1.43e-02 (3.16e-
prior.infoprecise			03)***
Random effects variances (SD)			
pi intercept	1.15e-06 (1.07e-03)	2.39e-06 (1.54e-03)	8.97e-05 (9.47e-03)
se1 intercept	3.30e-04 (1.82e-02)	1.57e-04 (1.25e-02)	1.75e-04 (1.32e-02)
sp1 intercept	2.26e-04 (1.50e-02)	2.17e-04 (1.47e-02)	4.04e-04 (2.01e-02)
Residual	1.61e-03 (4.01e-02)	4.89e-04 (2.21e-02)	4.98e-04 (2.23e-02)
Number of data points †	200	200	200

† A single data point is equal to one simulation.

Table 10-9: Outputs for Chapter 4 LMM's 9 to 11 which test the influence of constraint on error.

	LMM 9 **(errors of Phat)**	**LMM 10** **(errors of Sehat)**	**LMM 11** **(errors of Sphat)**
Fixed regression coefficients (SE)			
Intercept	9.67e-02 (9.19e-03)***	3.40e-02 (5.69e-03)***	4.82e-02 (5.32e-03)***
n.tests	-1.21e-02 (1.53e-03)***	-2.79e-03 (7.64e-04)***	-1.87e-03 (7.82e-04)*
con.infosp	7.19e-03 (1.06e-02)	1.14e-02 (5.76e-03)	-9.57e-03 (4.70e-03)*
coninfosppi	4.97e-03 (1.06e-02)	1.81e-02 (5.76e-03)**	-1.65e-02 (4.70e-03)**
Random effects variances (SD)			
pi intercept	6.67e-05 (8.17e-03)	5.09e-05 (7.13e-03)	1.63e-04 (1.28e-02)
se1 intercept	2.71e-11 (5.20e-06)	2.16e-04 (1.47e-02)	2.45e-04 (1.57e-02)
sp1 intercept	1.11e-03 (3.34e-02)	3.09e-04 (1.76e-02)	5.62e-05 (7.50e-03)
Residual	3.75e-03 (4.18e-02)	4.38e-04 (2.09e-02)	4.58e-04 (2.14e-02)
Number of data points †	600	600	600

† A single data point is equal to one simulation.

Table 10-10: ANOVA outputs showing the relationship between error and prior information (constraint and prior precision) for Chapter 4 LMM's 6 to 11. Full models contain the number of diagnostic tests available as a fixed parameter, and null models drop either prior precision or constraint, where indicated, as a fixed effect.

		Nested Model	DOF	AIC	Log-likelihood	deviance	χ2 value	p-value
Prior precision	P errors	Full	7	-674.58	344.29	-688.58	30.763	<0.0001
		Null	6	-645.81	328.91	-657.81		
	Se1 errors	Full	7	-896.75	455.37	-910.75	2.4526	0.1173
		Null	6	-896.30	454.15	-908.30		
	Sp1 errors	Full	7	-880.71	447.35	-894.71	19.65	<0.0001
		Null	6	-863.06	437.53	-875.06		
Constraint	P errors	Full	8	-1981.2	998.6	-1997.2	0.5854	0.7462
		Null	6	-1984.6	998.3	-1996.6		
	Se1 errors	Full	8	-2775.1	1395.5	-2791.1	11.428	0.0033
		Null	6	-2767.7	1389.8	-2779.7		

Sp1 errors	Full	8	-2750.9	1383.5	-2766.9	11.19	0.003717
	Null	6	-2743.8	1377.9	-2755.8		

Table 10-11: ANOVA outputs showing the relationship between error and the number of diagnostic tests available based on LMM's 6 to 11. Full models contain "n.tests" as a fixed effect, and null models drop "n.tests" as a fixed effect.

		Nested Model	DOF	AIC	Log-likelihood	deviance	χ2 value	p-value
Prior precision	P errors	Full	6	-645.81	328.91	-657.81	16.91	<0.0001
		Null	5	-630.9	320.45	-640.90		
	Se1 errors	Full	6	-896.3	454.15	-908.3	5.6497	0.01746
		Null	5	-892.65	451.32	-902.65		
	Sp1 errors	Full	6	-863.06	437.53	-875.06	4.5144	0.03361
		Null	5	-860.55	435.27	-870.55		
Constraint	P errors	Full	6	-1984.6	998.3	-1996.6	59.739	<0.0001
		Null	5	-1926.9	968.44	-1936.9		
	Se1 errors	Full	6	-2767.7	1389.8	-2779.7	13.219	0.0002772
		Null	5	-2756.4	1383.2	-2766.4		
	Sp1 errors	Full	6	-2743.8	1377.9	-2755.8	5.6897	0.01707
		Null	5	-2740.1	1375.0	-2750.1		

Table 10-12: The proportion of total variance explained by each random effect and the residual effects, expressed as a percentage, for LMM's 6 to 11.

Random effects	**% variance of total random effects**					
	Prior Precision			**Constraint**		
	LMM 6 (errors of Phat)	**LMM 7** (errors of Sehat)	**LMM 8** (errors of Sphat)	**LMM 9** (errors of Phat)	**LMM 10** (errors of Sehat)	**LM11** (errors of Sphat)
pi intercept	0.1	0.3	7.7	2.3	5.0	17.7
se1 intercept	15.2	18.1	15.0	NEGL	21.3	26.6
sp1 intercept	10.4	25.1	34.6	38.0	30.5	6.1
Residual	74.3	56.6	42.7	59.7	43.2	49.7

Table 10-13: A summary of the LMM's used in Chapter 5. LMM's 13 to 24 and 34 to 42 inclusive belong to the 15% scenario.

LMM	Response			Subset							Fixed						
	Bias	Error	Standard deviation	Normal	Precise	Uniform	Prior Distribution	Se1	Sp1	P	Prior Precision	Constraint	Number of samples	Number of tests	Extreme	Extreme * tests	Prior Distribution
1	●			●							2e-03	-6e-03	-1e-06	5e-03	4e-02	-2e-03	
2		●		●							-1e-02	3e-03	-6e-06	-9e-03	-7e-03	-1e-04	
3			●	●							-2e-02	6e-03	-9e-06	-2e-02	5e-04	1e-05	
4	●				●							-9e-03	-2e-06	6e-03	4e-02	-2e-03	
5		●			●							5e-03	-6e-06	-1e-02	-4e-03	-9e-04	
6			●		●							1e-02	-1e-05	-2e-02	3e-03	-7e-04	
7	●					●						1e-02	-2e-07	-2e-02	-3e-02	2e-02	
8		●				●						2e-01	-1e-06	-2e-02	-8e-02	2e-02	
9			●			●						1e-01	-2e-05	-3e-02	-2e-02	5e-03	
10	●						●										5e-03
11		●					●										1e-01
12			●				●										8e-02
13	●			●							-3e-02	6e-03	-4e-03	-5e-07			
14		●		●							-1e-02	5e-03	-5e-06	-1e-02			
15			●	●							--3e-02	5e-03	-9e-06	-2e-02			
16	●				●							-7e-03	-1e-06	-1e-02			
17		●			●							1e-02	-5e-06	-1e-02			
18			●		●							1e-02	-1e-05	-2e-02			

19	●					●						-4e-03	-5e-07	-7e-03			
20		●				●						5e-03	-5e-06	-1e-02			
21			●			●						5e-03	-9e-06	-2e-02			
22	●						●										-2e-02
23		●					●										1e-01
24			●				●										9e-02
25	●							●			-5e-03	2e-03	-1e-06	2e-03	7e-02	-6e-03	
26		●						●			-1e-02	1e-03	-6e-06	-5e-03	-1e-02	-7e-04	
27			●					●			-3e-02	3e-04	-1e-05	-1e-02	1e-03	-4e-04	
28	●								●		-1e-03	-1e-03	-1e-06	3e-03	5e-02	-4e-03	
29		●							●		-3e-03	6e-04	-6e-06	-3e-03	-1e-02	3e-04	
30			●						●		-1e-02	3e-03	-8e-06	-8e-03	2e-03	-4e-04	
31	●									●	1e-02	-2e-02	-2e-06	9e-03	8e-03	2e-03	
32		●								●	-1e-02	7e-03	-6e-06	-2e-02	3e-03	2e-05	
33			●							●	-3e-02	1e-02	-1e-05	-3e-03	-2e-03	8e-04	
34	●							●			-1e-03	2e-04	-2e-06	-4e-03			
35		●						●			-2e-02	2e-03	-6e-06	-6e-03			
36			●					●			-4e-02	7e-04	-1e-05	-1e-02			
37	●								●		3e-04	5e-04	-1e-05	-1e-02			
38		●							●		-2e-03	1e-03	-5e-06	-2e-03			
39			●						●		-1e-02	3e-03	-7e-06	-7e-03			
40	●									●	3e-02	-1e-02	-1e-06	-2e-02			
41		●								●	-2e-02	1e-02	-4e-06	-2e-02			
42			●							●	-4e-02	1e-02	-9e-06	-3e-02			

Table 10-14: A summary of the LMM's used in Chapter 6.

LMM	Response		Subset								Fixed						
	Logit error	Logit absolute error	Global	Se1	Sp1	P	Normal	Normal and Imprecise	Uniform	No other subset	Prior Precision	Constraint	Number of samples	Number of tests	Extreme	Extreme * tests	Prior Distribution
1a	●		●				●				-2e-02	3e-02	9e-06	-3e-02	-2e-01	1e-02	
1b	●					●	●				-1e-01	9e-02	1e-06	-8e-02	5e-02	-2e-02	
1c	●			●			●				3e-02	-1e-02	1e-05	-5e-03	-4e-01	3e-01	
1d	●				●		●				1e-02	6e-03	1e-05	-1e-02	-3e-01	2e-02	
1e		●	●				●				-7e-02	2e-02	-4e-05	-5e-02	-2e-02	-6e-03	
1f		●				●	●				-9e-02	4e-02	-4e-05	-1e-01	5e-02	-7e-03	
1g		●		●			●				-9e-02	8e-03	-4e-05	-3e-02	-3e-02	-8e-03	
1h		●			●		●				-2e-02	4e-03	-5e-05	-2e-02	-7e-02	-2e-03	
2a	●		●					●				5e-02	1e-05	-3e-02	-2e-01	5e-03	
2b	●					●		●				2e-01	7e-06	-1e-01	1e-01	-4e-02	
2c	●			●				●				-2e-02	1e-05	6e-03	-4e-01	3e-03	
2d	●				●			●				7e-03	1e-05	-3e-03	-3e-01	2e-02	
2e		●	●					●				3e-02	-4e-05	-6e-02	2e-02	-1e-01	
2f		●				●		●				6e-02	-3e-05	-1e-01	7e-02	-1e-02	
2g		●		●				●				1e-02	-5e-05	-4e-02	2e-02	-2e-02	
2h		●			●			●				5e-03	-4e-05	-2e-02	-4e-02	-1e-02	
3a	●		●						●			-9e-02	-2e-06	8e-02	1e-01	-1e-01	
3b	●					●			●			1e+00	4e-05	-2e-02	-7e-02	3e-02	
3c	●			●					●			-8e-01	-3e-05	9e-02	1e-01	-1e-01	

3d	●				●				●			-8e-01	-1e-05	2e-01	4e-01	-2e-01	
3e		●	●						●			1e+00	2e-05	-6e-02	-1e-01	5e-02	
3f		●				●			●			1e+00	2e-05	-6e-02	-1e-01	5e-02	
3g		●		●					●			4e-01	-2e-05	9e-05	1e-01	-3e-02	
3h		●			●				●			1e+00	2e-05	-6e-02	-1e-01	5e-02	
4a	●		●							●							-2e-02
4b	●					●				●							7e-01
4c	●			●						●							-4e-01
4d	●				●					●							-4e-01
4e		●	●							●							7e-01
4f		●				●				●							7e-01
4g		●		●						●							7e-01
4h		●			●					●							8e-01

11. Bibliography

Agresti, A. (2003) *An introduction to categorical data analysis, second edition*. John Wiley & Sons. doi: 10.1002/0470114754.

Akobeng, A. K. (2007) 'Understanding diagnostic tests 3: Receiver operating characteristic curves', *Acta Paediatrica, International Journal of Paediatrics*, pp. 644–647. doi: 10.1111/j.1651-2227.2006.00178.x.

