Use of kernel embeddings in the context of spatial disease mapping



Rohan Arambepola

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

Abstract

Over the last 20 years, high resolution mapping of estimated disease risk has become an important tool for disease control, facilitating the efficient allocation of limited resources. Maps of malaria risk, in particular, are widely used for operational planning and informing public health policy. Contemporary maps are typically made using Bayesian geostatistical models, informed by prevalence information from national health surveys and remote-sensed environmental and socioeconomic covariates. Increasingly, additional sources of data on malaria burden are available, including incidence data from routine surveillance, as are modelled products that may explain variation in risk, such as high resolution population estimates and intervention and travel time surfaces. Using these data, there is potential to improve estimates of risk and apply other statistical tools to answer policy-relevant questions. I develop novel statistical methods with two main aims—to learn the causal relationships between spatial and epidemiological variables from observational data and to improve estimates of malaria risk by combining prevalence and incidence data. These methods frequently necessitate the use of non-parametric approaches to accurately model complex relationships and account for a variety of possible sampling distributions. I use non-parametric methods based on the theory of kernel embeddings, which are computationally efficient and remain effective in high dimensions.

Learning causal relationships using spatial and epidemiological data is challenging for a number of reasons. Many existing causal inference meth-

ods rely on restrictive parametric assumptions that may not be satisfied in these cases. These data also include spatiotemporal structure that can confound relationships. Furthermore, variables are often observed at different resolutions, making comparisons difficult without aggregating to the lowest observed resolution. Such aggregation is an inefficient use of information and can interfere with learning fine-scale relationships. I develop a causal discovery procedure that combines constraint-based causal discovery algorithms, kernel-based independence tests and random Fourier feature approximations. I extend the theory of kernel-based tests to allow for variables observed at different resolutions. The result is a scalable, non-parametric causal discovery method that can be applied to data that includes spatiotemporal structure, non-linear relationships, and variables observed on different spatial or temporal scales. This is applied as a model-free feature selection method ('causal feature selection') and shown to perform favourably in the case study of forecasting malaria incidence in Madagascar. In addition, I discuss the challenges of estimating sub-national variation in the effectiveness of malaria control methods from observational data and approaches for estimating these effects at the community and individual level.

I also develop a Bayesian modelling methodology for combining prevalence data from national health surveys with routine surveillance incidence data to improve estimates of prevalence and make these estimates at a higher temporal cadence. Health facility catchment populations and treatment-seeking behaviour are modelled based on travel time to health facilities and a modular structure is used to limit the impact of bias in the incidence data on the final results. This method is used to produce high resolution estimates of monthly malaria prevalence in Madagascar from 2013 to 2016.

Acknowledgements

The work in this thesis could not have been achieved without the help and support of many people. First, I would like to thank my supervisors Dr Ewan Cameron and Prof. Peter Gething for all your support, advice and encouragement over the past four years. It has been a truly enjoyable experience to work with and get to know you and none of this work would have been possible without you. Also thank you to Prof. Christophe Fraser for taking over as my primary supervisor when MAP moved to Perth and welcoming me into your group.

I also want to thank everyone in MAP for being such a great group to be a part of. In particular, Tim, for answering any and all of my questions from day one; Kate Battle for all the help and always being so supportive; Ros, Suzanne, Emma and Kate Twohig for making all of the Madagascar work possible; Jen for the maps and many jokes that made little sense; and also Dan, Lisa, Harry, Colin, Joe and many others for all the coffee breaks, games nights and pub quizzes.

I am extremely grateful to the National Malaria Control Program, Ministry of Health, and Institut Pasteur of Madagascar and the Clinton Health Access Initiative for sharing the routine surveillance data that was the basis of many applications in this thesis. In particular I would like to thank Arsène Ratsimbasoa, Mauricette Andriamananjara, Saraha Rabeherisoa, Fanjasoa Rakotomanana and their teams for making this data sharing possible and for collaborating on the work mapping malaria prevalence

in Madagascar. I am also grateful to the Systems Biology DTC and St Cross College. My studies were funded by an Engineering and Physical Sciences Research Council studentship. Parts of the work in this thesis have been previously published [1, 2], where this is the case this has been clearly noted and all co-authors have given permission for this material to be reproduced.

Outside of my research, I would like to thank everyone who made my time at Oxford such a memorable experience. All of the SysBios for a great DTC year and beyond, especially Ravinda, Naomi, Lance and Jacob for all the squash and pints. Dylan, Amber, Krithika, Lance (again) and Kris for the parties at Divinity Road, dinners at St Cross, and that time we went to Chennai. Thom, Ben, Tom and Andrea for the too short time we were at Howard Street. Everyone in the Oxford Singers and Intermezzo, especially Beatrice for giving me many excuses to play the piano. Lily, for being my oldest friend in Oxford until you left! And to Oli and Cal for making the past few months of writing up and multiple lockdowns feel not so bad with the many great flat nights.

Finally, I would like to thank my parents, and my sisters Niri and Devi, for all their love and support, without which I would not be where I am.

Contents

Abstract **i**

Acknowledgements **iii**

List of Tables **ix**

List of Figures **x**

Abbreviations **xvi**

1 Introduction **1**

1.1 Causal inference 5

1.2 Kernel methods 11

2 Bayesian geostatistical models of malaria risk **15**

2.1 Malaria prevalence and incidence 17

2.2 Mapping prevalence in Madagascar 18

2.2.1 Data 19

2.2.2 Model specification 20

2.2.3 Sampling from the posterior 22

2.2.4 Results 26

2.3 Mapping incidence in Ayeyarwady, Myanmar 30

2.3.1 Background 30

2.3.2 Data 31

2.3.3 Model specification 33

2.3.4 Results 36

2.4 Mapping incidence from aggregated data in four states in India 38

2.4.1 Background 40

2.4.2 Data 41

2.4.3 Model specification 41
2.4.4 Results 44
2.5 Conclusion 47

3 Causal discovery with spatiotemporal data 48
3.1 Graphical models of causality 50
3.2 Causal discovery algorithms 54
3.2.1 Causal Markov condition 55
3.2.2 Faithfulness 56
3.2.3 d-separation 58
3.2.4 Score-based algorithms 65
3.2.5 Constraint-based algorithms 66
3.2.5.1 The SGS algorithm 67
3.2.5.2 The PC algorithm 69
3.2.6 Two-step discovery algorithm 74
3.3 Kernel-based independence tests 75
3.3.1 Reproducing kernel Hilbert spaces 76
3.3.2 Bochner's theorem 82
3.3.3 Kernel mean embeddings of distributions 89
3.3.4 Hilbert-Schmidt independence criterion 94
3.3.5 Randomized Conditional Independence Test 100
3.3.6 Independence tests with aggregated data 107
3.3.6.1 Random Fourier approximations with aggregated data 111
3.4 Spatiotemporal prewhitening 115
3.5 Causal discovery procedure 118

4 An updated map of prevalence in Madagascar informed by health survey and routine case data 121
4.1 Malaria in Madagascar 122
4.2 Risk mapping 124
4.3 Methods 128
4.3.1 Study data 128
4.3.2 Catchment population 129
4.3.3 Incidence model 133
4.3.4 Prevalence model 134
4.3.5 Causal feature selection 136
4.4 Results 138

4.4.1 Catchment populations . . . 138
4.4.2 Incidence . . . 138
4.4.3 Prevalence . . . 139
4.5 Discussion . . . 147
4.5.1 Limitations . . . 152
4.6 Conclusion . . . 154

5 Causal feature selection for forecasting incidence in Madagascar 157
5.1 Methods . . . 158
5.1.1 Modelling incidence . . . 158
5.1.1.1 Catchment model . . . 159
5.1.1.2 Disaggregation regression model . . . 160
5.1.1.3 Forward temporal estimation task . . . 163
5.1.2 Causal selection . . . 164
5.1.2.1 Prewhitening . . . 165
5.1.2.2 Aggregated independence tests . . . 166
5.1.3 Spike-and-slab regression . . . 167
5.1.4 Data . . . 169
5.2 Results . . . 172
5.2.1 Feature importance . . . 172
5.2.2 Model performance . . . 173
5.3 Discussion . . . 179
5.3.1 Limitations . . . 183
5.4 Conclusion . . . 184

6 Evaluating the effectiveness of malaria control methods: challenges and opportunities 186
6.1 Average and individual treatment effects . . . 188
6.2 Effect of malaria interventions . . . 195
6.2.1 Simulated data example 1: the impact of unobserved confounders 198
6.3 Increasing model flexibility . . . 209
6.3.1 Simulated data example 2: treatment effect interactions . . . 211
6.3.2 Gaussian process models . . . 218
6.4 Estimating individual-level effects . . . 225
6.5 Model validation . . . 228
6.6 Conclusion . . . 233

7 Conclusion **235**

Appendix **241**

Bibliography **247**

List of Tables

2.1 Mean and marginal standard deviations for the posterior distributions of model parameters when calculated using the No-U-Turn Sampler and a normal approximation (reproduced from Table 1, [68]). 26

2.2 Mean and standard deviation for the model parameters in the model of prevalence in Madagascar. 29

2.3 Mean and standard deviation for the coefficients for the occurrence of transmission surface in the model of incidence in Ayeyarwady. 39

2.4 Mean and standard deviation for the coefficients for the negative binomial rate surface in the model of incidence in Ayeyarwady. 39

2.5 Mean and standard deviation of the coefficients in the *P. vivax* model for the four states in India. 45

2.6 Mean and standard deviation of the coefficients in the *P. falciparum* model for the four states in India. 45

4.1 List of covariates. 130

4.2 Mean and 95% credible intervals for the prevalence model parameters. 141

6.1 True covariate coefficients from the data generating process and posterior mean values in the three model fits. 205

List of Figures

1.1 Observed points, shown in red with label 1 and blue with label 0. . . 13

2.1 Parasite prevalence rates in different clusters in the 2016 MIS in Madagascar. . . . 19
2.2 Standardised covariates used to inform prevalence estimates in Madagascar. . . . 21
2.3 Estimated *P. falciparum* prevalence for individuals between 6 and 59 months of age in Madagascar in 2016. . . . 27
2.4 The predicted values of the (a) Gaussian process, (b) linear contribution of covariates and (c) logit prevalence in the Madagascar prevalence model. . . . 28
2.5 Lower (2.5%, left) and upper (97.5%, right) quantiles of the posterior distribution of the estimates of prevalence in Madagascar in 2016. . . 29
2.6 MAP estimates of incidence of *P. falciparum* (left) and *P. vivax* (right) malaria in Myanmar in 2017 [29, 27, 74]. Figure made by Jennifer Rozier. 31
2.7 Average (over 2016 and 2017) number of confirmed cases per 1,000 people each year at health facilities in Ayeyarwady. Figure made by Katherine Battle. . . . 32
2.8 (a) Predicted probability of transmission and (b) predicted incidence rate (averaged over 2016 and 2017) in Ayeyarwady. Figure made by Katherine Battle. . . . 37
2.9 MAP estimates of incidence of *P. falciparum* (top) and *P. vivax* (bottom) malaria in India in 2015 [29, 27, 74]. Figure made by Jennifer Rozier. . . . 42
2.10 Observed incidence rates for *P. falciparum* (top) and *P. vivax* (bottom) in the four states in the study area in 2015. Figure made by Katherine Battle. . . . 43

2.11 Predicted incidence rates for *P. falciparum* (top) and *P. vivax* (bottom) in the four states in the study area in 2015. Figure made by Katherine Battle. 46

3.1 An example of a DAG. 53
3.2 An example of a PDAG. 54
3.3 An example of a DAG with three paths between X and Y, none of which contain a collider. There are two directed paths between X and Y and one which contains a common cause, Z_5, of X and Y. 60
3.4 Corresponding DAG for collider bias example. 61
3.5 Illustration of the negative correlation induced by only considering the subset of the population (shown in red) with sporting or academic ability above certain thresholds. The rest of the population is shown in grey. 62
3.6 Mock data (black points) fit using kernel ridge regression with a linear kernel (red), polynomial kernel with $c = 2, d = 2$ (green), polynomial kernel with $c = 1, d = 5$ (blue) and radial basis function kernel with $\sigma = 1$ (purple). 81
3.7 The embedding of $x = 0$ when using a radial basis function kernel ($\sigma = 1$) is shown by the black line. The dashed lines show approximations to this function when using different numbers of weights (different values of D), with the different colours representing different random samples of weights. 87
3.8 Estimates $\mu[\mathbf{x}]$ (black line) compared to the true function μ_P (grey dashed line) samples of different sizes, N, from a standard normal distribution when using a radial basis function kernel ($\sigma = 1$). The sample values are shown as red points. 92
3.9 Estimates $\mu[\mathbf{x}]$ (black line) compared to the true function μ_P (grey dashed line) samples of different sizes, N, from a standard normal distribution when using a Laplace kernel ($\sigma = 1$). The sample values are shown as red points. 92
3.10 (a) Joint observations of variables Z and Y (black points) and Gaussian process regression of Y on Z (red line). (b) Residuals r_{YZ} from the Gaussian process regression of Y against Z, plotted against Z. 101

3.11 Density of correlations between independent variables when both have spatial structure (green), one has spatial structure and the other does not (blue) or neither have spatial structure (red). 117

4.1 (a) Ecozones defined by Howes et al. [54] representing contiguous areas with distinct patterns of transmission. (b) Prevalence rates at survey sites in the 2013 and 2016 MIS. 124
4.2 Locally smoothed estimates of treatment-seeking rates (at public, private and any facility) against travel time to the nearest health facility from the 2013 and 2016 MISs and the relationship used in the catchment model. 132
4.3 Annual incidence rates at each health facility based on routine case data and modelled catchment populations. 138
4.4 Monthly incidence rates at each health facility based on routine case data and modelled catchment populations stratified by ecozone (2013-2016). 140
4.5 Incidence surfaces for all ages aggregated annually. 141
4.6 Prevalence estimates for individuals between 6 and 59 months of age, (a) aggregated annually and (b) population-weighted mean over time with 95% credible intervals. 143
4.7 Population-weighted mean prevalence over time stratified by ecozone with 95% credible interval. 145
4.8 Prevalence estimates for 2016 with ecozones borders shown in black. . 146
4.9 Number of months in 2016 with estimated prevalence (a) below 5% and (b) above 20%. 146
4.10 Annual incidence surfaces using three alternative catchment population estimates: estimates from the catchment model assuming higher treatment-seeking (top), estimates from the catchment model assuming lower treatment-seeking (middle) and estimates supplied by the NMCP (bottom). 148
4.11 Annual prevalence estimates using three alternative catchment population estimates: estimates from the catchment model assuming higher treatment-seeking (top), estimates from the catchment model assuming lower treatment-seeking (middle) and estimates supplied by the NMCP (bottom). 149

4.12 Uncertainty in 2016 annual prevalence estimates expressed with (a) interquartile range, (b) probability of prevalence exceeding 0.15 and (c) probability of prevalence not exceeding 0.05. 150

5.1 Geolocated health facilities. 170
5.2 Proportion of health facilities in each district that could not be precisely geolocated. 171
5.3 Feature rankings by importance (quantified by the proportion of repeats in which that variable was a direct cause of incidence) according to the causal inference approach, across each of the thirteen data iterations. 173
5.4 Feature rankings by importance (quantified by the marginal posterior inclusion probability) according to the linear spike-and-slab model, across each of the thirteen data iterations. 174
5.5 Feature rankings by importance (quantified by the marginal posterior inclusion probability) according to the Gaussian process spike-and-slab model, across each of the thirteen data iterations. 175
5.6 Comparison of correlation between predicted and true incidence rates for the linear (left) and Gaussian process (right) models when using spike-and-slab and causal selection. 176
5.7 Comparison of correlation between predicted and true incidence rates for linear (left) and Gaussian process (right) models when using all features (no selection) and causal selection. 177
5.8 Comparison of temporal correlation for the linear (top) and Gaussian process (bottom) models when using spike-and-slab and causal selection. The time-series considered are all 24 months forecasted (left) and months 13-24 (right). 178
5.9 Comparison of temporal correlation for the linear (top) and Gaussian process (bottom) models when using all features (no selection) and causal selection. The time-series considered are all 24 months forecasted (left) and months 13-24 (right) 179

6.1 Covariates used for simulated prevalence datasets. 199
6.2 Observed prevalence (left) and IRS (right) in simulated data example 1. For IRS distribution, blue points represents locations where IRS was used and red where it was not. 202

6.3 A DAG illustrating the causal relationships between variables in simulated data example 1. Here 'elev' and 'pr' represent elevation and prevalence (observed in a survey), respectively. 204
6.4 (a) True effect of IRS on prevalence, and predicted effect when the model was fit with (b) all covariates, (c) without rainfall, and (d) without accessibility. 206
6.5 Observed and predicted prevalence at validation locations in simulated data example 1 when the model was fit with (a) all covariates, (b) without rainfall, and (c) without accessibility. 209
6.6 A DAG illustrating the causal relationships between variables in the second simulated data example. Here 'elev' and 'pr' represent elevation and prevalence (observed in a survey), respectively. 212
6.7 Observed prevalence (left) and IRS (right) in scenario 1, simulated data example 2. For IRS distribution, blue points represents locations where IRS was used and red where it was not. 215
6.8 Observed prevalence (left) and IRS (right) in scenario 2, simulated data example 2. For IRS distribution, blue points represents locations where IRS was used and red where it was not. 216
6.9 True (left) and predicted (right) effect of IRS in scenarios 1 and 2 of simulated data example 2. 219
6.10 True (left) and predicted (right) effect of IRS when the data from scenarios 1 and 2 in simulated data example 2 were instead fit with the Gaussian process model. 221
6.11 Proportion of individuals with access to an ITN in the 2013 MIS in Madagascar [178] 222
6.12 True (left) and predicted (right) effect of IRS when the data from scenarios 1 and 2 in simulated data example 2 were instead fit with the Gaussian process model with additional covariates. 224
6.13 A DAG illustrating the concept of a negative control. The relationship of interest is the effect of T on Y, and Y' is a negative control. X are observed confounders that are controlled for and U are unobserved confounding variables that may or may not exist. If a non-zero causal effect of T on Y' is inferred it suggests variables U may exist. 231
6.14 A DAG illustrating the causal relationships required to use the instrumental variables method. Here U represents unobserved confounders which may exist. 232

A1 Monthly interquartile ranges. 243
A2 Probability of prevalence exceeding 0.15 each month. 244
A3 Probability of prevalence not exceeding 0.05 each month. 245
A4 Comparison of root mean square error between predicted and true incidence rates for the linear (top) and Gaussian process (bottom) models when using spike-and-slab and causal selection. 246
A5 Comparison of root mean square error between predicted and true incidence rates for the linear (top) and Gaussian process (bottom) models when using no selection and causal selection. 246

Abbreviations

Abbreviation	Term
ACT	artemisinin-based combination therapy
AI	aridity index
CHAI	Clinton Health Access Initiative
CI	credible interval
CPDAG	completed partially directed acyclic graph
DAG	directed acyclic graph
DHS	Demographic and Health Surveys
EVI	enhanced vegetation index
FCI algorithm	Fast Causal Inference algorithm
GMRF	Gaussian Markov random field
GP	Gaussian process
HMC	Hamiltonian Monte Carlo
HMIS	health management information system
HSIC	Hilbert-Schmidt Independence Criterion
INLA	Integrated Nested Laplace Approximation
IPTp	intermittent preventive treatment in pregnancy
IRS	indoor residual spraying
ITN	insecticide-treated bednet
KCIT	Kernel Conditional Independence Test
LQ	lower quartile
LST	land surface temperature

MAP	Malaria Atlas Project
MDC	mass distribution campaign
MICS	Multiple Indicator Cluster Survey
MIS	Malaria Indicator Survey
NMCP	National Malaria Control Programme of Madagascar
PDAG	partially directed acyclic graph
PET	potential evapotranspiration
RCIT	Randomized Conditional Independence Test
RCT	randomised controlled trial
RDT	rapid diagnostic test
RESIT	regression with subsequent independence test
RFF	random Fourier feature
RKHS	reproducing kernel Hilbert space
SD	standard deviation
SPDE	stochastic partial differential equation
TCB	tasselled cap brightness
TCW	tasselled cap wetness
TSI	temperature suitability index
TWI	topographic wetness index
UQ	upper quartile

Chapter 1

Introduction

There is a long history of mapping of disease incidence and prevalence around the world in order to understand the geographical distribution of a disease or identify risk factors [3]. In the nineteenth century, interest in disease mapping was often focused on infectious diseases, such as yellow fever in the United States and cholera in Europe [4, 5]. Maps of the latter included John Snow's famous work on the Broad Street outbreak in London, which demonstrated that the disease was spreading through contaminated water [6]. Efforts to map chronic diseases increased during the twentieth century, particularly mapping cancer in the United Kingdom and United States [7, 8, 9]. Mapping malaria risk, the focus of this thesis, was also of increasing interest around this time [10]. In 1968, Lysenko and Semashko [11] produced a global map of malaria endemicity stratified into four classes (hypoendemic, mesoendemic, hyperendemic and holoendemic), with some temperate areas noted additionally as 'epidemic'. This work was a qualitative synthesis of historical records and previous maps of malaria metrics (including parasite rate, sickle cell incidence and vector distributions). In the following decades, most maps of malaria risk were produced at the country, rather than global, level [10, 12]. Country-level maps of raw malaria metrics (such as case counts from health facilities, prevalence data from household surveys, and information on vector abundance) are still commonly used to inform operational planning [13, 14, 15, 16]. This visualisation of raw epidemiological or

entomological data is an important first step in characterising spatiotemporal patterns in malaria risk in a given area, but such maps have a number of limitations. Incidence data, for example, is often in the form of observed case counts from health facilities and may suffer from biases and incompleteness due to low treatment-seeking rates and incomplete record keeping (amongst other factors) [17]. Information on parasite prevalence collected as part of health surveys may avoid some of these sources of bias but are likely to have lower spatial coverage and occur relatively infrequently, resulting in uncertainty when extrapolating these results to other areas or time points. Moreover, this uncertainty is difficult to quantify and is unlikely to be reflected in the maps themselves. Relying on raw data is also problematic in pre-elimination settings, where the low transmission rates lead to a very noisy sampling distribution for standard malaria metrics [18].