Albert, P. S. and Dodd, L. E. (2004) 'A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard', *Biometrics*, 60(2), pp. 427–435. doi: 10.1111/j.0006-341X.2004.00187.x.

Albert, P. S. and Dodd, L. E. (2008) 'On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation', *Journal of the American Statistical Association*, 103(481), pp. 61–73. doi: 10.1198/016214507000000329.

Allaby, M. (2015) *A Dictionary of Ecology*, *A Dictionary of Ecology*. Oxford University Press. doi: 10.1093/acref/9780191793158.001.0001.

Allen, A. R., Skuce, R. A. and Byrne, A. W. (2018) 'Bovine tuberculosis in Britain and Ireland - A perfect storm? The confluence of potential ecological and epidemiological impediments to controlling a chronic infectious disease', *Frontiers in Veterinary Science*, 5(JUN). doi: 10.3389/fvets.2018.00109.

Allen, A., Skuce, R. and McDowell, S. (2011) *Bovine TB: a review of badger-to-cattle transmission*, *Agri-food and Biosciences Institute*. Available at: https://www.bovinetb.info/docs/bovine-tb-a-review-of-badger-to-cattle-transmission.pdf.

Allen, T. F. H. and Starr, T. B. (2017) *Hierarchy: Perspectives for Ecological Complexity*. University of Chicago Press. doi: https://doi.org/10.7208/chicago/9780226489711.001.0001.

Amrhein, V., Greenland, S. and McShane, B. (2019) 'Scientists rise up against statistical significance', *Nature*, 567(7748), pp. 305–307. doi: 10.1038/d41586-019-00857-9.

van Andel, M. *et al.* (2020) 'Estimating foot-and-mouth disease (FMD) prevalence in central Myanmar: Comparison of village headman and farmer disease reports with serological findings', *Transboundary and Emerging Diseases*, 67(2), pp. 778–791. doi: 10.1111/tbed.13397.

Andersen, K. G. *et al.* (2020) 'The proximal origin of SARS-CoV-2', *Nature Medicine*, 26(4), pp. 450–452. doi: 10.1038/s41591-020-0820-9.

Andersen, R., Hagenaars, J. A. and McCutcheon, A. L. (2003) 'Applied Latent Class Analysis', *Canadian Journal of Sociology / Cahiers canadiens de sociologie*, 28(4), p. 584. doi: 10.2307/3341848.

Arango-Sabogal, J. C. *et al.* (2019) 'Accuracy of leukocyte esterase test, endometrial cytology and vaginal discharge score for diagnosing postpartum reproductive tract health status in dairy cows at the moment of sampling, using a latent class model fit within a Bayesian framework', *Preventive Veterinary Medicine*, 162, pp. 1–10. doi: 10.1016/j.prevetmed.2018.11.003.

Arnold, M. E. *et al.* (2021) 'A Bayesian analysis of a test and vaccinate or remove study to control bovine tuberculosis in badgers (Meles meles)', *PLoS ONE*, 16(1 January), p. e0246141. doi: 10.1371/journal.pone.0246141.

Auger-Méthé, M. *et al.* (2016) 'State-space models' dirty little secrets: Even

simple linear Gaussian models can have estimation problems', *Scientific Reports*, 6(1), pp. 1–10. doi: 10.1038/srep26677.

Augusiak, J., Van den Brink, P. J. and Grimm, V. (2014) 'Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach', *Ecological Modelling*, 280, pp. 117–128. doi: 10.1016/j.ecolmodel.2013.11.009.

Bachmann, L. M. *et al.* (2005) 'Consequences of different diagnostic "gold standards" in test accuracy research: Carpal Tunnel Syndrome as an example', *International Journal of Epidemiology*, 34(4), pp. 953–955. doi: 10.1093/IJE/DYI105.

Baker, L. *et al.* (2020) 'Local rabies transmission and regional spatial coupling in European foxes', *PLoS ONE*, 15(5). doi: 10.1371/journal.pone.0220592.

Banner, K. M., Irvine, K. M. and Rodhouse, T. J. (2020) 'The use of Bayesian priors in Ecology: The good, the bad and the not great', *Methods in Ecology and Evolution*, 11(8), pp. 882–889. doi: 10.1111/2041-210X.13407.

Barr, D. J. *et al.* (2013) 'Random effects structure for confirmatory hypothesis testing: Keep it maximal', *Journal of Memory and Language*, 68(3), pp. 255–278. doi: 10.1016/j.jml.2012.11.001.

Barreto, M. L., Teixeira, M. G. and Carmo, E. H. (2006) 'Infectious diseases epidemiology', *Journal of Epidemiology and Community Health*, 60(3), pp. 192–195. doi: 10.1136/jech.2003.011593.

Barroso, P., Acevedo, P. and Vicente, J. (2021) 'The importance of long-term studies on wildlife diseases and their interfaces with humans and domestic animals: A review', *Transboundary and Emerging Diseases*, 68(4), pp. 1895–

1909. doi: 10.1111/tbed.13916.

Bates, D. *et al.* (2015) ‘Fitting linear mixed-effects models using lme4’, *Journal of Statistical Software*, 67(1). doi: 10.18637/jss.v067.i01.

Bayes (1763) ‘LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S’, *Philosophical Transactions of the Royal Society of London*, 53, pp. 370–418. doi: 10.1098/rstl.1763.0053.

Becker, D. J. *et al.* (2019) ‘The problem of scale in the prediction and management of pathogen spillover’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1782). doi: 10.1098/rstb.2019.0224.

Begg, C. B. (1987) ‘Biases in the assessment of diagnostic tests’, *Statistics in Medicine*, 6(4), pp. 411–423. doi: 10.1002/sim.4780060402.

Beguin, J. *et al.* (2012) ‘Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation’, *Methods in Ecology and Evolution*, 3(5), pp. 921–929. doi: 10.1111/j.2041-210X.2012.00211.x.

Bell, B. A., Ferron, J. M. and Kromrey, J. D. (2008) ‘Cluster Size in Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models’, *JSM Proceedings*, pp. 1122–1129. Available at: http://www.amstat.org/sections/srms/proceedings/y2008/Files/300933.pdf (Accessed: 28 March 2023).

Bell, D. M. *et al.* (2018) ‘Visual interpretation and time series modeling of Landsat imagery highlight drought’s role in forest canopy declines’, *Ecosphere*, 9(6). doi: 10.1002/ecs2.2195.

Belmont, J. *et al.* (2022) ‘A new statistical approach for identifying rare species under imperfect detection’, *Diversity and Distributions*, 28(5), pp. 882–893. doi: 10.1111/ddi.13495.

Benavides, J. A. *et al.* (2017) ‘Estimating Loss of Brucella Abortus Antibodies from Age-Specific Serological Data In Elk’, *EcoHealth*, 14(2), pp. 234–243. doi: 10.1007/s10393-017-1235-z.

Benjamin-Fink, N. and Reilly, B. K. (2017) ‘A road map for developing and applying object-oriented bayesian networks to “WICKED” problems’, *Ecological Modelling*, 360, pp. 27–44. doi: 10.1016/j.ecolmodel.2017.06.028.

Bennett, A. J. *et al.* (2020) ‘Relatives of rubella virus in diverse mammals’, *Nature*, 586(7829), pp. 424–428. doi: 10.1038/s41586-020-2812-9.

Bennington, C. C. and Thayne, W. V. (1994) ‘Use and misuse of mixed model analysis of variance in ecological studies’, *Ecology*, 75(3), pp. 717–722. doi: 10.2307/1941729.

Benton, C. H. *et al.* (2018) ‘Inbreeding intensifies sex- and age-dependent disease in a wild mammal’, *Journal of Animal Ecology*, 87(6), pp. 1500–1511. doi: 10.1111/1365-2656.12878.

Berkvens, D. *et al.* (2006) ‘Estimating disease prevalence in a Bayesian framework using probabilistic constraints’, *Epidemiology*, 17(2), pp. 145–153. doi: 10.1097/01.ede.0000198422.64801.8d.

Berlow, E. L. *et al.* (2009) ‘Simple prediction of interaction strengths in complex food webs’, *Proceedings of the National Academy of Sciences of the United States of America*, 106(1), pp. 187–191. doi: 10.1073/pnas.0806823106.

Bermingham, M. L. *et al.* (2015) ‘Hui and Walter’s latent-class model extended

to estimate diagnostic test properties from surveillance data: A latent model for latent data', *Scientific Reports*, 5(1), pp. 1–14. doi: 10.1038/srep11861.

Beyer, H. L. *et al.* (2013) 'The effectiveness of Bayesian state-space models for estimating behavioural states from movement paths', *Methods in Ecology and Evolution*, 4(5), pp. 433–441. doi: 10.1111/2041-210X.12026.

Böhm, M., Hutchings, M. R. and White, P. C. L. (2009) 'Contact networks in a wildlife-livestock host community: Identifying high-risk individuals in the transmission of bovine TB among badgers and cattle', *PLoS ONE*, 4(4), p. e5016. doi: 10.1371/journal.pone.0005016.

Bolker, B. (2020) *GLMM FAQ*. Available at: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

Bolker, B. M. *et al.* (2009) 'Generalized linear mixed models: a practical guide for ecology and evolution', *Trends in Ecology and Evolution*, pp. 127–135. doi: 10.1016/j.tree.2008.10.008.

Bordier, M. *et al.* (2020) 'Characteristics of One Health surveillance systems: A systematic literature review', *Preventive veterinary medicine*, 181. doi: 10.1016/J.PREVETMED.2018.10.005.

Branscum, A. J., Gardner, I. A. and Johnson, W. O. (2005) 'Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling', *Preventive Veterinary Medicine*, 68(2–4), pp. 145–163. doi: 10.1016/j.prevetmed.2004.12.005.

Brennan, A. *et al.* (2017) 'Shifting brucellosis risk in livestock coincides with spreading seroprevalence in elk', *PLoS ONE*, 12(6), p. e0178780. doi: 10.1371/journal.pone.0178780.

Brenner, H. and Gefeller, O. (1997) ‘Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence’, *Statistics in Medicine*, 16(9), pp. 981–991. doi: 10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N.

von Brömssen, C. *et al.* (2018) ‘Statistical models for evaluating suspected artefacts in long-term environmental monitoring data’, *Environmental Monitoring and Assessment*, 190(9). doi: 10.1007/s10661-018-6900-3.

de Bronsvoort, B. M. C. *et al.* (2019) ‘Comparison of Two Rift Valley Fever Serological Tests in Cameroonian Cattle Populations Using a Bayesian Latent Class Approach’, *Frontiers in Veterinary Science*, 6, p. 258. doi: 10.3389/fvets.2019.00258.

Brown, E. R. (2010) ‘Bayesian Estimation of the Time-Varying Sensitivity of a Diagnostic Test with Application to Mother-to-Child Transmission of HIV’, *Biometrics*, 66(4), pp. 1266–1274. doi: 10.1111/j.1541-0420.2010.01398.x.

Bujang, M. A. and Adnan, T. H. (2016) ‘Requirements for minimum sample size for sensitivity and specificity analysis’, *Journal of Clinical and Diagnostic Research*, pp. YE01–YE06. doi: 10.7860/JCDR/2016/18129.8744.

Buzdugan, S. N. *et al.* (2017) ‘Inference of the infection status of individuals using longitudinal testing data from cryptic populations: Towards a probabilistic approach to diagnosis’, *Scientific Reports*, 7(1), pp. 1–11. doi: 10.1038/s41598-017-00806-4.

Byrne, A. W. *et al.* (2012) ‘Population Estimation and Trappability of the European Badger (Meles meles): Implications for Tuberculosis Management’, *PLoS ONE*, 7(12). doi: 10.1371/journal.pone.0050807.

Calenge, C. *et al.* (2021) ‘Estimating disease prevalence and temporal dynamics using biased capture serological data in a wildlife reservoir: The example of brucellosis in Alpine ibex (Capra ibex)’, *Preventive Veterinary Medicine*, 187, p. 105239. doi: 10.1016/j.prevetmed.2020.105239.

Cariboni, J. *et al.* (2007) ‘The role of sensitivity analysis in ecological modelling’, *Ecological Modelling*, 203(1–2), pp. 167–182. doi: 10.1016/j.ecolmodel.2005.10.045.