Partly in response to these issues, the last 20 years has seen the rise of mapping disease risk using Bayesian geostatistical models, mostly using and building on the model based geostatistics framework pioneered by Diggle and others [19, 20]. In this framework, the response variable (such as prevalence or incidence of a disease) is modelled as a function of a spatially autocorrelated process and explanatory variables. A number of examples of these models are described in Chapter 2. In the context of malaria risk mapping, this has been made possible by the increasing availability of malariometric data and high resolution satellite image-based measurements of environmental and socioeconomic variables. These environmental and socioeconomic variables (such as rainfall, temperature, population density and travel time to cities) are typically available globally on a fine grid of pixels (often 1km-by-1km or 5km-by-5km in size) and, where applicable, at a high temporal resolution (often weekly or monthly) [21]. Initially, the main source of information on malaria risk informing these models was parasite prevalence data, primarily from national cross-sectional household surveys carried out since the late 1980s in many malaria-endemic countries

(mostly in sub-Saharan Africa) as part of the UNICEF Multiple Indicator Cluster Survey (MICS) [22] and Demographic and Health Surveys (DHS) Program [23] (as well as research projects and smaller surveys). In a landmark study harnessing these data types in 2009, Hay et al. [24] used a geostatistical model to produce a global characterisation of *Plasmodium falciparum* endemicity, updating the Lysenko and Semashko map for the start of the 21st century. Compared to country-level maps of raw data, these results benefited from the sharing of information between countries, statistical shrinkage and regularisation, and improved estimates further from observed locations due to the use of explanatory covariates. Moreover, the uncertainty of these estimates was quantified using the posterior credible interval. In 2013, Bhatt et al. [25] used a similar approach to map the global endemicity of dengue. More recently, incidence data collected at health facilities has been the main source of information on malaria transmission outside of sub-Saharan Africa. These data were used in the construction of global maps of *P. falciparum* in 2015 and 2019 [26, 27] and *Plasmodium vivax* in 2012 and 2019 [28, 29]. Increasingly, incidence information is also available within sub-Saharan Africa but the sources of bias previously discussed, such as incompleteness and heterogeneous rates of treatment seeking, can make including these data in risk mapping models challenging [17].

In this thesis, I build on this previous risk mapping work with a focus on advancing the statistical methods for learning from the ever increasing suite of satellite-based covariates. My main contribution is the development of a framework for performing causal discovery when using spatial data. To do this, I make use of kernel-based non-parametric independence tests to apply established causal discovery algorithms to these sources of data. I also address the problems of inferring causation when the variables have spatiotemporal structure and testing for independence when variables have been observed at different resolutions. I investigate the utility of this causal discovery procedure as a feature selection tool (which I term 'causal feature selec-

tion') when mapping malaria incidence and prevalence. I also present a method for incorporating prevalence and incidence information within the same model to address some of the biases common in routine surveillance data. This method is used to make monthly maps of malaria prevalence in Madagascar between 2013 and 2016.

In Chapter 2, I begin by describing three malaria risk mapping problems. These examples establish further context for the work in the rest of the thesis, as well as introducing several concepts that are used in more complex models in the following chapters. Two of these examples—mapping incidence in regions of India and Myanmar—were risk mapping projects to guide read-world decision-making around resource allocation that I undertook in collaboration with the Clinton Health Access Initiative (CHAI). Chapter 3 introduces the mathematical background on causal discovery algorithms and kernel-based independence tests. I extend this theory to allow for independence testing between variables observed at different resolutions and then bring together the concepts discussed in this chapter to describe the causal discovery procedure for spatiotemporal data. In Chapter 4, I develop a method for combining incidence and prevalence data to map prevalence at a high temporal resolution while minimising the effect of bias in the incidence data and apply this method to map prevalence in Madagascar. Within this model, I also use the causal discovery procedure for causal feature selection. In Chapter 5, I compare the use of causal feature selection to a classical feature selection method (spike-and-slab regression) and the alternative of no feature selection when forecasting malaria incidence in Madagascar. This incidence model involves the representation of case occurrence on a fine-scale and accounting for variable rates of treatment-seeking and unknown catchment areas (which were partly learned within the model). Finally, Chapter 6 focuses on the use of causal inference to infer the effectiveness of malaria control methods in different settings. I discuss the challenges faced when estimating heterogeneity in effectiveness and methods that may address some of these problems. In the remainder of

this chapter, I briefly introduce the concepts of causal inference and kernel methods, which are covered in more detail in Chapter 3.

1.1 Causal inference

Many important questions in the field of public health are fundamentally about causal, rather than associative, relationships between different variables. For example, when we talk about driving factors of disease transmission, we usually mean variables that *cause* transmission. Similarly, evaluating the impact of an intervention involves distinguishing causal effects from associations. The most reliable way to answer causal questions is through experiments or randomised controlled trials (RCTs), but these are not always possible, practical, or ethical. Causal inference methods use "observational data"—here meaning retrospectively or opportunistically gathered, rather than designed experimental or RCT data—to infer causal structures or effects. Inferring causal relationships from observational data is known to be a difficult problem ("correlation does not imply causation"). This is due to two main sources of error—confounding and collider bias. Confounding occurs when two variables of interest have one or more common causes. These common causes are called *confounding variables* or *confounders*. If the relationship between two variables is confounded then the association between these variables observed in the data could be partly or entirely due to confounding variables. For example, suppose we are interested in the effect of drinking a glass of red wine a day on an individual's risk of heart disease and we observe in a sample of the population that, on average, the individuals who drink a glass of red wine a day are less likely to suffer from heart disease. From this we might hypothesise that drinking red wine reduces the likelihood of heart disease and that if the entire population drank a glass of red wine a day the incidence of heart disease would decrease. However, it may in fact be the case that this relationship

is confounded by socioeconomic status, and that higher socioeconomic status causes an individual to be more likely to drink red wine and, through improved access to healthcare, less likely to suffer from heart disease. In this case, red wine could decrease the risk of heart disease but it could also have no effect or even increase risk (to a degree outweighed by the socioeconomic effect). From joint observations of red wine consumption and heart disease status, it would be impossible to separate the true causal effect from the effect of the confounding variable on red wine consumption and heart disease. In contrast, in a hypothetical clinical trial designed to measure such an effect, individuals could be randomly assigned to a treatment group, who are asked to drink a glass of red wine a day, and a control group, who are not. In such a design, neither socioeconomic status nor any other variable would affect who drinks wine and therefore there would be no confounding factors, and any association between drinking wine and heart disease could be interpreted as a causal effect. When considering malaria control efforts, many interventions are targeted at high transmission areas. As a result, the use of the intervention and the strength of transmission (some time after when the intervention is expected to have had an effect) may appear positively correlated, even for an effective intervention. In this case, pre-intervention risk is a confounding variable, as it has a causal effect on intervention distribution and post-intervention transmission.

Collider bias refers to the phenomenon where conditioning on a common effect of two variables induces an association between these variables, even when no association exists in the unconditional distribution. In many cases this is known as selection bias (although collider bias is more general, as conditioning on a common effect is not always an obvious example of 'selection'). For example, in patients treated at health facilities in a region, it might be observed that there is a positive correlation between the distance travelled to the facility and disease severity. This could suggest that there is a higher incidence of severe disease in more remote areas. However, even if

there was no spatial variation in disease severity, this trend could still occur if people were more likely to seek treatment if they lived close to a facility or if the disease was more severe. In this case, disease severity and distance to a health facility both have a causal effect on the likelihood of an individual seeking treatment at a health facility. Therefore, considering only individuals who attended a facility is equivalent to selecting or conditioning on this common effect, which induces a positive association between severity and distance to a facility. Collider bias can occur in RCTs based on the enrolment process for participants but it is generally easier to control this selection process compared to when using observational data.

Accounting for these sources of bias can be difficult. While measured confounding variables can be included in an analysis, unmeasured confounding variables can still bias results and in general it is not possible to detect the existence of such unmeasured confounding variables from observational data alone. Removing the effect of collider bias often requires a detailed understanding of the data collection process, which may not always be available. Moreover, accounting for confounding variables typically involves conditioning on these variables, but conditioning on the wrong variables could inadvertently induce collider bias. While these factors suggest we should be cautious when inferring causal relationships from observational data, it has been argued that many 'causation blind' statistical analyses ultimately aim to answer causal questions (whether acknowledged explicitly or not) and therefore it is better to attempt to understand and account for causal structure than to ignore it [30, 31]. Examples of causal questions in malaria risk mapping include the work of Bhatt et al. [26] on estimating the effect of different interventions on malaria burden between 2000 and 2015. Estimating these effects required counterfactual predictions of what prevalence would have been in the absence of these interventions. Similarly, recent work by my research group on the indirect effects of the COVID-19 pandemic on malaria morbidity and mortality involved predicting the causal effect of reduced levels of

control measures [32]. Both of these analyses employed geospatial models.

Defining the concept of causation is not straightforward and has been the subject of much philosophical debate [33]. Here I introduce three common approaches from the field of causal inference. The first is the 'target trial' definition of a causal effect proposed by Hernán and Robins [30, 34], which is based on the Potential Outcomes framework of Neyman and Rubin [35, 36]. Here the causal effect of a variable X on variable Y is defined as the effect that would be observed in a randomised controlled trial in which the variable X was manipulated. For example, under this definition the causal effect of obesity on heart disease is not well defined, as it would not be possible to conduct a trial in which individuals were randomly assigned to be obese or not be obese. On the other hand, the causal effect of a specific diet and exercise regimen on heart disease could be defined. The target trial definition is therefore concrete and easily interpretable but cannot define some causal effects that may be of interest, such as the effect of race on health outcomes or geographical factors on malaria risk. Alternatively, Pearl and collaborators model causal relationships using structural equation models [37, 38, 31], which specify a functional form for each variable in terms of its causes. In this setting, a counterfactual outcome (and therefore a causal effect) can be obtained by modifying the variable of interest while holding all other variables (that are not caused by the variable of interest) constant. The interpretation of the counterfactual produced by this mechanical procedure is similar to the 'closest world' interpretation of causation [39]. In this interpretation, the counterfactual value of variable Y when the value of the variable X is set to x is defined by first considering every possible world in which the variable of X could have taken the value x. The counterfactual scenario is then defined to be the closest of these worlds to reality and the counterfactual value of Y is its value in this closest world. This definition acknowledges the fact that in a world where a variable takes on a different value, other differences may be necessary (for example, a counterfactual world in which a

red car is instead blue is a world in which the paint used on the car was blue instead of red, and so on) but the counterfactual scenario should strictly only have these necessary differences. Using the structural equation or closest world approaches, it is possible to define the causal effect of obesity on heart disease, although this causal effect will not necessarily be the same as the one defined through diet or exercise using the target trial approach. Finally, an alternative approach is to avoid defining the notions of causality or a causal effect at all but instead set out an agreed set of axioms which a notional causality is assumed to satisfy (in analogy to axiomatic treatments of geometry or probability, for example). This is the approach taken by Spirtes et al. [40]. They make the fundamental assumptions that the relation "A is a cause of B" is transitive, irreflexive, and antisymmetric. In other words, they assume that

1. if A causes B and B causes C, then A causes C,
2. any event cannot be a cause of itself, and
3. if A is a cause of B then B cannot be a cause of A.

Further assumptions can be used to connect the causal relationships between random variables to properties of their joint distribution (see Chapter 3).

Causal inference methods can be categorised into methods for inferring the causal structure between variables (discovering which variable is a cause of which others but not necessarily quantifying these relationships), often known as causal discovery, and methods for estimating causal effects (for example, in terms of effect sizes or functional relationships) when the causal structure is assumed or known. In this thesis I primarily focus on the former, although the latter is discussed in Chapter 6. Causal discovery methods aim to infer a graph that represents the causal structure of a set of random variables based on joint observations. In these graphs, nodes represent the different variables and a directed edge from variable X to variable Y is

included if X is determined to be a direct cause of Y (the notion of a direct cause is defined in Chapter 3). Causal discovery methods can be again divided into two broad classes, constraint-based and score-based algorithms [41]. Score-based algorithms associate each candidate graph with a model, which can then be fit to the observed data. Graphs are scored based on some measure of goodness of fit of these models (such as the Bayesian Information Criterion) and an efficient search algorithm is used to find a graph with a high (ideally the highest) score. A notable example is the Greedy Equivalence Search algorithm [42]. Score based methods naturally depend on the assumptions that are used to construct the model from each graph (for example, the functional relationship between each variable and its causes is often assumed to be linear) which limits their utility when relationships between variables are potentially complex and non-linear. Constraint-based algorithms infer independence and conditional independence relationships from the observed data and reject causal structures that are not consistent with these relationships. For example, if variables X and Y appear (unconditionally) independent then these algorithms may reject graphs in which X is a direct cause of Y (and vice versa). A number of algorithms have been proposed to efficiently identify the graph (or set of graphs) that is consistent with the independence relations learned from the data. While in theory these constraint-based algorithms require fewer assumptions than score-based methods, in practice commonly used conditional independence tests require similar strong parametric assumptions and are hence at risk of misspecification errors. However, as we will see, these algorithms can be made non-parametric by using non-parametric kernel-based independence tests. In my work, I focus on one of the most commonly used constraint-based algorithms, the PC algorithm [43, 44].

1.2 Kernel methods

Methods using kernel functions have been a popular tool in machine learning for addressing a variety of different problems in a non-parametric way. A kernel function on some set $\mathcal{X}$ is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. These functions can be thought of as similarity measures. For example, the radial basis function kernel on $\mathcal{X} = \mathbb{R}^n$ is defined as

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{1.1}$$

for any $x, y \in \mathbb{R}^n$, where $\sigma \in \mathbb{R}_{>0}$ is a fixed parameter. Here the value of this function is highest, representing greatest similarity, when the Euclidean distance between points is small. Initially these functions were used to transform linear algorithms (for problems such as regression, cluster analysis and classification) into non-parametric methods by replacing all inner products in the original algorithm with kernel function evaluations. The underlying theory is that this is equivalent to mapping the data points into a high (often infinite) dimensional Hilbert space in which the inner product of the image of points $x, y \in \mathcal{X}$ is equal to $k(x, y)$. Applying the linear algorithm in this high dimensional space is equivalent to a non-linear algorithm in the original space. In recent years, there has been increasing interest in embedding probability distributions, rather than individual data points, into these Hilbert spaces. Under certain conditions this embedding is injective and distributions can be compared by comparing their embeddings. This has led to a number of applications, including the independence tests that I make use of in this thesis [45, 46, 47]. Another important application of kernel methods is Gaussian process regression, which is a fundamental tool in spatial statistics and can be thought of as a kernel version of Bayesian linear regression (this is discussed in Section 3.3.2). Computing the posterior mean of a Gaussian process is also equivalent to kernel ridge regression, which is commonly used for non-parametric regression and used within one of the independence tests that I

implement in this thesis, the Randomized Conditional Independence Test [47].

To illustrate the utility of mapping data points into higher-dimensional spaces, consider the following example of a kernel support vector machine. Suppose we observe points in a two dimensional space, $(x_1, y_1), ..., (x_{200}, y_{200}) \in \mathbb{R}^2$, with labels $l_1, ..., l_{200}$ where $l_i \in \{0, 1\}$. These points with their labels are shown in Figure 1.1. The task is to find a boundary that can be used to classify future observations into one of these two groups. A conventional support vector machine tries to find a hyperplane (in this case a line) which separates the observed groups and which will accurately classify new observations. In this case, it is clear from the labels of the observed data that the classification appears to be based on the distance of each point from the origin and there is no line that will separate these two groups well. However, we can consider the map $\phi : \mathbb{R}^2 \to \mathbb{R}^3$,

$$\phi : (x, y) \longmapsto (x, y, z) = (x, y, x^2 + y^2).$$

We refer to ϕ as a *feature map* and $\mathbb{R}^3$ as the *feature space*. If we map the data points into $\mathbb{R}^3$ using ϕ, we can use a support vector machine in this new space. In this case, the algorithm will identify that the plane $z = 1$ separates the two groups as required. This plane corresponds to a separating *curve* in the original space, defined by $x^2 + y^2 = 1$. Therefore applying the map ϕ to the data allows us to apply a linear algorithm to produce a non-linear result.

This non-linear classification task can be framed in terms of kernel operations as follows. The support vector machine algorithm is built on an optimisation problem involving the evaluation of inner products in this new space, $\phi(x_i, y_i)^T \phi(x_j, y_j)$, for all pairs of observed points. This inner product can be used to define a kernel on

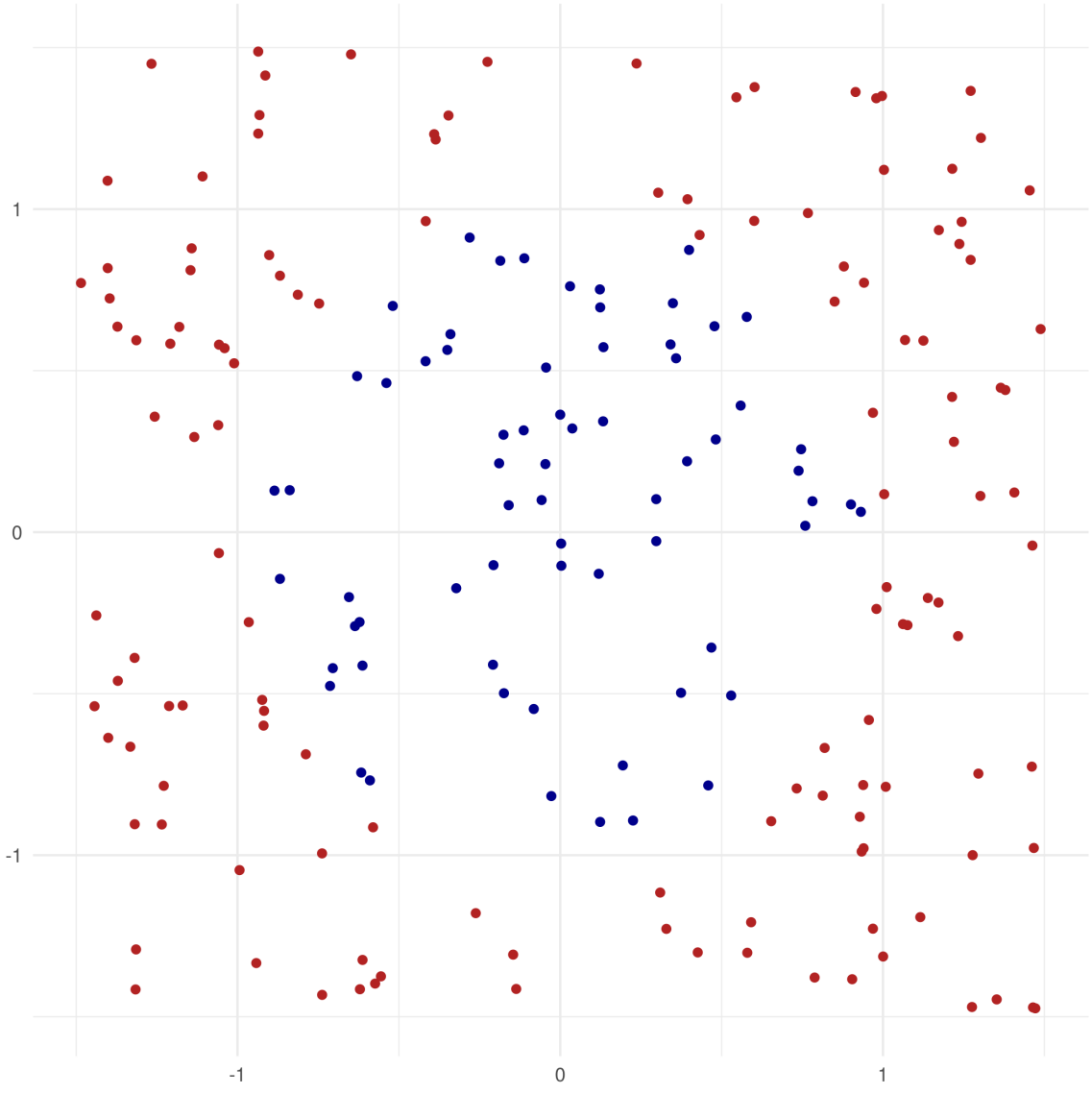


Figure 1.1: Observed points, shown in red with label 1 and blue with label 0.