Carlson, C. J. *et al.* (2021) ‘The future of zoonotic risk prediction’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1837). doi: 10.1098/RSTB.2020.0358.

Carter, S. P. *et al.* (2007) ‘Culling-induced social perturbation in Eurasian badgers Meles meles and the management of TB in cattle: An analysis of a critical problem in applied ecology’, *Proceedings of the Royal Society B: Biological Sciences*, 274(1626), pp. 2769–2777. doi: 10.1098/rspb.2007.0998.

Carter, S. P. *et al.* (2012) ‘BCG Vaccination Reduces Risk of Tuberculosis Infection in Vaccinated Badgers and Unvaccinated Badger Cubs’, *PLoS ONE*. Edited by J. L. Herrmann, 7(12), p. e49833. doi: 10.1371/journal.pone.0049833.

Casaubon, J. *et al.* (2012) ‘Bovine viral diarrhea virus in free-ranging wild ruminants in Switzerland: low prevalence of infection despite regular interactions with domestic livestock’, *BMC Veterinary Research*, 8. doi: 10.1186/1746-6148-8-204.

Celeux, G. (1998) ‘Bayesian Inference for Mixture: The Label Switching Problem’, *Compstat*, pp. 227–232. doi: 10.1007/978-3-662-01131-7_26.

Chalom, A. and de Prado, P. I. de K. L. (2012) ‘Parameter space exploration of

ecological models', *arXiv:1210.6278*. Available at: http://arxiv.org/abs/1210.6278 (Accessed: 20 March 2022).

Chambers, M. A. *et al.* (2008) 'Validation of the BrockTB Stat-Pak assay for detection of tuberculosis in Eurasian badgers (Meles meles) and influence of disease severity on diagnostic accuracy', *Journal of Clinical Microbiology*, 46(4), pp. 1498–1500. doi: 10.1128/JCM.02117-07.

Chen, W. C. *et al.* (2020) 'Non-parametric Bayesian density estimation for biological sequence space with applications to pre-mRNA splicing and the karyotypic diversity of human cancer', *bioRxiv*, pp. 1–18. doi: 10.1101/2020.11.25.399253.

Cheng, T. L. *et al.* (2021) 'The scope and severity of white-nose syndrome on hibernating bats in North America', *Conservation Biology*, 35(5), pp. 1586–1597. doi: 10.1111/cobi.13739.

Chivers, C., Leung, B. and Yan, N. D. (2014) 'Validation and calibration of probabilistic predictions in ecology', *Methods in Ecology and Evolution*, 5(10), pp. 1023–1032. doi: 10.1111/2041-210X.12238.

Clark, J. S. (2005) 'Why environmental scientists are becoming Bayesians', *Ecology Letters*, 8(1), pp. 2–14. doi: 10.1111/j.1461-0248.2004.00702.x.

Clark, J. S. and Gelfand, A. E. (2006) 'A future for models and data in environmental science', *Trends in Ecology and Evolution*, pp. 375–380. doi: 10.1016/j.tree.2006.03.016.

Cleasby, I. R. and Nakagawa, S. (2011) 'Neglected biological patterns in the residuals', *Behavioral Ecology and Sociobiology*, 65(12), pp. 2361–2372. doi: 10.1007/s00265-011-1254-7.

Clifton-Hadley, R. S., Wilesmith, J. W. and Stuart, F. A. (1993) ‘Mycobacterium bovis in the European badger (Meles meles): Epidemiological findings in tuberculous badgers from a naturally infected population’, *Epidemiology and Infection*, 111(1), pp. 9–19. doi: 10.1017/S0950268800056624.

Cochran, W. G. (1977) *Sampling Techniques third edition*. John Wiley & Sons.

Collard, K. J. (2023) ‘A study of the incidence of bovine tuberculosis in the wild red deer herd of Exmoor’, *European Journal of Wildlife Research*, 69(1), pp. 1–8. doi: 10.1007/s10344-022-01638-y.

Collins, J. and Huynh, M. (2014) ‘Estimation of diagnostic test accuracy without full verification: A review of latent class methods’, *Statistics in Medicine*, 33(24), pp. 4141–4169. doi: 10.1002/sim.6218.

Conn, P. B. *et al.* (2018) ‘A guide to Bayesian model checking for ecologists’, *Ecological Monographs*, 88(4), pp. 526–542. doi: 10.1002/ecm.1314.

Cordes, C. L. (1980) ‘Adaptive environmental assessment and management: an overview.’, *Proc. Gulf of Mexico coastal ecosystem workshop, Sept 1979, Port Aransas Texas*, pp. 185–189. Available at: https://keep.lib.asu.edu/items/149155 (Accessed: 17 March 2023).

Craft, M. E. *et al.* (2008) ‘Dynamics of a multihost pathogen in a carnivore community’, *Journal of Animal Ecology*, 77(6), pp. 1257–1264. doi: 10.1111/j.1365-2656.2008.01410.x.

Crawley, M. J. (2012) ‘The R Book: Second Edition’, *The R Book: Second Edition*, pp. 1–1051. doi: 10.1002/9781118448908.

Crispell, J. *et al.* (2017) ‘Using whole genome sequencing to investigate transmission in a multi-host system: Bovine tuberculosis in New Zealand’, *BMC*

Genomics, 18(1), pp. 1–12. doi: 10.1186/s12864-017-3569-x.

Crispell, J. *et al.* (2019) 'Combining genomics and epidemiology to analyse bi-directional transmission of mycobacterium bovis in a multi-host system', *eLife*, 8. doi: 10.7554/eLife.45833.

Cross, P. C. *et al.* (2010) 'Mapping brucellosis increases relative to elk density using hierarchical bayesian models', *PLoS ONE*, 5(4), p. e10322. doi: 10.1371/journal.pone.0010322.

Cross, P. C. *et al.* (2018) 'Estimating distemper virus dynamics among wolves and grizzly bears using serology and Bayesian state-space models', *Ecology and Evolution*, 8(17), pp. 8726–8735. doi: 10.1002/ece3.4396.

Cumming, G. (2014) 'The New Statistics: Why and How', *Psychological Science*, 25(1), pp. 7–29. doi: 10.1177/0956797613504966.

Cushman, S. A. (2010) 'Space and time in ecology: Noise or fundamental driver?', in *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer Japan, pp. 19–41. doi: 10.1007/978-4-431-87771-4_2.

Dalley, D. *et al.* (2008) 'Development and evaluation of a gamma-interferon assay for tuberculosis in badgers (Meles meles)', *Tuberculosis*, 88(3), pp. 235–243. doi: 10.1016/j.tube.2007.11.001.

Dannemiller, N. G. *et al.* (2020) 'Diagnostic Uncertainty and the Epidemiology of Feline Foamy Virus in Pumas (Puma concolor)', *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-58350-7.

Darwin, C. (1859) *On the Origin of Species*. doi: doi: 10.4324/9780203509104/ORIGIN-SPECIES-1859-CHARLES-DARWIN.

Defries, R. and Nagendra, H. (2017) 'Ecosystem management as a wicked

problem', *Science*, 356(6335), pp. 265–270. doi: 10.1126/science.aal1950.

Delahay, R. J., Langton, S., *et al.* (2000) 'The spatio-temporal distribution of Mycobacterium boris (bovine tuberculosis) infection in a high-density badger population', *Journal of Animal Ecology*, 69(3), pp. 428–441. doi: 10.1046/j.1365-2656.2000.00406.x.

Delahay, R. J., Brown, J. A., *et al.* (2000) 'The use of marked bait in studies of the territorial organization of the European Badger (Meles meles)', *Mammal Review*, 30(2), pp. 73–87. doi: 10.1046/j.1365-2907.2000.00058.x.

Delahay, R. J., Cheeseman, C. L. and Clifton-Hadley, R. S. (2001) 'Wildlife disease reservoirs: The epidemiology of Mycobacterium bovis infection in the Eeuropean badger (Meles meles) and other British mammals', in *Tuberculosis*, pp. 43–49. doi: 10.1054/tube.2000.0266.

Dendukuri, N. *et al.* (2004) 'Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test', *Biometrics*, 60(2), pp. 388–397. doi: 10.1111/j.0006-341X.2004.00183.x.

Dendukuri, N., Bélisle, P. and Joseph, L. (2010) 'Bayesian sample size for diagnostic test studies in the absence of a gold standard: Comparing identifiable with non-identifiable models', *Statistics in Medicine*, 29(26), pp. 2688–2697. doi: 10.1002/sim.4037.

Dietze, M. C. (2017) 'Prediction in ecology: A first-principles framework: A', *Ecological Applications*, 27(7), pp. 2048–2060. doi: 10.1002/eap.1589.

DiRenzo, G. V. *et al.* (2018) 'Imperfect pathogen detection from non-invasive skin swabs biases disease inference', *Methods in Ecology and Evolution*, 9(2), pp. 380–389. doi: 10.1111/2041-210X.12868.

DiRenzo, G. V., Hanks, E. and Miller, D. A. W. (2023) ‘A practical guide to understanding and validating complex models using data simulations’, *Methods in Ecology and Evolution*, 14(1), pp. 203–217. doi: 10.1111/2041-210X.14030.

Donnelly, C. A. and Nouvellet, P. (2013) ‘The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England’, *PLoS Currents*, 5(Outbreaks). doi: 10.1371/currents.outbreaks.097a904d3f3619db2fe78d24bc776098.

Drewe, J. A. *et al.* (2010) ‘Diagnostic accuracy and optimal use of three tests for tuberculosis in live badgers’, *PLoS ONE*, 5(6), p. e11196. doi: 10.1371/journal.pone.0011196.

Duan, L. L., Johndrow, J. E. and Dunson, D. B. (2018) ‘Scaling up data augmentation MCMC via calibration’, *Journal of Machine Learning Research*, 19, pp. 1–34. Available at: https://www.jmlr.org/papers/volume19/17-573/17-573.pdf.

Dunson, D. B. (2001) ‘Commentary: Practical advantages of Bayesian analysis of epidemiologic data’, *American Journal of Epidemiology*, 153(12), pp. 1222–1226. doi: 10.1093/aje/153.12.1222.

Düx, A. *et al.* (2020) ‘Measles virus and rinderpest virus divergence dated to the sixth century BCE’, *Science*, 368(6497), pp. 1367–1370. doi: 10.1126/science.aba9411.

Eisler, Z., Bartos, I. and Kertész, J. (2008) ‘Fluctuation scaling in complex systems: Taylor’s law and beyond’, *Advances in Physics*, 57(1), pp. 89–142. doi: 10.1080/00018730801893043.

Elith, J. and Leathwick, J. R. (2009) ‘Species distribution models: Ecological

explanation and prediction across space and time', *Annual Review of Ecology, Evolution, and Systematics*, 40, pp. 677–697. doi: 10.1146/annurev.ecolsys.110308.120159.

Ellwanger, J. H. and Chies, J. A. B. (2021) 'Zoonotic spillover: Understanding basic aspects for better prevention', *Genetics and Molecular Biology*, 44(1), p. 20200355. doi: 10.1590/1678-4685-GMB-2020-0355.

Enøe, C., Georgiadis, M. P. and Johnson, W. O. (2000) 'Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown', *Preventive Veterinary Medicine*, 45(1–2), pp. 61–81. doi: 10.1016/S0167-5877(00)00117-3.

Enright, J. A. and O'Hare, A. (2017) 'Reconstructing disease transmission dynamics from animal movements and test data', *Stochastic Environmental Research and Risk Assessment*, 31(2), pp. 369–377. doi: 10.1007/s00477-016-1354-z.

Escobar, L. E. *et al.* (2020) 'The ecology of chronic wasting disease in wildlife', *Biological Reviews*, 95(2), pp. 393–408. doi: 10.1111/brv.12568.

Evans, T. S. *et al.* (2016) 'Habitat influences distribution of chronic wasting disease in white-tailed deer', *Journal of Wildlife Management*, 80(2), pp. 284–291. doi: 10.1002/jwmg.1004.

Ezenwa, V. O. *et al.* (2016) 'Host behaviour-parasite feedback: An essential link between animal behaviour and disease ecology', *Proceedings of the Royal Society B: Biological Sciences*, 283(1828), p. 20153078. doi: 10.1098/rspb.2015.3078.