$\mathcal{X} = \mathbb{R}^2$,

$$\begin{aligned} k((x_1, y_1), (x_2, y_2)) &= \phi(x_1, y_1)^T \phi(x_2, y_2) \\ &= x_1 x_2 + y_1 y_2 + (x_1^2 + y_1^2)(x_2^2 + y_2^2) \end{aligned}$$

for any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}$. This is a kernel as it is a symmetric function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}^2$. The process of generating a non-linear method by first mapping data points under ϕ and then applying the support vector machine algorithm can be reinterpreted as simply replacing all inner product evaluations in the linear algorithm with kernel function evaluations. This suggests this technique can be generalised by using arbitrary kernel functions on $\mathcal{X}$. These kernels may correspond to high or infinite dimensional feature spaces but since the algorithm can be expressed in terms of kernel evaluations, explicitly computing the image of data points under the feature map can be avoided. The use of these higher-dimensional and more expressive kernels, such as the radial basis function kernel, removes the need to hand tune the embedding as above and, by only using kernel evaluations, is computationally efficient.

A similar approach can be used for any linear algorithm that involves data points only through inner product evaluations. This has led to a wide array of non-linear 'kernalised' methods, including kernel ridge regression and kernel principal component analysis. The ability to perform all computations through the kernel function without having to evaluate the feature map is known as the 'kernel trick' [48].

Chapter 2

Bayesian geostatistical models of malaria risk

Many of the statistical models used for malaria risk mapping are based on the modern model-based geostatistics framework pioneered by Diggle and others [19, 20]. These models assume that there exists a smooth latent risk surface, which exhibits spatial autocorrelation, and that observations at survey locations are produced by a sampling distribution based on this underlying risk. For example, the number of cases of a disease occurring in a location over some period of time could be modelled with a Poisson sampling distribution possessing a rate parameter that depends on the risk at (or near) that location. The latent risk, η, is typically modelled as the sum of a mean function and a zero-mean Gaussian process, that is

$$\eta(s) = \mu(s) + f(s)$$

where μ is the mean function, f is a Gaussian process and $s \in \mathbb{R}^2$. This risk surface is conceptualised as continuous but is modelled computationally as a fine grid (for example 1km-by-1km or 5km-by-5km) of pixels. A link function connects the relevant parameter in the sampling distribution to the underlying risk, in a similar way to the relationship between the linear predictor and sampling distribution in generalised linear models. In the example of case counts, a log link function is usually used to

connect the rate parameter of the Poisson distribution to the latent risk surface. In the simplest case, the mean function is a constant and any spatial heterogeneity in risk is captured by the Gaussian process term. Subsequent predictions are the result of the regularisation and spatial smoothing of this process. In this case, contemporary geostatistical models often produce results similar to other spatial interpolation or smoothing techniques (collectively referred to as 'kriging' methods [19, 49]) but have the important advantages of weighting observations in a principled manner through the likelihood and quantifying the uncertainty of point estimates. More commonly, explanatory covariates are included to potentially explain some of the observed variation in risk. The inclusion of covariates that are associated with risk can reduce uncertainty and improve predictions in locations far from observations. The Gaussian process term then represents the remaining spatial structure, which can be thought of as the combined effect of factors not included in the explanatory variables. A common way to include covariates is to assume a linear effect on risk,

$$\mu(s) = \beta_0 + \beta^T X(s)$$

where β_0 and β are parameters and $X(s)$ are covariate values at location s. Such models are sometimes referred to as generalised linear geostatistical models (see Chapter 4 in [20] for more details). More complex behaviour can also be considered, such as non-linear effects and interactions between covariates (see Chapters 5 and 6).

In this chapter, I describe three models for mapping malaria risk in different contexts. The aim is to illustrate how a common modelling framework can be applied to different real-world risk mapping problems and to provide context for the methodological innovations in the following chapters. All of the models in the chapter were designed and implemented by myself. Some of the figures of final outputs were produced by other members of the Malaria Atlas Project for use in project reports, where

this is the case they are credited.

2.1 Malaria prevalence and incidence

In this thesis, geostatistical models are employed to map both malaria prevalence and incidence. It is useful to define how these terms are used in this context, which differs slightly from their more general definitions. In a general epidemiological context, the prevalence of a disease at some point in time refers to the proportion of the population suffering from the disease at that time point. Incidence of a disease, on the other hand, is the number of new cases of a disease that occur over a given time period, for example annual incidence is the number of new cases that occur each year. Incidence is often also standardised by the population, in which case it is the number of new cases per unit time per person. The relationship between these two quantities depends on the rate at which people recover or die due to the disease—for example, if this rate is smaller than the incidence rate then prevalence will increase.

prevalence

incidence

In a malaria-specific context, *prevalence* is defined as the proportion of the population with a detectable malaria infection at any given time, i.e. the proportion of the population with malaria parasites (at a level that can be detected) in their bloodstream. On the other hand, malaria *incidence* is the number of *clinical cases* of malaria occurring in a time period. A clinical case is an infection that leads to symptoms that would be diagnosed as malaria in a clinical context. The difference between this definition and the more general one is that not every infection is symptomatic, therefore the proportion of the population with an infection (what we refer to as prevalence of malaria) is usually higher than the proportion with an infection and clinical symptoms. These definitions are the result of how each metric is typically measured. Prevalence data are often generated in cross-sectional surveys where individuals are tested for infection. In contrast, incidence data are typically recorded

at health facilities and therefore only capture symptomatic cases for which treatment was sought. These sources of data are discussed in more detail in Chapter 4. The relationship between incidence and prevalence is known to be complex and depends on past and present transmission in an area and treatment levels. [50, 51, 52]. Having defined these terms, we can now consider the risk mapping examples.

2.2 Mapping prevalence in Madagascar

This example is chosen to illustrate a relatively straightforward risk mapping problem, using malaria prevalence data from a national health survey to make a smooth map of predicted prevalence. These surveys are designed to produce representative samples using a consistent methodology between survey sites, therefore avoiding issues of selection bias, incomplete reporting and inconsistent standards of diagnosis that can occur in other sources of data (see Chapter 4 for more details). These are household surveys and so typically take place in the area where the participants live and therefore this can be treated as point-level data. This is in contrast to incidence data, which is often an aggregation of cases over a larger geographical area. Here I used data from a national health survey conducted in 2016 and environmental and socioeconomic covariates within a spatial model to produce a map of estimated prevalence. Kang et al. [53] implemented a similar spatiotemporal model using data from multiple health surveys to predict prevalence in Madagascar in multiple years. I build on the work of Kang et al. [53] in Chapter 4 by incorporating malaria case data into these models to make monthly maps of prevalence between 2013 and 2016. More context for this work and interpretation of prevalence estimates can be found in Chapter 4 (and in [54, 53]).

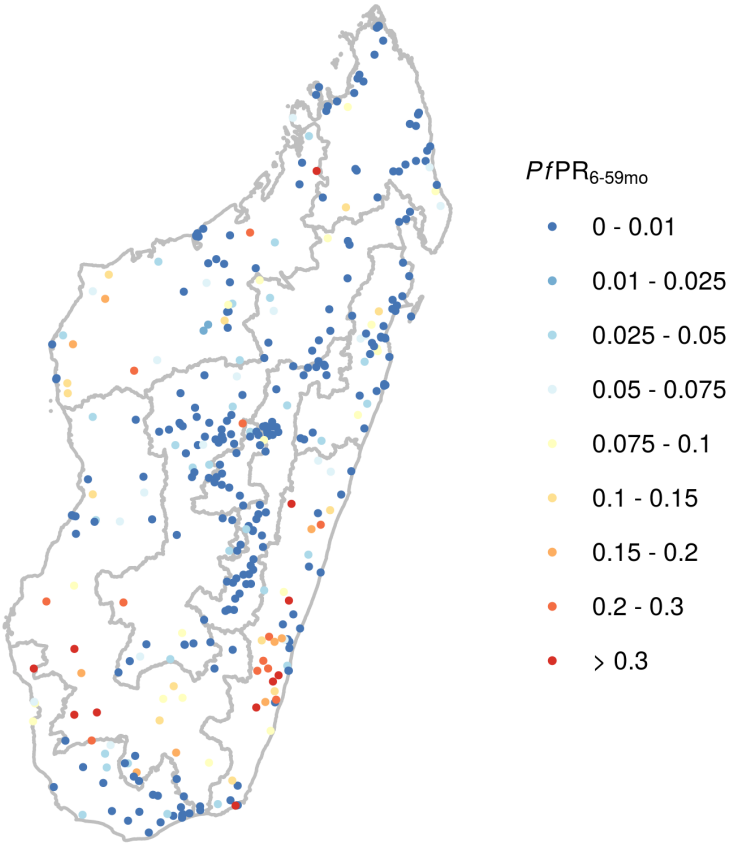


Figure 2.1: Parasite prevalence rates in different clusters in the 2016 MIS in Madagascar.

2.2.1 Data

The response data in this example come from the Malaria Indicator Survey (MIS) carried out in Madagascar in 2016 [55]. The indicator of interest here was the prevalence of *P. falciparum* infection in children between 6 and 59 months of age, determined by microscopy. For each child tested, the result is a binary variable representing presence or absence of infection. Figure 2.1 shows the proportion of children with an infection at each location (cluster) in the survey. The number of children examined in each cluster ranged from 2 to 53 with a mean of 19.3.

The covariates used to inform the model were elevation [56], accessibility to

cities [57] and distance to water [58]. Elevation may be informative as malaria transmission in Madagascar is typically (and has historically been) higher in the low-lying coastal regions than in the highlands in the centre of the country [59, 54]. Distance to water [58] is a measure of the distance to lakes, rivers and other wetland areas, which is likely to influence the presence of breeding areas for the *Anopheles* mosquitoes that transmit malaria [60, 61]. Finally, accessibility to cities measures the travel time to cities (defined as built up areas with a population centre of at least 50,000 inhabitants) and can be considered a proxy for urbanicity (or conversely remoteness). In sub-Saharan Africa, transmission is often higher in rural locations, potentially due to increased numbers of vector breeding sites and more limited access to healthcare [62, 63]. These covariates are satellite image-based products—elevation being directly measured by radar on dedicated satellites and accessibility and distance to water being modelled surfaces derived from images of landcover, surface water, roads and man-made structures—and were available on an approximately 1km-by-1km grid. The exact resolution of these products was measured in degrees and therefore the precise size of each pixel in kilometres depends on the latitude. Throughout this thesis, I refer to pixels approximately 1km-by-1km or 5km-by-5km in size, which correspond to square pixels of length 0.00833 and 0.0417 degrees, respectively. These conversions become less accurate further from the equator but are acceptable for the geographical areas considered here.