Farnsworth, M. L. *et al.* (2005) 'Human land use influences chronic wasting

disease prevalence in mule deer', *Ecological Applications*, 15(1), pp. 119–126. doi: 10.1890/04-0194.

Feinleib, M. and Zar, J. H. (1975) *Biostatistical Analysis.*, *Journal of the American Statistical Association*. Pearson Education India. doi: 10.2307/2285423.

Feki-Sahnoun, W. *et al.* (2018) 'Using general linear model, Bayesian Networks and Naive Bayes classifier for prediction of Karenia selliformis occurrences and blooms', *Ecological Informatics*, 43, pp. 12–23. doi: 10.1016/j.ecoinf.2017.10.017.

Fischer, J. E., Bachmann, L. M. and Jaeschke, R. (2003) 'A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis', *Intensive Care Medicine*, 29(7), pp. 1043–1051. doi: 10.1007/s00134-003-1761-8.

Fitzgerald, S. D. and Kaneene, J. B. (2013) 'Wildlife Reservoirs of Bovine Tuberculosis Worldwide: Hosts, Pathology, Surveillance, and Control', *Veterinary Pathology*, 50(3), pp. 488–499. doi: 10.1177/0300985812467472.

Flegal, J. *et al.* (2017) 'Package "mcmcse"'. Available at: http://r.meteo.uni.wroc.pl/web/packages/mcmcse/mcmcse.pdf.

Flor, M. *et al.* (2020) 'Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification', *BMC Public Health*, 20(1), pp. 1–10. doi: 10.1186/s12889-020-09177-4.

Fogel, N. (2015) 'Tuberculosis: A disease without boundaries', *Tuberculosis*, 95(5), pp. 527–531. doi: 10.1016/j.tube.2015.05.017.

Forst, C. V. (2010) 'Host-pathogen systems biology', in *Infectious Disease*

Informatics. Springer New York, pp. 123–147. doi: 10.1007/978-1-4419-1327-2_6.

Fountain-Jones, N. M. *et al.* (2018) 'Towards an eco-phylogenetic framework for infectious disease ecology', *Biological Reviews*, 93(2), pp. 950–970. doi: 10.1111/brv.12380.

Fountain-Jones, N. M. *et al.* (2019) 'How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure', *Journal of Animal Ecology*, 88(10), pp. 1447–1461. doi: 10.1111/1365-2656.13076.

Gardner, I. A. *et al.* (2000) 'Conditional dependence between tests affects the diagnosis and surveillance of animal diseases', *Preventive Veterinary Medicine*, 45(1–2), pp. 107–122. doi: 10.1016/S0167-5877(00)00119-7.

Gardner, I. A. *et al.* (2011) 'Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants', *Preventive Veterinary Medicine*, 101(1–2), pp. 18–34. doi: 10.1016/j.prevetmed.2011.04.002.

Gardner, I. A. *et al.* (2021) 'Introduction Validation of tests for OIE-listed diseases as fit-for-purpose in a world of evolving diagnostic technologies', *OIE Revue Scientifique et Technique*, 40(1), pp. 19–28. doi: 10.20506/rst.40.1.3207.

Gardner, I. A., Johnson, W. O. and Norris, M. (2009) 'Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard', *Statistics and Its Interface*, 2(2), pp. 171–185. doi: 10.4310/sii.2009.v2.n2.a7.

Gaughran, A. *et al.* (2018) 'Super-ranging. A new ranging strategy in European badgers', *PLoS ONE*, 13(2), p. e0191818. doi: 10.1371/journal.pone.0191818.

Gavier-Widén, D. *et al.* (2009) 'A review of infection of wildlife hosts with mycobacterium bovis and the diagnostic difficulties of the "no visible lesion"

presentation', *New Zealand Veterinary Journal*, 57(3), pp. 122–131. doi: 10.1080/00480169.2009.36891.

Gelman, A. (2006) 'Multilevel (hierarchical) modeling: What It can and cannot do', *Technometrics*, 48(3), pp. 432–435. doi: 10.1198/004017005000000661.

Gelman, A. *et al.* (2010) 'Bridges between deterministic and probabilistic models for binary data', *Statistical Methodology*, 7(3), pp. 187–209. doi: 10.1016/j.stamet.2009.08.005.

Gelman, A. (2019) 'Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size', *Annals of Surgery*, 269(1), pp. E9–E10. doi: 10.1097/SLA.0000000000002908.

Gelman, A. and Carlin, J. (2014) 'Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors', *Perspectives on Psychological Science*, 9(6), pp. 641–651. doi: 10.1177/1745691614551642.

Gelman, A. and Carpenter, B. (2020) 'Bayesian analysis of tests with unknown specificity and sensitivity', *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 69(5), pp. 1269–1283. doi: 10.1111/rssc.12435.

Gelman, A. and Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. doi: 10.1017/cbo9780511790942.

Gelman, A., Meng, X. L. and Stern, H. (1996) 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica Sinica*, 6(4), pp. 733–807. Available at: https://www.jstor.org/stable/24306036 (Accessed: 19 March 2022).

Gelman, A. and Rubin, D. B. (1992) 'Inference from iterative simulation using

multiple sequences', *Statistical Science*, 7(4), pp. 457–472. doi: 10.1214/ss/1177011136.

Gelman, A., Simpson, D. and Betancourt, M. (2017) 'The prior can often only be understood in the context of the likelihood', *Entropy*, 19(10). doi: 10.3390/e19100555.

Geremia, C. *et al.* (2015) 'Bayesian modeling of prion disease dynamics in mule deer using population monitoring and capture-recapture data', *PLoS ONE*, 10(10), p. e0140687. doi: 10.1371/journal.pone.0140687.

Gibb, R. *et al.* (2020) 'Zoonotic host diversity increases in human-dominated ecosystems', *Nature 2020 584:7821*, 584(7821), pp. 398–402. doi: 10.1038/s41586-020-2562-8.

Gilbert, A. T. *et al.* (2013) 'Deciphering serology to understand the ecology of infectious diseases in wildlife', *EcoHealth*, 10(3), pp. 298–313. doi: 10.1007/s10393-013-0856-0.

Godfroid, J. *et al.* (2005) 'From the discovery of the Malta fever's agent to the discovery of a marine mammal reservoir, brucellosis has continuously been a re-emerging zoonosis', *Veterinary Research*, 36(3), pp. 313–326. doi: 10.1051/vetres:2005003.

Gonçalves, L. *et al.* (2012) 'Bayesian latent class models in malaria diagnosis', *PLoS ONE*, 7(7), p. e40633. doi: 10.1371/journal.pone.0040633.

Goodman, L. A. (1974) 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika*, 61(2), pp. 215–231. doi: 10.1093/biomet/61.2.215.

Graham, J. *et al.* (2013) 'Multi-state modelling reveals sex-dependent

transmission, progression and severity of tuberculosis in wild badgers', *Epidemiology and Infection*, 141(7), pp. 1429–1436. doi: 10.1017/S0950268812003019.

Green, D. M. and Swets, J. A. (1966) *Signal detection theory and psychophysics*. John Wiley.

Greenwald, R. *et al.* (2003) 'Improved serodetection of Mycobacterium bovis infection in badgers (Meles meles) using multiantigen test formats', *Diagnostic Microbiology and Infectious Disease*, 46(3), pp. 197–203. doi: 10.1016/S0732-8893(03)00046-4.

Greiner, M. and Gardner, I. A. (2000) 'Epidemiologic issues in the validation of veterinary diagnostic tests', *Preventive Veterinary Medicine*, 45(1–2), pp. 3–22. doi: 10.1016/S0167-5877(00)00114-8.

Gustafson, P. *et al.* (2005) 'On model expansion, model contraction, identifiability and prior Information: Two illustrative scenarios involving mismeasured variables', *Statistical Science*, 20(2), pp. 111–140. doi: 10.1214/088342305000000098.

Habibzadeh, F., Habibzadeh, P. and Yadollahie, M. (2016) 'On determining the most appropriate test cut-off value: The case of tests with continuous results', *Biochemia Medica*, 26(3), pp. 297–307. doi: 10.11613/BM.2016.034.

Hahn, A., Schwarz, N. G. and Frickmann, H. (2019) 'Comparison of screening tests without a gold standard—A pragmatic approach with virtual reference testing', *Acta Tropica*, 199, p. 105118. doi: 10.1016/J.ACTATROPICA.2019.105118.

Hallman, T. A. and Robinson, W. D. (2020) 'Deciphering ecology from statistical

artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance', *Diversity and Distributions*, 26(3), pp. 315–328. doi: 10.1111/ddi.13030.

Halsey, L. G. *et al.* (2015) 'The fickle P value generates irreproducible results', *Nature Methods*, 12(3), pp. 179–185. doi: 10.1038/nmeth.3288.

Halsey, L. G. (2019) 'The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum?', *Biology Letters*. The Royal Society. doi: 10.1098/rsbl.2019.0174.

Hanley, J. A. and McNeil, B. J. (1982) 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, 143(1), pp. 29–36. doi: 10.1148/radiology.143.1.7063747.

Harper, E. B., Stella, J. C. and Fremier, A. K. (2011) 'Global sensitivity analysis for complex ecological models: A case study of riparian cottonwood population dynamics', *Ecological Applications*, 21(4), pp. 1225–1240. doi: 10.1890/10-0506.1.

Harrison, X. A. *et al.* (2018) 'A brief introduction to mixed effects modelling and multi-model inference in ecology', *PeerJ*, 2018(5), p. e4794. doi: 10.7717/peerj.4794.

Hartig, F. *et al.* (2017) 'BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics', *R package version 0.1.7*.

Hastings, A. *et al.* (1993) 'Chaos in ecology: Is mother nature a strange attractor?', *Annual Review of Ecology and Systematics*, pp. 1–33. doi: 10.1146/annurev.es.24.110193.000245.

Hawking, S. and Mlodinow, L.- (2010) ‘The (elusive) theory of everything’, *JSTOR*. Available at: https://www.jstor.org/stable/26002214.

Hayes, A. F. and Cai, L. (2007) ‘Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation’, *Behavior Research Methods*, 39(4), pp. 709–722. doi: 10.3758/BF03192961.

Hefley, T. J. *et al.* (2017) ‘When mechanism matters: Bayesian forecasting using models of ecological diffusion’, *Ecology Letters*, 20(5), pp. 640–650. doi: 10.1111/ele.12763.

Heisey, D. M. *et al.* (2010) ‘Rejoinder: Sifting through model space’, *Ecology*, 91(12), pp. 3503–3514. doi: 10.1890/10-0894.1.

Hellmann, J. J. and Fowler, G. W. (1999) ‘Bias, precision, and accuracy of four measures of species richness’, *Ecological Applications*, 9(3), pp. 824–834. doi: 10.1890/1051-0761(1999)009[0824:BPAAOF]2.0.CO;2.

Helman, S. K. *et al.* (2020) ‘Estimating prevalence and test accuracy in disease ecology: How Bayesian latent class analysis can boost or bias imperfect test results’, *Ecology and Evolution*, 10(14), pp. 7221–7232. doi: 10.1002/ece3.6448.

Hines, D. E., Ray, S. and Borrett, S. R. (2018) ‘Uncertainty analyses for Ecological Network Analysis enable stronger inferences’, *Environmental Modelling and Software*, 101, pp. 117–127. doi: 10.1016/j.envsoft.2017.12.011.

Hobbs, N. T. and Hooten, M. B. (2015) *Bayesian models: A statistical primer for ecologists*, *Bayesian Models: A Statistical Primer for Ecologists*. Princeton University Press.

Hodgson, D. J. (2022) ‘Email communication from David Hodgson, 21 April’.

Holyoak, M. and Wetzel, C. (2020) ‘Variance-Explicit Ecology: A Call for Holistic Study of the Consequences of Variability at Multiple Scales’, in *Unsolved Problems in Ecology*, pp. 25–42. doi: 10.1515/9780691195322-005.

Hooten, M. B., Hobbs, N. T. and Ellison, A. M. (2015) ‘A guide to Bayesian model selection for ecologists’, *Ecological Monographs*, 85(1), pp. 3–28. doi: 10.1890/14-0661.1.