2.2.2 Model specification

Parasite prevalence data was available in the form of the number of individuals with an infection, Y_i, and the total number tested, N_i, at location s_i for surveys $i = 1, ..., 358$. The covariates were available on a grid of 1km-by-1km pixels and survey i was associated with the covariate values in the corresponding pixel, X_i. The number

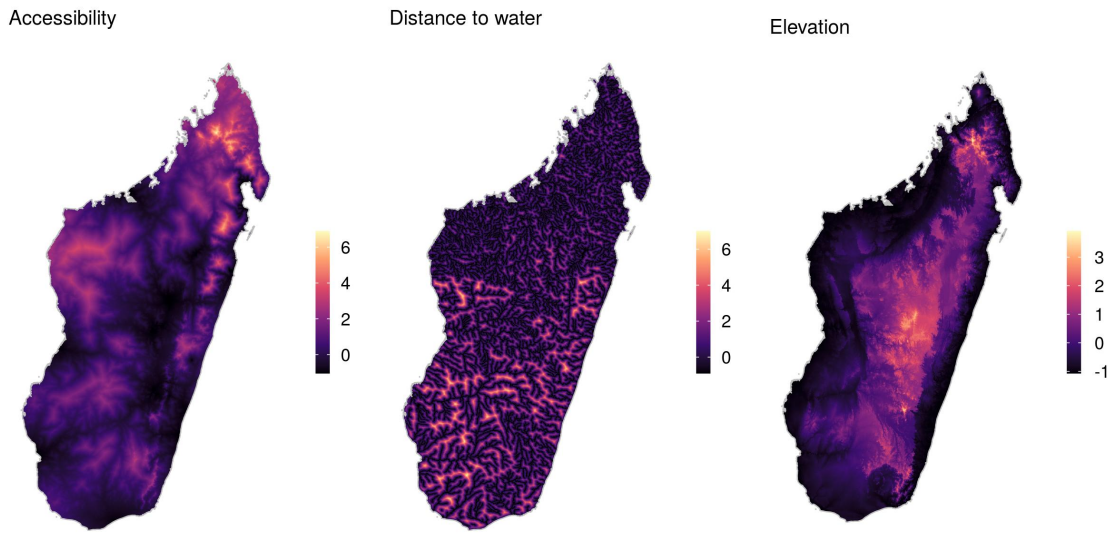


Figure 2.2: Standardised covariates used to inform prevalence estimates in Madagascar.

of individuals testing positive in survey i was modelled as a binomial random variable,

$$Y_i \sim \text{Binomial}(N_i, p_i) \tag{2.1}$$

where p_i was the true prevalence at this location. The true prevalence was modelled as

$$l(p_i) = \beta_0 + \beta^T X_i + f(s_i) \tag{2.2}$$

where l is the logit function,

$$l(p) = \log\left(\frac{p}{1-p}\right)$$

and β_0 and β were regression parameters. Here f was a spatial Gaussian process with a Mátern covariance function, parameterised by its range, ρ, marginal standard deviation, σ, and smoothness, ν. The range parameter, ρ, determines the extent of the spatial correlation, with a correlation of approximately 0.1 for points that are a distance ρ apart. The smoothness parameter ν, which controls the mean-square

differentiability of the process, is generally fixed due to identifiability issues, while the range and standard deviation were given hyperpriors and learned within the model. Gaussian priors were used for the regression coefficients,

$$\beta_0 \sim \text{Normal}(0, 2)$$
$$\beta \sim \text{Normal}(0, 1)$$

which regularise parameter estimates towards smaller coefficients and a simpler model. Penalized complexity priors [64] were used for σ and ρ. These are priors that shrink σ towards zero and ρ towards infinity. In other words, they shrink the model towards a lower complexity model that does not contain a spatial component. The joint penalized complexity prior on (σ, ρ) is such that

$$\text{P}(\sigma > \sigma_{\text{upper}}) = \sigma_{\text{prob}}, \text{ and}$$
$$\text{P}(\rho < \rho_{\text{lower}}) = \rho_{\text{prob}}$$

where σ_{upper}, σ_{prob}, ρ_{lower} and ρ_{prob} are context-specific constants to be chosen. The values chosen in this setting were $\sigma_{\text{upper}} = 0.5, \sigma_{\text{prob}} = 0.01, \rho_{\text{lower}} = 1$ and $\rho_{\text{prob}} = 0.01$. Locations were specified using latitude and longitude and therefore a range of 1 implies that the covariance between locations around 110km apart is approximately 0.1.

2.2.3 Sampling from the posterior

With the likelihood and prior distributions specified, the posterior density (up to a constant factor) can be calculated using Bayes' Theorem,

$$\text{P}(\beta_0, \beta, f, \rho, \sigma|D) \propto \text{P}(\beta_0, \beta, f, \rho, \sigma)\,\text{P}(D|\beta_0, \beta, f, \rho, \sigma)$$

where D represents the data $\{Y_i, N_i, X_i, s_i\}_{i=1}^{358}$. The prior is specified with Markovian structure,

$$\mathrm{P}(\beta_0, \beta, f, \rho, \sigma) = \mathrm{P}(\beta_0)\mathrm{P}(\beta)\mathrm{P}(f|\rho, \sigma)\mathrm{P}(\rho, \sigma)$$

and the likelihood can be obtained by first calculating p_i, for $i = 1, ..., 358$, using (2.2) and then evaluating

$$\mathrm{P}(D|\beta_0, \beta, f, \rho, \sigma) = \prod_{i=1}^{358} \mathrm{P}(Y_i|N_i, p_i)$$

where the likelihood $\mathrm{P}(Y_i|N_i, p_i)$ is here the density of the binomial sampling distribution for N_i trials with outcome probability p_i evaluated at the observed Y_i.

The resulting posterior density does not have a simple analytic form that would allow us to draw samples directly from the posterior, nor does it have closed form expressions for conditional densities that would facilitate, for example, Gibbs sampling. The typical strategy (when not working with spatial data) for sampling from this kind of posterior would be some kind of Hamiltonian Monte Carlo (HMC) approach. This could be used to explore the joint hyperparameter and random field posterior but for large spatial datasets, where each location is effectively a new model dimension, this can pose challenges for mixing and sampling efficiency. For this reason, the random field values are often marginalised out, greatly reducing the dimensionality of the parameter space. When using a normal likelihood, the random field values can be integrated out analytically, however this is in general not possible with other likelihoods (including binomial and Poisson). In these cases, we can use the Laplace approximation to approximate the marginalised posterior $\mathrm{P}(\beta_0, \beta, \rho, \sigma|D)$. To derive this approximation, suppose that we wish to compute the integral of some smooth density g,

$$I = \int_{x \in \mathbb{R}^n} g(x)\,\mathrm{d}x.$$

Assume that x_0 is the value at which g achieves its maximum and let $f(x) = \log g(x)$.

Then using the second-order Taylor series approximation of f at x_0 we have

$$\int_{x\in\mathbb{R}^n} g(x)\,\mathrm{d}x = \int_{x\in\mathbb{R}^n} \exp(f(x))\,\mathrm{d}x \tag{2.3}$$

$$\approx \int_{x\in\mathbb{R}^n} \exp\left(f(x_0) + \frac{1}{2}(x-x_0)^T Hf(x_0)(x-x_0)\right)\,\mathrm{d}x. \tag{2.4}$$

where $Hf(x_0)$ is the Hessian of f at x_0. Note that the second term of the expansion is zero because f achieves its maximum at x_0 and therefore the gradient of f in all directions is zero at this point. Furthermore, note that the right hand side of (2.4) is equal to

$$\exp(f(x_0)) \int_{x\in\mathbb{R}^n} \exp\left(\frac{1}{2}(x-x_0)^T Hf(x_0)(x-x_0)\right)\,\mathrm{d}x$$

and this integrand is proportional to a multivariate normal density. Therefore, we have that

$$I \approx -\frac{(2\pi)^{\frac{n}{2}} \exp(f(x_0))}{\det(Hf(x_0))^{\frac{1}{2}}}.$$

In other words, the Laplace approximation allows us to approximately calculate the integral I if we can find the maximum of g and calculate the Hessian of $\log g$ at this point. This can be used to approximate the marginal density,

$$\mathrm{P}(\beta_0, \beta, \rho, \sigma|D) = \int_{\mathbb{R}^n} \mathrm{P}(\beta_0, \beta, f, \rho, \sigma|D)\,\mathrm{d}f$$

for any values of $\beta_0, \beta, \rho, \sigma$. This approximate posterior density can be sampled from using HMC, a normal approximation (again using the theory underlying the Laplace approximation) or the Integrated Nested Laplace Approximation (INLA) approach [65, 66]. In the work in this thesis, I used normal approximations to obtain posterior samples. For spatial models this is often significantly faster than HMC (see discussion below) and is more widely applicable that INLA approaches (for example, it is more straightforward to use a normal approximation for disaggregation regression models, see Section 2.4).

Regardless of the sampling method, calculating the Gaussian process likelihood remains a challenge. For n observations, this requires solving a linear system involving an $n \times n$ covariance matrix, which has a cost of $O(n^3)$ and therefore scales very poorly as the number of observations increases. One solution is to approximate the Mátern Gaussian process with a Gaussian Markov random field (GMRF) as proposed by Lindgren et al. [67]. A GMRF is a Gaussian process on a discretely indexed set of points with the property that the conditional distribution of the value at any location given observations at all other locations only depends on a finite set of neighbours. This is similar to autoregressive models commonly used when modelling time-series—for example in an AR(1) process the distribution of an observation at some time t conditional on all other observations is in fact only dependent on the values observed at times $t-1$ and $t+1$. This is in contrast to a Gaussian process defined by a Mátern or exponential kernel, for example, where the conditional distribution at any location depends on the values observed at all other locations. The Markov property of the GMRF conditional distributions leads to a sparse precision matrix, which in turn allows for more efficient likelihood evaluations. Lindgren et al. [67] showed a link between Mátern Gaussian processes and the solutions to a class of stochastic partial differential equations (SPDEs). A finite element approximation to the solution of these SPDEs yields a GMRF approximation to the full Gaussian process. This provides an efficient modelling strategy that scales well with increasing numbers of observations. One limitation of this method is that the connection between Gaussian processes and GMRFs described is specific to the Mátern kernel. Another approach is to use a low rank approximation to the kernel (usually via Random Fourier Features), which is discussed in Chapter 3.