Horne, J. K. and Schneider, D. C. (1995) ‘Spatial Variance in Ecology’, *Oikos*, 74(1), p. 18. doi: 10.2307/3545670.

Hu, B., Gonzales, J. L. and Gubbins, S. (2017) ‘Bayesian inference of epidemiological parameters from transmission experiments’, *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-17174-8.

Hudson, D. W. *et al.* (2019) ‘Analysis of lifetime mortality trajectories in wildlife disease research: BaSTA and Beyond’, *Diversity*, 11(10), p. 182. doi: 10.3390/d11100182.

Hui, S. L. and Walter, S. D. (1980) ‘Estimating the Error Rates of Diagnostic Tests’, *Biometrics*, 36(1), p. 167. doi: 10.2307/2530508.

Ings, T. C. *et al.* (2009) ‘Ecological networks - Beyond food webs’, *Journal of Animal Ecology*, 78(1), pp. 253–269. doi: 10.1111/j.1365-2656.2008.01460.x.

Islam, M. A. *et al.* (2020) ‘Bayesian latent class evaluation of three tests for the screening of subclinical caprine mastitis in Bangladesh’, *Tropical Animal Health and Production*, 52(6), pp. 2873–2881. doi: 10.1007/s11250-020-02263-0.

Jia, B. *et al.* (2020) ‘Validation of laboratory tests for infectious diseases in wild mammals: review and recommendations’, *Journal of Veterinary Diagnostic*

Investigation, 32(6), pp. 776–792. doi: 10.1177/1040638720920346.

Jiang, J. and Lahiri, P. (2006) 'Mixed model prediction and small area estimation', *Test*, 15(1), pp. 1–96. doi: 10.1007/BF02595419.

Johnson, P. C. D. (2014) 'Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models', *Methods in Ecology and Evolution*, 5(9), pp. 944–946. doi: 10.1111/2041-210X.12225.

Johnson, P. T. J., Ostfeld, R. S. and Keesing, F. (2015) 'Frontiers in research on biodiversity and disease', *Ecology Letters*, 18(10), pp. 1119–1133. doi: 10.1111/ele.12479.

Johnson, W. O. *et al.* (2009) 'On the interpretation of test sensitivity in the two-test two-population problem: Assumptions matter', *Preventive Veterinary Medicine*, 91(2–4), pp. 116–121. doi: 10.1016/j.prevetmed.2009.06.006.

Johnson, W. O., Gastwirth, J. L. and Pearson, L. M. (2001) 'Screening without a "gold standard": The Hui-Walter paradigm revisited', *American Journal of Epidemiology*, 153(9), pp. 921–924. doi: 10.1093/aje/153.9.921.

Jones, G. *et al.* (2010) 'Identifiability of models for multiple diagnostic testing in the absence of a gold standard', *Biometrics*, 66(3), pp. 855–863. doi: 10.1111/j.1541-0420.2009.01330.x.

Jones, O. R. and Vaupel, J. W. (2017) 'Senescence is not inevitable', *Biogerontology*, 18(6), pp. 965–971. doi: 10.1007/s10522-017-9727-3.

Joseph, L., Gyorkos, T. W. and Coupal, L. (1995) 'Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard', *American Journal of Epidemiology*, 141(3), pp. 263–272. doi: 10.1093/oxfordjournals.aje.a117428.

Jovani, R. and Tella, J. L. (2006) ‘Parasite prevalence and sample size: misconceptions and solutions’, *Trends in Parasitology*, 22(5), pp. 214–218. doi: 10.1016/j.pt.2006.02.011.

Kaldor, N. (1961) ‘Capital Accumulation and Economic Growth’, *The Theory of Capital*, pp. 177–222. doi: 10.1007/978-1-349-08452-4_10.

Kao, Y. H. and Eisenberg, M. C. (2018) ‘Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment’, *Epidemics*, 25, pp. 89–100. doi: 10.1016/j.epidem.2018.05.010.

Kéfi, S. *et al.* (2012) ‘More than a meal... integrating non-feeding interactions into food webs’, *Ecology Letters*, 15(4), pp. 291–300. doi: 10.1111/j.1461-0248.2011.01732.x.

Kellner, K. F. (2015) ‘jagsUI: a wrapper around rjags to streamline JAGS analyses’. Available at: https://cran.r-project.org/web/packages/jagsUI/index.html.

Kellner, K. F. and Swihart, R. K. (2014) ‘Accounting for imperfect detection in ecology: A quantitative review’, *PLoS ONE*. Public Library of Science, p. e111436. doi: 10.1371/journal.pone.0111436.

Kéry, M. and Schaub, M. (2011) *Bayesian Population Analysis using WinBUGS: A Hierarchical Perspective*, *Bayesian Population Analysis using WinBUGS: A Hierarchical Perspective*. doi: 10.1016/C2010-0-68368-4.

Khomtchouk, B. B., Hennessy, J. R. and Wahlestedt, C. (2017) ‘Shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics’, *PLoS ONE*, 12(5). doi: 10.1371/journal.pone.0176334.

King, A. W. (1997) ‘Hierarchy Theory: A Guide to System Structure for Wildlife Biologists’, in *Wildlife and Landscape Ecology*. Springer New York, pp. 185–212. doi: 10.1007/978-1-4612-1918-7_7.

Kosmala, M. *et al.* (2016) ‘Estimating wildlife disease dynamics in complex systems using an Approximate Bayesian Computation framework’, *Ecological Applications*, 26(1), pp. 295–308. doi: 10.1890/14-1808.

Kostoulas, P. *et al.* (2017) ‘STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models’, *Preventive Veterinary Medicine*, 138, pp. 37–47. doi: 10.1016/j.prevetmed.2017.01.006.

Krebs, J. R. *et al.* (1998) ‘Badgers and bovine TB: Conflicts between conservation and health’, *Science*, 279(5352), pp. 817–818. doi: 10.1126/science.279.5352.817.

Krolewiecki, A. J. *et al.* (2018) ‘Transrenal DNA-based diagnosis of Strongyloides stercoralis (Grassi, 1879) infection: Bayesian latent class modeling of test accuracy’, *PLoS Neglected Tropical Diseases*, 12(6), p. e0006550. doi: 10.1371/journal.pntd.0006550.

Kruschke, J. K. (2013) ‘Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, “Philosophy and the practice of Bayesian statistics”’, *British Journal of Mathematical and Statistical Psychology*, 66(1), pp. 45–56. doi: 10.1111/j.2044-8317.2012.02063.x.

Kruschke, J. K. (2014) *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition*, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Edition*. doi: 10.1016/B978-0-12-405888-0.09999-2.

Kruschke, J. K. and Liddell, T. M. (2018) ‘The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective’, *Psychonomic Bulletin and Review*, 25(1), pp. 178–206. doi: 10.3758/s13423-016-1221-4.

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017) ‘lmerTest Package: Tests in Linear Mixed Effects Models’, *Journal of Statistical Software*, 82(13), pp. 1–26. doi: 10.18637/JSS.V082.I13.

Lachish, S. *et al.* (2012) ‘Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests’, *Methods in Ecology and Evolution*, 3(2), pp. 339–348. doi: 10.1111/j.2041-210X.2011.00156.x.

Lachish, S. and Murray, K. A. (2018) ‘The certainty of uncertainty: Potential sources of bias and imprecision in disease ecology studies’, *Frontiers in Veterinary Science*, 5(MAY), p. 90. doi: 10.3389/fvets.2018.00090.

Lau, C. L. *et al.* (2017) ‘Unravelling infectious disease eco-epidemiology using Bayesian networks and scenario analysis: A case study of leptospirosis in Fiji’, *Environmental Modelling and Software*, 97, pp. 271–286. doi: 10.1016/j.envsoft.2017.08.004.

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*. Houghton-Mifflin, Boston.

Lazic, S. E. *et al.* (2020) ‘A Bayesian predictive approach for dealing with pseudoreplication’, *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-59384-7.

Leeflang, M. M. G. *et al.* (2013) ‘Variation of a test’s sensitivity and specificity with disease prevalence’, *CMAJ. Canadian Medical Association Journal*,

185(11), p. E537. doi: 10.1503/cmaj.121286.

Lewis, F. I. and Torgerson, P. R. (2012) ‘A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic’, *Emerging Themes in Epidemiology*, 9(1), pp. 1–8. doi: 10.1186/1742-7622-9-9.

Li, J. and Fine, J. P. (2011) ‘Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis’, *Biostatistics*, 12(4), pp. 710–722. doi: 10.1093/biostatistics/kxr008.

Li, W. *et al.* (2005) ‘Bats are natural reservoirs of SARS-like coronaviruses’, *Science*, 310(5748), pp. 676–679. doi: 10.1126/science.1118391.

Li, Y. *et al.* (2018) ‘Bayesian Latent Class Analysis Tutorial’, *Multivariate Behavioral Research*, 53(3), pp. 430–451. doi: 10.1080/00273171.2018.1428892.

Lindén, A. and Mäntyniemi, S. (2011) ‘Using the negative binomial distribution to model overdispersion in ecological count data’, *Ecology*, 92(7), pp. 1414–1421. doi: 10.1890/10-1831.1.

Link, W. A. and Eaton, M. J. (2012) ‘On thinning of chains in MCMC’, *Methods in Ecology and Evolution*, 3(1), pp. 112–115. doi: 10.1111/j.2041-210X.2011.00131.x.

Liu, J. *et al.* (2014) ‘A two-stage Bayesian method for estimating accuracy and disease prevalence for two dependent dichotomous screening tests when the status of individuals who are negative on both tests is unverified’, *BMC Medical Research Methodology*, 14(1), pp. 1–11. doi: 10.1186/1471-2288-14-110.

Liu, Y. L. *et al.* (2022) ‘Extending Hui-Walter framework to correlated outcomes

with application to diagnosis tests of an eye disease among premature infants', *Statistics in Medicine*, 41(3), pp. 433–448. doi: 10.1002/sim.9269.

Lohr, C. *et al.* (2017) 'Predicting island biosecurity risk from introduced fauna using Bayesian Belief Networks', *Science of the Total Environment*, 601–602, pp. 1173–1181. doi: 10.1016/j.scitotenv.2017.05.281.

Lorah, J. (2018) 'Effect size measures for multilevel models: definition, interpretation, and TIMSS example', *Large-Scale Assessments in Education*, 6(1), pp. 1–11. doi: 10.1186/s40536-018-0061-2.

Lunn, D. J. *et al.* (2000) 'WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility', *Statistics and Computing*, 10(4), pp. 325–337. doi: 10.1023/A:1008929526011.

Lütkenhöner, B. and Basel, T. (2013) 'Predictive modeling for diagnostic tests with high specificity, but low sensitivity: A study of the glycerol test in patients with suspected Menière's disease', *PLoS ONE*. Edited by X. Wang, 8(11), p. e79315. doi: 10.1371/journal.pone.0079315.

Lynch, T. *et al.* (2010) 'A systematic review on the diagnosis of pediatric bacterial pneumonia: When gold is bronze', *PLoS ONE*, 5(8). doi: 10.1371/journal.pone.0011989.

Lynn, P. and Healey, J. F. (1992) *Statistics: A Tool for Social Research.*, *The Statistician*. Wadsworth Pub. Co. doi: 10.2307/2348267.

Maas, C. J. M. and Hox, J. J. (2005) 'Sufficient sample sizes for multilevel modeling', *Methodology*, 1(3), pp. 86–92. doi: 10.1027/1614-2241.1.3.86.

Manyweathers, J. *et al.* (2020) 'Understanding the vulnerability of beef producers in Australia to an FMD outbreak using a Bayesian Network predictive

model', *Preventive Veterinary Medicine*, 175, p. 104872. doi: 10.1016/j.prevetmed.2019.104872.

Manzoli, D. E. *et al.* (2013) 'Multi-Level Determinants of Parasitic Fly Infection in Forest Passerines', *PLoS ONE*, 8(7), p. e67104. doi: 10.1371/journal.pone.0067104.

Mariën, J. *et al.* (2017) 'Arenavirus Dynamics in Experimentally and Naturally Infected Rodents', *EcoHealth*, 14(3), pp. 463–473. doi: 10.1007/s10393-017-1256-7.

Martin, A. M. *et al.* (2019) 'Population-scale treatment informs solutions for control of environmentally transmitted wildlife disease', *Journal of Applied Ecology*, 56(10), pp. 2363–2375. doi: 10.1111/1365-2664.13467.

Martin, L. E. R. *et al.* (2017) 'Weather influences trapping success for tuberculosis management in European badgers (Meles meles)', *European Journal of Wildlife Research*, 63(1), p. 30. doi: 10.1007/s10344-017-1089-2.

Mazeri, S. *et al.* (2016) 'Evaluation of the performance of five diagnostic tests for Fasciola hepatica infection in naturally infected cattle using a Bayesian no gold standard approach', *PLoS ONE*, 11(8), p. e0161621. doi: 10.1371/journal.pone.0161621.

McAloon, C. G. *et al.* (2019) 'A review of paratuberculosis in dairy herds — Part 1: Epidemiology', *Veterinary Journal*, pp. 59–65. doi: 10.1016/j.tvjl.2019.01.010.

McCallum, H. *et al.* (2009) 'Transmission dynamics of Tasmanian devil facial tumor disease may lead to disease-induced extinction', *Ecology*, 90(12), pp. 3379–3392. doi: 10.1890/08-1763.1.

McClintock, B. T. *et al.* (2010) 'Seeking a second opinion: Uncertainty in

disease ecology', *Ecology Letters*, 13(6), pp. 659–674. doi: 10.1111/j.1461-0248.2010.01472.x.

McDonald, J. L. *et al.* (2014) 'Mortality trajectory analysis reveals the drivers of sex-specific epidemiology in natural wildlife - Disease interactions', *Proceedings of the Royal Society B: Biological Sciences*, 281(1790), p. 20140526. doi: 10.1098/rspb.2014.0526.

McDonald, J. L. *et al.* (2016) 'Demographic buffering and compensatory recruitment promotes the persistence of disease in a wildlife population', *Ecology Letters*, 19(4), pp. 443–449. doi: 10.1111/ele.12578.

McDonald, J. L. and Hodgson, D. J. (2018) 'Prior precision, prior accuracy, and the estimation of disease prevalence using imperfect diagnostic tests', *Frontiers in Veterinary Science*, 5(MAY), pp. 2297–1769. doi: 10.3389/fvets.2018.00083.

McDonald, J. L., Robertson, A. and Silk, M. J. (2018) 'Wildlife disease ecology from the individual to the population: Insights from a long-term study of a naturally infected European badger population', *Journal of Animal Ecology*, 87(1), pp. 101–112. doi: 10.1111/1365-2656.12743.

McKay, M. D., Beckman, R. J. and Conover, W. J. (2000) 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics*, 42(1), pp. 55–61. doi: 10.1080/00401706.2000.10485979.

McLachlan, G. J., Lee, S. X. and Rathnayake, S. I. (2019) 'Finite mixture models', *Annual Review of Statistics and Its Application*, 6, pp. 355–378. doi: 10.1146/annurev-statistics-031017-100325.

Merkle, J. A. *et al.* (2018) 'Linking spring phenology with mechanistic models of

host movement to predict disease transmission risk', *Journal of Applied Ecology*, 55(2), pp. 810–819. doi: 10.1111/1365-2664.13022.

Meyer, H. and Pebesma, E. (2022) 'Machine learning-based global maps of ecological variables and the challenge of assessing them', *Nature Communications*. Nature Publishing Group, pp. 1–4. doi: 10.1038/s41467-022-29838-9.

Miller, D. A. W. *et al.* (2012) 'Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs', *Methods in Ecology and Evolution*, 3(5), pp. 850–859. doi: 10.1111/j.2041-210X.2012.00216.x.

Miller, W. C. (2012) 'Commentary: Reference-test Bias in Diagnostic-test Evaluation: A Problem for Epidemiologists, too', 23(1), pp. 83–85. doi: 10.1097/EDE.0b013e31823b5b5b.

Milner-Gulland, E. J. and Shea, K. (2017) 'Embracing uncertainty in applied ecology', *Journal of Applied Ecology*, 54(6), pp. 2063–2068. doi: 10.1111/1365-2664.12887.

Milns, I., Beale, C. M. and Anne Smith, V. (2010) 'Revealing ecological networks using Bayesian network inference algorithms', *Ecology*, 91(7), pp. 1892–1899. doi: 10.1890/09-0731.1.

Mitchell, D. J., Beckmann, C. and Biro, P. A. (2021) 'Understanding the unexplained: The magnitude and correlates of individual differences in residual variance', *Ecology and Evolution*, 11(12), pp. 7201–7210. doi: 10.1002/ece3.7603.

Mollentze, N. and Streicker, D. G. (2020) 'Viral zoonotic risk is homogenous

among taxonomic orders of mammalian and avian reservoir hosts', *Proceedings of the National Academy of Sciences of the United States of America*, 117(17), pp. 9423–9430. doi: 10.1073/pnas.1919176117.

Møller, A. P. and Jennions, M. D. (2002) 'How much variance can be explained by ecologists and evolutionary biologists?', *Oecologia*, 132(4), pp. 492–500. doi: 10.1007/s00442-002-0952-2.

Monello, R. J. *et al.* (2017) 'Pathogen-mediated selection in free-ranging elk populations infected by chronic wasting disease', *Proceedings of the National Academy of Sciences of the United States of America*, 114(46), pp. 12208–12212. doi: 10.1073/pnas.1707807114.

Monnahan, C. C., Thorson, J. T. and Branch, T. A. (2017) 'Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo', *Methods in Ecology and Evolution*. British Ecological Society, pp. 339–348. doi: 10.1111/2041-210X.12681.

Montoya, J. M., Pimm, S. L. and Solé, R. V. (2006) 'Ecological networks and their fragility', *Nature*, 442(7100), pp. 259–264. doi: 10.1038/nature04927.

More, S. J. *et al.* (2018) 'Further description of bovine tuberculosis trends in the United Kingdom and the Republic of Ireland, 2003-2015', *Veterinary Record*, 183(23), p. 717. doi: 10.1136/vr.104718.

Morris, R. S. and Pfeiffer, D. U. (1995) 'Directions and issues in bovine tuberculosis epidemiology and control in New Zealand', *New Zealand Veterinary Journal*, 43(7), pp. 256–265. doi: 10.1080/00480169./1995.35904.

Mouquet, N. *et al.* (2015) 'Predictive ecology in a changing world', *Journal of Applied Ecology*, 52(5), pp. 1293–1310. doi: 10.1111/1365-2664.12482.

Muma, J. B. *et al.* (2007) ‘Evaluation of three serological tests for brucellosis in naturally infected cattle using latent class analysis’, *Veterinary Microbiology*, 125(1–2), pp. 187–192. doi: 10.1016/j.vetmic.2007.05.012.

Munch, S. B., Poynor, V. and Arriaza, J. L. (2017) ‘Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology’, *Ecological Complexity*, 32, pp. 134–143. doi: 10.1016/j.ecocom.2016.08.006.

Nakagawa, S. and Schielzeth, H. (2013) ‘A general and simple method for obtaining R2 from generalized linear mixed-effects models’, *Methods in Ecology and Evolution*, 4(2), pp. 133–142. doi: 10.1111/j.2041-210x.2012.00261.x.

Naujokaitis-Lewis, I. R. *et al.* (2009) ‘Sensitivity analyses of spatial population viability analysis models for species at risk and habitat conservation planning’, *Conservation Biology*, 23(1), pp. 225–229. doi: 10.1111/j.1523-1739.2008.01066.x.

Newman, K. B. (1998) ‘State-Space Modeling of Animal Movement and Mortality with Application to Salmon’, *Biometrics*, 54(4), p. 1290. doi: 10.2307/2533659.

Noonan, M. J. (2015) ‘The socio-ecological functions of fossoriality in a group-living carnivore, the European badger (Meles meles)’, *Oxford University Research Archive*. Available at: https://ora.ox.ac.uk/objects/uuid:69ea12af-f012-41ec-9359-6983cee8590a.

O’Hagan, M. J. H. *et al.* (2019) ‘Test characteristics of the tuberculin skin test and post-mortem examination for bovine tuberculosis diagnosis in cattle in Northern Ireland estimated by Bayesian latent class analysis with adjustments for covariates’, *Epidemiology and Infection*, 147, pp. 1–8. doi: 10.1017/S0950268819000888.

O'Hare, A. *et al.* (2014) 'Estimating epidemiological parameters for bovine tuberculosis in British cattle using a Bayesian partial-likelihood approach', *Proceedings of the Royal Society B: Biological Sciences*, 281(1783), p. 20140248. doi: 10.1098/rspb.2014.0248.

Ochola, L. *et al.* (2006) 'The reliability of diagnostic techniques in the diagnosis and management of malaria in the absence of a gold standard', *Lancet Infectious Diseases*, 6(9), pp. 582–588. doi: 10.1016/S1473-3099(06)70579-5.

Olsen, A. *et al.* (2022) 'Determination of an optimal ELISA cut-off for the diagnosis of Toxoplasma gondii infection in pigs using Bayesian latent class modelling of data from multiple diagnostic tests', *Preventive Veterinary Medicine*, 201, p. 105606. doi: 10.1016/j.prevetmed.2022.105606.

Omurtag, A. and Fenton, A. A. (2012) 'Assessing Diagnostic Tests: How to Correct for the Combined Effects of Interpretation and Reference Standard', *PLoS ONE*, 7(12), p. e52221. doi: 10.1371/journal.pone.0052221.

Palmer, M. A., Hakenkamp, C. C. and Nelson-Baker, K. (1997) 'Ecological heterogeneity in streams: Why variance matters', *Journal of the North American Benthological Society*, 16(1), pp. 189–202. doi: 10.2307/1468251.

Pandit, P. and A. Han, B. (2020) 'Rise of Machines in Disease Ecology', *The Bulletin of the Ecological Society of America*, 101(1), p. e01625. doi: 10.1002/bes2.1625.

Patel, K. K. *et al.* (2022) 'Bayesian evaluation of temporal changes in sensitivity and specificity of three serological tests for multiple circulating strains of rabbit haemorrhagic disease virus', *Authorea Preprints*. doi: 10.22541/AU.166065663.30872438/V1.

Patterson, T. A. *et al.* (2008) ‘State-space models of individual animal movement’, *Trends in Ecology and Evolution*, 23(2), pp. 87–94. doi: 10.1016/j.tree.2007.10.009.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, *Probabilistic Reasoning in Intelligent Systems*. Elsevier. doi: 10.1016/c2009-0-27609-4.

Peek, M. S. *et al.* (2003) ‘How much variance is explained by ecologists? Additional perspectives’, *Oecologia*, 137(2), pp. 161–170. doi: 10.1007/s00442-003-1328-y.

Peel, A. J. *et al.* (2018) ‘Support for viral persistence in bats from age-specific serology and models of maternal immunity’, *Scientific Reports*, 8(1), pp. 1–11. doi: 10.1038/s41598-018-22236-6.

Pepin, K. M. *et al.* (2017) ‘Inferring infection hazard in wildlife populations by linking data across individual and population scales’, *Ecology Letters*, 20(3), pp. 275–292. doi: 10.1111/ele.12732.

Pereira, G. D. A. *et al.* (2012) ‘A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: An application to Chagas disease’, *Computational and Mathematical Methods in Medicine*, 2012. doi: 10.1155/2012/487502.

Pfukenyi, D. M. *et al.* (2020) ‘Evaluation of the sensitivity and specificity of the lateral flow assay, Rose Bengal test and the complement fixation test for the diagnosis of brucellosis in cattle using Bayesian latent class analysis’, *Preventive Veterinary Medicine*, 181. doi: 10.1016/j.prevetmed.2020.105075.

Plotkin, J. B. (2017) ‘Molecular evolution: No escape from the tangled bank’, *Nature*, 551(7678), pp. 42–43. doi: 10.1038/nature24152.

Plummer, M. (2003) *JAGS: a program for analysis of bayesian graphical models using Gibbs sampling*, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22. Vienna, Austria.* Available at: https://mcmc-jags.sourceforge.io/.

Plummer, M. *et al.* (2006) 'CODA: convergence diagnosis and output analysis for MCMC', *journal.r-project.org*. Available at: https://journal.r-project.org/articles/RN-2006-002/RN-2006-002.pdf.