In [68], we made a direct comparison between using a normal approximation and the No-U-Turn Sampler [69] (an extension to HMC) to sample from the posterior when performing disaggregation regression. Using the normal approximation took

56 seconds compared to the No-U-Turn Sampler which took 94 hours (run until diagnostics suggested it had converged to the target distribution). Comparing the mean and standard deviation of each parameter (Table 2.1, reproduced from Table 1 in [68]) suggests that the resulting posteriors from the two methods were similar.

Table 2.1: Mean and marginal standard deviations for the posterior distributions of model parameters when calculated using the No-U-Turn Sampler and a normal approximation (reproduced from Table 1, [68]).

	No-U-Turn Sampler		Normal approx.	
Parameter	Mean	SD	Mean	SD
Intercept	-3.09	0.33	-3.12	0.27
Slope 1	0.20	0.28	0.24	0.27
Slope 2	0.32	0.21	0.30	0.21
Slope 3	-0.24	0.23	-0.24	0.21
$\log(\tau_u)$	1.08	0.27	1.10	0.27
$\log(\sigma)$	-3.34	0.59	-3.55	0.59
$\log(\rho)$	0.72	0.31	0.56	0.30

2.2.4 Results

The modelled estimates of prevalence are shown in Figure 2.3. We can see that these estimates resemble a smoothing and interpolation of the observed data points (Figure 2.1). The trends of higher prevalence on the southwest and southeast coasts and lower rates in the highlands and desert south are generally in agreement with historical and more recent spatial patterns (see Chapter 4 for more discussion). Figure 2.4 shows the separate, additive contributions of the Gaussian process and covariates to the logit prevalence. Much of the fine-scale detail of the risk surface comes from the covariate information, while the Gaussian process captures a mix of longer scale spatial trends and local "hotspots". The latter are particularly seen in areas in the southeast and southwest suggesting endemicity patterns that could not be explained by the available covariates. The mean and standard deviation of the posterior distribution of each model parameter is shown in Table 2.2. The positive coefficient

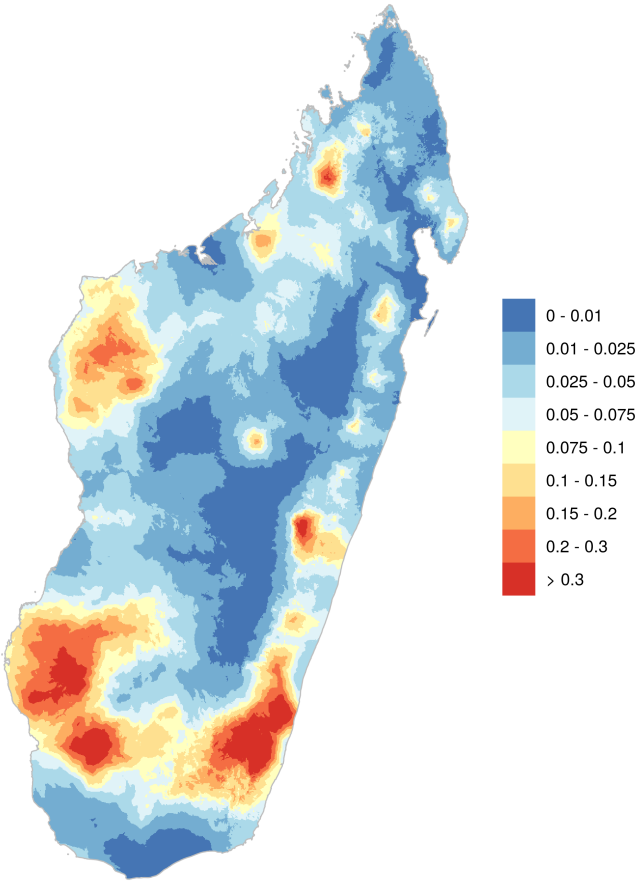


Figure 2.3: Estimated *P. falciparum* prevalence for individuals between 6 and 59 months of age in Madagascar in 2016.

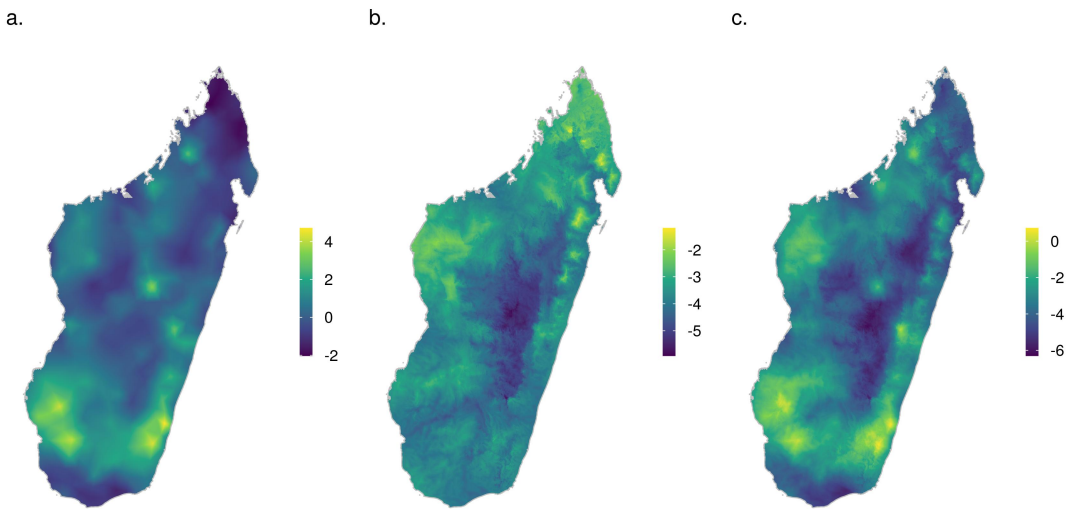


Figure 2.4: The predicted values of the (a) Gaussian process, (b) linear contribution of covariates and (c) logit prevalence in the Madagascar prevalence model.

for accessibility to cities, and its relatively small standard deviation, supports the hypothesis that transmission is higher in remote areas. Similarly, the negative coefficient for elevation suggests lower transmission rates in highland areas, which is also as expected. There is no clear effect of distance to water in the model, with large proportions of the posterior distribution above and below zero. Figure 2.5 shows the 2.5% and 97.5% quantiles of the posterior distribution of the prevalence. We can see that while the spatial patterns are consistent, there is a fairly large amount of uncertainty in the prevalence values within the 95% credible interval. This is in large part a reflection of the small survey sizes per cluster in this dataset and the large variance of the binomial sampling distribution at low prevalence. In Chapter 4, I also quantify uncertainty using 'exceedance surfaces' which may communicate uncertainty in a more intuitive and useful way.

It is useful to consider the assumptions that have gone into this model. Clearly one is that the effect of each covariate on logit prevalence is assumed to be linear, with no interactions between covariates. In relatively low data settings, this is a pragmatic

Table 2.2: Mean and standard deviation for the model parameters in the model of prevalence in Madagascar.

Parameter	Mean	SD
β_0	-4.67	0.37
$\beta_{\text{Accessibility}}$	0.58	0.15
$\beta_{\text{Distance to water}}$	-0.05	0.07
$\beta_{\text{Elevation}}$	-0.55	0.24
$\log \rho$	0.09	0.25
$\log \sigma$	0.35	0.11

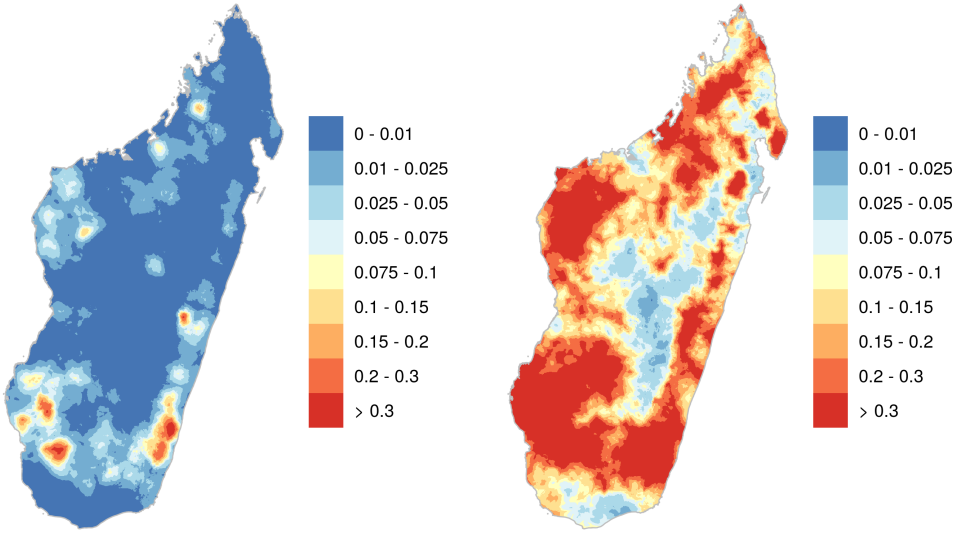


Figure 2.5: Lower (2.5%, left) and upper (97.5%, right) quantiles of the posterior distribution of the estimates of prevalence in Madagascar in 2016.

choice as the inclusion of more flexible relationships and interaction terms can lead to overfitting. Another model choice is the use of a stationary covariance function, which supposes the covariance between any two locations is simply a function of the distance between them, regardless of their absolute positions. Thinking of the Gaussian field as representing unobserved, spatially-structured risk factors, this assumption may not always hold. As we can see in Figure 2.2, this stationarity is not exhibited in the observed covariates, as in each case there are some areas where covariate values vary greatly over short distances and other large areas where values are very similar.

The final assumption I will highlight here is the identification of each survey point with a single pixel in the 1km-by-1km grid. One issue with this approach is that in many health surveys, including the MIS that these data come from, random noise is added to the survey locations (up to a distance of 2km in urban locations and 5km in rural locations) to protect the privacy of respondents. Furthermore, associating the risk to an individual with risk in the pixel where they live is a simplification, neglecting the movement of populations during the day. We would expect risk in nearby areas to be similar and so not incorporating small-scale movement into the model may be unlikely to have a large effect on results. However, systematic travel in a population to areas with very different risk profiles can have a greater effect (for example, if a large proportion of the population in an otherwise low risk area work in a high risk area). For this and a number of other reasons, there has been much interest in using mobile phone data and other information (such as social media, travel history surveys and genomic data) to better understand human mobility patterns and their impact on malaria risk [70, 71, 72, 73].