Pollock, J. M., Welsh, M. D. and McNair, J. (2005) 'Immune responses in bovine tuberculosis: Towards new strategies for the diagnosis and control of disease', in *Veterinary Immunology and Immunopathology*, pp. 37–43. doi: 10.1016/j.vetimm.2005.08.012.

Ponciano, J. M. *et al.* (2012) 'Assessing Parameter Identifiability in Phylogenetic Models Using Data Cloning', *Systematic Biology*, 61(6), pp. 955–972. doi: 10.1093/sysbio/sys055.

Pool, R. (1989) 'Ecologists flirt with chaos', *Science*, 243(4889), pp. 310–313. doi: 10.1126/science.243.4889.310.

Porta, M. (2016) *A Dictionary of Epidemiology*, *International Journal of Epidemiology*. Oxford University Press. doi: 10.1093/acref/9780199976720.001.0001.

Pourakbari, B. *et al.* (2018) 'Evaluation of the QuantiFERON®-TB Gold In-Tube assay and tuberculin skin test for the diagnosis of latent tuberculosis infection in an Iranian referral hospital', *Infectious Disorders - Drug Targets*, 18(2). doi: 10.2174/1871526518666180228164036.

Prowse, T. A. A. *et al.* (2016) 'An efficient protocol for the global sensitivity

analysis of stochastic ecological models', *Ecosphere*, 7(3). doi: 10.1002/ecs2.1238.

R Core Team (2023) 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org/.

Raghavan, R. K. *et al.* (2016) 'Bayesian Spatiotemporal Pattern and Eco-climatological Drivers of Striped Skunk Rabies in the North Central Plains', *PLoS Neglected Tropical Diseases*, 10(4), p. e0004632. doi: 10.1371/journal.pntd.0004632.

Ragonnet-Cronin, M. *et al.* (2021) 'Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions', *Nature Communications*, 12(1). doi: 10.1038/s41467-021-22366-y.

Rahman, A. K. M. A. *et al.* (2019) 'Bayesian evaluation of three serological tests for the diagnosis of bovine brucellosis in bangladesh', *Epidemiology and Infection*, 147. doi: 10.1017/S0950268818003503.

Ransohoff, D. F. and Feinstein, A. R. (1978) 'Problems of Spectrum and Bias in Evaluating the Efficacy of Diagnostic Tests', *New England Journal of Medicine*, 299(17), pp. 926–930. doi: 10.1056/nejm197810262991705.

Ratner, B. (2009) 'The correlation coefficient: Its values range between 1/1, or do they', *Journal of Targeting, Measurement and Analysis for Marketing*, 17(2), pp. 139–142. doi: 10.1057/jt.2009.5.

van Ravenzwaaij, D., Cassey, P. and Brown, S. D. (2018) 'A simple introduction to Markov Chain Monte–Carlo sampling', *Psychonomic Bulletin and Review*,

25(1), pp. 143–154. doi: 10.3758/s13423-016-1015-8.

Regan, H. M., Colyvan, M. and Burgman, M. A. (2002) ‘A taxonomy and treatment of uncertainty for ecology and conservation biology’, *Ecological Applications*, 12(2), pp. 618–628. doi: 10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2.

Richardson, D. M. and Pyšek, P. (2007) ‘Elton, C.S. 1958: The ecology of invasions by animals and plants. London: Methuen’, *Progress in Physical Geography*, 31(6), pp. 659–666. doi: 10.1177/0309133307087089.

Ries, L. *et al.* (2004) ‘Ecological responses to habitat edges: Mechanisms, models, and variability explained’, *Annual Review of Ecology, Evolution, and Systematics*, 35, pp. 491–522. doi: 10.1146/annurev.ecolsys.35.112202.130148.

Ries, L. and Sisk, T. D. (2004) ‘A predictive model of edge effects’, *Ecology*. Ecological Society of America, pp. 2917–2926. doi: 10.1890/03-8021.

Rights, J. D. and Sterba, S. K. (2019) ‘Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures’, *Psychological Methods*, 24(3), pp. 309–338. doi: 10.1037/met0000184.

Rindskopf, D. and Rindskopf, W. (1986) ‘The value of latent class analysis in medical diagnosis’, *Statistics in Medicine*, 5(1), pp. 21–27. doi: 10.1002/sim.4780050105.

Rittel, H. W. J. and Webber, M. M. (1973) *Dilemmas in a general theory of planning*, *Policy Sciences*. doi: 10.1007/BF01405730.

Roberts, D. W. (2017) ‘Distance, dissimilarity, and mean–variance ratios in ordination’, *Methods in Ecology and Evolution*, 8(11), pp. 1398–1407. doi:

10.1111/2041-210X.12739.

Roberts, G. O. (1995) *Markov Chain Monte Carlo in Practice*. CRC Press. doi: 10.1201/b14835-8.

Rodríguez, R. A. *et al.* (2019) ‘Degrees of freedom: Definitions and their minimum and most meaningful combination for the modelling of ecosystem dynamics with the help of physical principles’, *Ecological Modelling*, 392, pp. 226–235. doi: 10.1016/j.ecolmodel.2018.11.021.

Rogers, L. M. *et al.* (1999) ‘The increase in badger (Meles meles) density at Woodchester Park, south-west England: A review of the implications for disease (Mycobacterium bovis) prevalence’, *Mammalia*, 63(2), pp. 183–192. doi: 10.1515/mamm.1999.63.2.183.

Roosa, K. and Chowell, G. (2019) ‘Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models 01 Mathematical Sciences 0104 Statistics 01 Mathematical Sciences 0102 Applied Mathematics’, *Theoretical Biology and Medical Modelling*, 16(1). doi: 10.1186/s12976-018-0097-6.

Rothman, K. J., Greenland, S. and Lash, T. (2014) *Modern Epidemiology, 3rd Edition*. Lippincott Williams & Wilkins. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24644503 (Accessed: 3 July 2023).

Royle, J. A. and Young, K. V. (2008) ‘A hierarchical model for spatial capture recapture data’, *Ecology*, 89(8), pp. 2281–2289. doi: 10.1890/07-0601.1.

Rydevik, G., Innocent, G. T. and McKendrick, I. J. (2018) ‘Evaluating diagnostic tests with near-perfect specificity: Use of a Hui-Walter approach when designing a trial of a DIVA test for bovine tuberculosis’, *Frontiers in Veterinary Science*,

5(AUG), p. 192. doi: 10.3389/fvets.2018.00192.

Rykiel, E. J. (1996) ‘Testing ecological models: The meaning of validation’, *Ecological Modelling*, 90(3), pp. 229–244. doi: 10.1016/0304-3800(95)00152-2.

Saint-Mont, U. (2022) ‘Induction: A Logical Analysis’, *Foundations of Science*, 27(2), pp. 455–487. doi: 10.1007/s10699-020-09683-z.

Saltelli, A. *et al.* (2020) ‘Five ways to ensure that models serve society: a manifesto’, *Nature*, 582(7813), pp. 482–484. doi: 10.1038/d41586-020-01812-9.

Samuel, M. D. *et al.* (2015) ‘Avian malaria in Hawaiian forest birds: Infection and population impacts across species and elevations’, *Ecosphere*, 6(6). doi: 10.1890/ES14-00393.1.

Sander, E. L., Wootton, J. T. and Allesina, S. (2017) ‘Ecological Network Inference from Long-Term Presence-Absence Data’, *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-07009-x.

Scheiner, S. M. and Willig, M. R. (2013) *The Theory of Ecology*, *The Theory of Ecology*. doi: 10.7208/chicago/9780226736877.001.0001.

Schielzeth, H. *et al.* (2020) ‘Robustness of linear mixed-effects models to violations of distributional assumptions’, *Methods in Ecology and Evolution*, 11(9), pp. 1141–1152. doi: 10.1111/2041-210X.13434.

Schofield, M. R. *et al.* (2021) ‘On the robustness of latent class models for diagnostic testing with no gold standard’, *Statistics in Medicine*, 40(22), pp. 4751–4763. doi: 10.1002/sim.8999.

van de Schoot, R. *et al.* (2021) ‘Bayesian statistics and modelling’, *Nature Reviews Methods Primers*, 1(1), pp. 1–26. doi: 10.1038/s43586-020-00001-2.

Scott, J. and Marshall, G. (2009) *A Dictionary of Sociology*, *A Dictionary of Sociology*. Oxford University Press. doi: 10.1093/acref/9780199533008.001.0001.

Sharkey, K. J. (2008) 'Deterministic epidemiological models at the individual level', *Journal of Mathematical Biology*, 57(3), pp. 311–331. doi: 10.1007/s00285-008-0161-7.

Shinkins, B. and Perera, R. (2013) 'Diagnostic uncertainty: Dichotomies are not the answer', *British Journal of General Practice*, pp. 122–123. doi: 10.3399/bjgp13X664090.

Shreffler, J. and Huecker, M. R. (2020) 'Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios', *StatPearls*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/32491423 (Accessed: 28 March 2023).

Siegel, S. and Castellan, N. J. J. (1988) *Nonparametric statistics for the behavioral sciences, 2nd ed.* Mcgraw-Hill Book Company.

Silk, M. J. *et al.* (2017) 'The application of statistical network models in disease research', *Methods in Ecology and Evolution*, 8(9), pp. 1026–1041. doi: 10.1111/2041-210X.12770.

Simmons, M. P. *et al.* (2006) 'The relative performance of Bayesian and parsimony approaches when sampling characters evolving under homogeneous and heterogeneous sets of parameters', *Cladistics*, 22(2), pp. 171–185. doi: 10.1111/j.1096-0031.2006.00098.x.

Simpson, E. H. (1951) 'The Interpretation of Interaction in Contingency Tables', *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2), pp.

238–241. doi: 10.1111/j.2517-6161.1951.tb00088.x.

Smith, G. C. and Cheeseman, C. L. (2007) ‘Efficacy of trapping during the initial proactive culls in the randomised badger culling trial’, *Veterinary Record*. British Veterinary Association, pp. 723–726. doi: 10.1136/vr.160.21.723.

Smith, L. H. and Vanderweele, T. J. (2019) ‘Bounding Bias Due to Selection’, *Epidemiology*, 30(4), pp. 509–516. doi: 10.1097/EDE.0000000000001032.

Snow, J. (1856) ‘On the Mode of Communication of Cholera’, *Edinburgh Medical Journal*, 1(7), p. 668. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5307547/.

Spake, R. *et al.* (2022) ‘Improving quantitative synthesis to achieve generality in ecology’, *Nature Ecology and Evolution*, pp. 1818–1828. doi: 10.1038/s41559-022-01891-z.

Stevenson, M. *et al.* (2020) ‘epiR: Tools for the Analysis of Epidemiological Data’, *https://cran.r-project.org/web/packages/epiR/index.html*. Available at: https://cran.r-project.org/package=epiR (Accessed: 16 March 2023).

Stocks, T., Britton, T. and Höhle, M. (2021) ‘Model selection and parameter estimation for dynamic epidemic models via iterated filtering: Application to rotavirus in Germany’, *Biostatistics*, 21(3), pp. 400–416. doi: 10.1093/BIOSTATISTICS/KXY057.

Strayer, D. L. *et al.* (2003) ‘Effects of Land Cover on Stream Ecosystems: Roles of Empirical Models and Scaling Issues’, *Ecosystems*, 6(5), pp. 407–423. doi: 10.1007/pl00021506.

Strelioff, C. C. *et al.* (2013) ‘Inferring patterns of influenza transmission in swine from multiple streams of surveillance data’, *Proceedings of the Royal Society B:*

Biological Sciences, 280(1762), p. 20130872. doi: 10.1098/rspb.2013.0872.

Swets, J. A. (1988) ‘Measuring the accuracy of diagnostic systems’, *Science Science*, 240(4857), pp. 1285–1293. doi: 10.1126/science.3287615.

Swift, B. M. C. *et al.* (2021) ‘Tuberculosis in badgers where the bovine tuberculosis epidemic is expanding in cattle in England’, *Scientific Reports 2021 11:1*, 11(1), pp. 1–10. doi: 10.1038/s41598-021-00473-6.

Tabak, M. A., Pedersen, K. and Miller, R. S. (2019) ‘Detection error influences both temporal seroprevalence predictions and risk factors associations in wildlife disease models’, *Ecology and Evolution*, 9(18), pp. 10404–10414. doi: 10.1002/ece3.5558.

Taylor, L. R. (1961) ‘Aggregation, Variance and the Mean’, *Nature*, 189(4766), pp. 732–735. doi: 10.1038/189732a0.

Tian, H. *et al.* (2017) ‘Anthropogenically driven environmental changes shift the ecological dynamics of hemorrhagic fever with renal syndrome’, *PLoS Pathogens*, 13(1), p. e1006198. doi: 10.1371/journal.ppat.1006198.

Ting, Z. and Shaolin, P. (2008) ‘Spatial scale types and measurement of edge effects in ecology’, *Acta Ecologica Sinica*, 28(7), pp. 3322–3333. doi: 10.1016/S1872-2032(08)60071-2.

Toft, N. *et al.* (2007) ‘Evaluation of three serological tests for diagnosis of Maedi-Visna virus infection using latent class analysis’, *Veterinary Microbiology*, 120(1–2), pp. 77–86. doi: 10.1016/j.vetmic.2006.10.025.

Toft, N., Jørgensen, E. and Højsgaard, S. (2005) ‘Diagnosing diagnostic tests: Evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard’, in *Preventive Veterinary Medicine*.

Elsevier, pp. 19–33. doi: 10.1016/j.prevetmed.2005.01.006.

Tompkins, D. M. *et al.* (2011) ‘Wildlife diseases: From individuals to ecosystems’, *Journal of Animal Ecology*. John Wiley & Sons, Ltd, pp. 19–38. doi: 10.1111/j.1365-2656.2010.01742.x.

Tonnang, H. E. Z. *et al.* (2017) ‘Advances in crop insect modelling methods—Towards a whole system approach’, *Ecological Modelling*. Elsevier, pp. 88–103. doi: 10.1016/j.ecolmodel.2017.03.015.

Tredennick, A. T. *et al.* (2021) ‘A practical guide to selecting models for exploration, inference, and prediction in ecology’, *Ecology*, 102(6). doi: 10.1002/ecy.3336.

Tuncer, Y., Tanik, M. M. and Allison, D. B. (2008) ‘An overview of statistical decomposition techniques applied to complex systems’, *Computational Statistics and Data Analysis*, 52(5), pp. 2292–2310. doi: 10.1016/j.csda.2007.09.012.

Tuyttens, F. A. M. *et al.* (1999) ‘Differences in trappability of European badgers Meles meles in three populations in England’, *Journal of Applied Ecology*, 36(6), pp. 1051–1062. doi: 10.1046/j.1365-2664.1999.00462.x.

Upton, G. and Cook, I. (2014) *A Dictionary of Statistics*, *A Dictionary of Statistics*. Oxford University Press. doi: 10.1093/acref/9780199679188.001.0001.

Uusitalo, L. *et al.* (2015) ‘An overview of methods to evaluate uncertainty of deterministic models in decision support’, *Environmental Modelling and Software*, 63, pp. 24–31. doi: 10.1016/j.envsoft.2014.09.017.

VanderWaal, K. L. and Ezenwa, V. O. (2016) ‘Heterogeneity in pathogen

transmission: mechanisms and methodology', *Functional Ecology*, 30(10), pp. 1606–1622. doi: 10.1111/1365-2435.12645.

Varela-Castro, L. *et al.* (2021) 'A long-term survey on Mycobacterium tuberculosis complex in wild mammals from a bovine tuberculosis low prevalence area', *European Journal of Wildlife Research*, 67(3), pp. 1–8. doi: 10.1007/s10344-021-01489-z.

Vaseghi, S. V. (2008) *Advanced Digital Signal Processing and Noise Reduction: Fourth Edition*, *Advanced Digital Signal Processing and Noise Reduction: Fourth Edition*. wiley. doi: 10.1002/9780470740156.

Vats, D., Flegal, J. M. and Jones, G. L. (2019) 'Multivariate output analysis for Markov chain Monte Carlo', *Biometrika*, 106(2), pp. 321–337. doi: 10.1093/biomet/asz002.

Vehtari, A. *et al.* (2020) 'Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data', *Journal of Machine Learning Research*, 21, pp. 1–53.

Vernon, I. *et al.* (2018) 'Bayesian uncertainty analysis for complex systems biology models: Emulation, global parameter searches and evaluation of gene functions', *BMC Systems Biology*, 12(1). doi: 10.1186/s12918-017-0484-3.

Vesely, A., Finos, L. and Goeman, J. J. (2021) 'Permutation-Based True Discovery Guarantee by Sum Tests', *Journal of the Royal Statistical Society Series B: Statistical Methodology*. doi: 10.1093/JRSSSB/QKAD019.

Viana, M. *et al.* (2015) 'Dynamics of a morbillivirus at the domestic -wildlife interface: Canine distemper virus in domestic dogs and lions', *Proceedings of the National Academy of Sciences of the United States of America*, 112(5), pp.

1464–1469. doi: 10.1073/pnas.1411623112.

Volodina, V. and Challenor, P. (2021) ‘The importance of uncertainty quantification in model reproducibility’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197). doi: 10.1098/rsta.2020.0071.

De Waele, V. *et al.* (2011) ‘Age-stratified Bayesian analysis to estimate sensitivity and specificity of four diagnostic tests for detection of Cryptosporidium oocysts in neonatal calves’, *Journal of Clinical Microbiology*, 49(1), pp. 76–84. doi: 10.1128/JCM.01424-10.

Wagner, H. M. (1995) ‘Global Sensitivity Analysis’, *Operations Research*, 43(6), pp. 948–969. doi: 10.1287/opre.43.6.948.

Waller, L. and Carlin, B. (2010) ‘Disease Mapping’, in *Chapman & Hall/CRC handbooks of modern statistical methods*. NIH Public Access, pp. 217–243. doi: 10.1201/9781420072884-c14.

Walter, S. D. and Irwig, L. M. (1988) ‘Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review’, *Journal of Clinical Epidemiology*, 41(9), pp. 923–937. doi: 10.1016/0895-4356(88)90110-2.

Waltner-Toews, D. (2017) ‘Zoonoses, one health and complexity: Wicked problems and constructive conflict’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1725), p. 20160171. doi: 10.1098/rstb.2016.0171.

Wang, C., Lin, X. and Nelson, K. P. (2020) ‘Bayesian hierarchical latent class models for estimating diagnostic accuracy’, *Statistical Methods in Medical Research*, 29(4), pp. 1112–1128. doi: 10.1177/0962280219852649.

Warton, D. I. and Hui, F. K. C. (2017) ‘The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)’, *Methods in Ecology and Evolution*, 8(11), pp. 1408–1414. doi: 10.1111/2041-210X.12843.

Warton, D. I., Wright, S. T. and Wang, Y. (2012) ‘Distance-based multivariate analyses confound location and dispersion effects’, *Methods in Ecology and Evolution*, 3(1), pp. 89–101. doi: 10.1111/j.2041-210X.2011.00127.x.

Watts, R. G. (2008) *Introduction to Perturbation Methods*. New York, NY: Springer New York (Texts in Applied Mathematics). doi: 10.1007/978-3-031-79300-4_10.

Weber, N. *et al.* (2013) ‘Badger social networks correlate with tuberculosis infection’, *Current Biology*, 23(20), pp. R915–R916. doi: 10.1016/j.cub.2013.09.011.

Webster, R. G. *et al.* (2005) ‘The spread of the H5N1 bird flu epidemic in Asia in 2004.’, *Archives of virology. Supplementum*, (19), pp. 117–129. doi: 10.1007/3-211-29981-5_10.

Wells, K. *et al.* (2017) ‘Infection of the fittest: devil facial tumour disease has greatest effect on individuals with highest reproductive output’, *Ecology Letters*, 20(6), pp. 770–778. doi: 10.1111/ele.12776.

White, L. A., Forester, J. D. and Craft, M. E. (2017) ‘Using contact networks to explore mechanisms of parasite transmission in wildlife’, *Biological Reviews*, 92(1), pp. 389–409. doi: 10.1111/brv.12236.

White, P. C. and Harris, S. (1995) ‘Bovine tuberculosis in badger (Meles meles) populations in southwest England: An assessment of past, present and possible

future control strategies using simulation modelling', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 349(1330), pp. 415–432. doi: 10.1098/rstb.1995.0127.

WHO (2020) *WHO Director-General's opening remarks at the media briefing on COVID-19 - 12 October 2020*, *World Health Organization*. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---12-october-2020 (Accessed: 16 March 2023).

WHO (2022) *Tuberculosis*, *https://www.who.int/news-room/fact-sheets/detail/tuberculosis*. Available at: https://www.who.int/news-room/fact-sheets/detail/tuberculosis (Accessed: 16 March 2023).

Wickham, H. (2011) *The split-apply-combine strategy for data analysis*, *Journal of Statistical Software*. doi: 10.18637/jss.v040.i01.

Wickham, H. (2014) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. doi: 10.18637/jss.v035.b01.

Wijesiri, B. *et al.* (2018) 'Use of surrogate indicators for the evaluation of potential health risks due to poor urban water quality: A Bayesian Network approach', *Environmental Pollution*, 233, pp. 655–661. doi: 10.1016/j.envpol.2017.10.076.

Wikle, C. K. (2003) 'Hierarchical Bayesian models for predicting the spread of ecological processes', *Ecology*, 84(6), pp. 1382–1394. doi: 10.1890/0012-9658(2003)084[1382:HBMFPT]2.0.CO;2.

Winkler, M. S. *et al.* (2021) 'Bridging animal and clinical research during SARS-CoV-2 pandemic: A new-old challenge', *EBioMedicine*, 66, p. 103291. doi:

10.1016/j.ebiom.2021.103291.

Wolfe, N. D., Dunavan, C. P. and Diamond, J. (2007) ‘Origins of major human infectious diseases’, *Nature*, 447(7142), pp. 279–283. doi: 10.1038/nature05775.

Woodroffe, R. *et al.* (2006) ‘Effects of culling on badger Meles meles spatial organization: Implications for the control of bovine tuberculosis’, *Journal of Applied Ecology*, 43(1), pp. 1–10. doi: 10.1111/j.1365-2664.2005.01144.x.

Wu, J. and David, J. L. (2002) ‘A spatially explicit hierarchical approach to modeling complex ecological systems: Theory and applications’, *Ecological Modelling*, 153(1–2), pp. 7–26. doi: 10.1016/S0304-3800(01)00499-9.

Wu, J. and Li, H. (2006) ‘Uncertainty analysis in ecological studies:An overview’, *Scaling and Uncertainty Analysis in Ecology: Methods and Applications*, pp. 45–66. doi: 10.1007/1-4020-4663-4_3.

Wu, Z. *et al.* (2016) ‘Partially latent class models for case-control studies of childhood pneumonia aetiology’, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 65(1), pp. 97–114. doi: 10.1111/rssc.12101.

Wu, Z. *et al.* (2021) ‘A Bayesian approach to restricted latent class models for scientifically structured clustering of multivariate binary outcomes’, *Biometrics*, 77(4), pp. 1431–1444. doi: 10.1111/biom.13388.

Xu, C. *et al.* (2004) ‘Sensitivity analysis in ecological modeling’, *Chinese Journal of Applied Ecology*, 15(6), pp. 1056–1062.

Yanai, R. D., See, C. R. and Campbell, J. L. (2018) ‘Current Practices in Reporting Uncertainty in Ecosystem Ecology’, *Ecosystems*, 21(5), pp. 971–981. doi: 10.1007/s10021-017-0197-x.

Yang, Y. and Atkinson, P. M. (2008) ‘Parameter exploration of the raster space activity bundle simulation’, *Journal of Geographical Systems*, 10(3), pp. 263–289. doi: 10.1007/s10109-008-0062-8.

Yates, K. L. *et al.* (2018) ‘Outstanding Challenges in the Transferability of Ecological Models’, *Trends in Ecology and Evolution*, 33(10), pp. 790–802. doi: 10.1016/j.tree.2018.08.001.

Zhuang, L. *et al.* (2013) ‘Multi-species SIR models from a dynamical Bayesian perspective’, *Theoretical Ecology*, 6(4), pp. 457–473. doi: 10.1007/s12080-013-0180-x.