2.3 Mapping incidence in Ayeyarwady, Myanmar

2.3.1 Background

Malaria risk in Myanmar is highly heterogeneous, with no or low rates of transmission in large parts of the country (covering 44% and 40% of the population, respectively, in 2017) and some smaller high transmission areas (defined as areas with estimates of more than one case per 1000 people per year) [75]. *P. falciparum* and *P. vivax* are both endemic, with *P. falciparum* being slightly more common, accounting for around 60% of cases in 2017 [75]. Figure 2.6 shows the Malaria Atlas Project (MAP) estimates of incidence in 2017. Transmission has decreased significantly in the past decade, with a reduction from around 420,000 cases (confirmed by rapid diagnostic

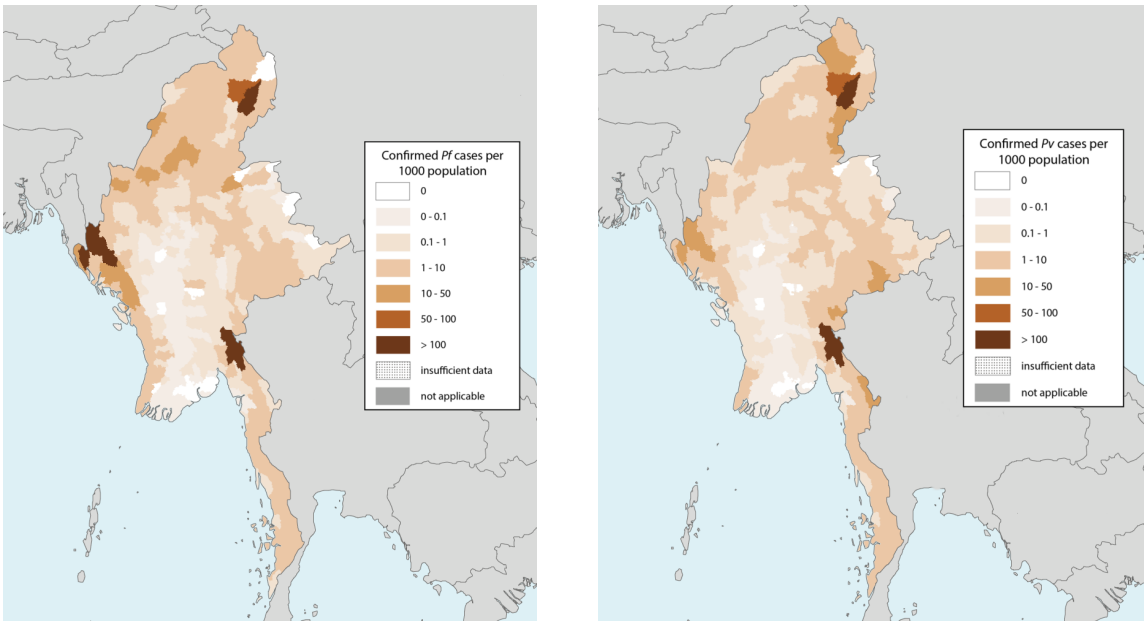


Figure 2.6: MAP estimates of incidence of *P. falciparum* (left) and *P. vivax* (right) malaria in Myanmar in 2017 [29, 27, 74]. Figure made by Jennifer Rozier.

test (RDT) or microscopy) in 2010 to 73,000 in 2018 [75]. Ayeyarwady region, in the south west, is a high transmission region for both *P. falciparum* and *P. vivax* with an estimated 1-10 cases per 1000 people caused by each species each year. This mapping project was conducted for the Clinton Health Access Initiative with the aim of producing refined estimates of malaria incidence in Ayeyarwady in 2016 and 2017 using health facility incidence data.

2.3.2 Data

Annual case data from 2016 and 2017 was available from 2212 health facilities, of which 1368 facilities had both precise location information and estimated catchment populations. These 1368 health facilities were the main source of response data in the model. However, for all facilities (including those not precisely located) the township (the administrative division below region and district) in which they reside was known and therefore the observed case totals in each township could be calculated, based on data from all health facilities. This information was also used in the model, allowing

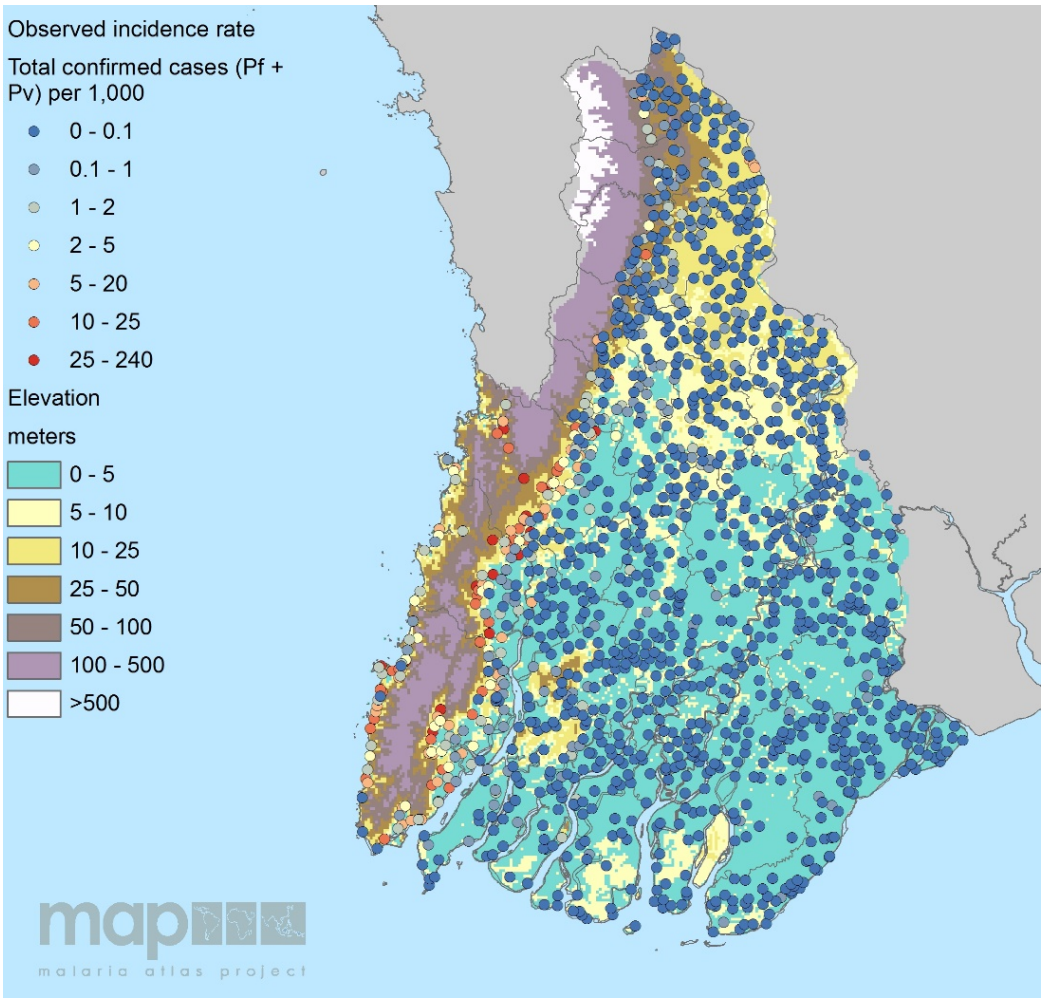


Figure 2.7: Average (over 2016 and 2017) number of confirmed cases per 1,000 people each year at health facilities in Ayeyarwady. Figure made by Katherine Battle.

case data from all health facilities to inform the incidence estimates.

Figure 2.7 shows the average annual rate of confirmed cases over the two years at each health facility. There was a large amount of spatial heterogeneity in the observed incidence rate, with almost all cases occurring near the mountains in the west and northwest. The mountains themselves have a very low population density, assumed to be too low for stable transmission, so this area was not modelled. The majority of health facilities (1849 out of 2212) saw no confirmed cases in the two years.

2.3.3 Model specification

One approach to model this incidence data would be to suppose the number of cases observed at health facility i (in location s_i) at time t, Y_{it}, follows a Poisson sampling distribution,

$$Y_{it} \sim \text{Poisson}(\lambda_{it}\text{pop}_i) \tag{2.5}$$

where pop_i is the catchment population of facility i and λ_{it} is the incidence rate to be represented by a latent risk surface. As with the model of prevalence in Madagascar, a linear effect of covariates could be assumed, this time with a log link function, such that

$$\log \lambda_{it} = \beta_0 + \beta^T X_{it} + f(s_i)$$

where β_0, β are parameters, X_{it} are covariate values at location s_i at time t, and f is a zero-mean spatial Gaussian process. However, the large number of health facilities that did not observe any cases in either year suggests that there may be areas in the region where no transmission occurs, rather than occurring at very low rates as the above data generating process would imply. Therefore, I instead used a hurdle model, which distinctly represents both the probability that the case generating process (i.e. transmission) occurs and the rate of this process where it does occur. A negative binomial was used for the sampling distribution under transmission to allow for overdispersion (beyond that of the Poisson distribution, which has a variance exactly equal to its mean). For facility i at time t, the number of cases observed, Y_{it}, was modelled as a random variable with

$$P(Y_{it} = y_{it}) = \begin{cases} p_{it}\, \pi(y_{it}|\lambda_{it}\text{pop}_i, \delta) & y_{it} > 0 \\ (1 - p_{it}) + p_{it}\, \pi(0|\lambda_{it}\text{pop}_i, \delta) & y_{it} = 0 \end{cases}$$

where p_{it} was the probability of transmission occurring at that location and time and π was the probability mass function of the negative binomial distribution with

rate $\lambda_{it}\text{pop}_i$ and dispersion δ. In this data generating process, zero cases can be observed if transmission does not occur or if it occurs but produces zero cases. On the other hand, a positive number of cases can only be observed if the transmission process occurs. The probability of transmission, p_{it}, was modelled as

$$\text{logit}(p_{it}) = \beta_0 + \beta^T X_{it} + f(s_i)$$

where X_{it} were covariate values and f was a spatial Gaussian process. The incidence rate under transmission λ_{it} was modelled as

$$\log(\lambda_{it}) = \bar{\beta}_0 + \bar{\beta}^T \bar{X}_{it} + \bar{f}(s_i)$$

where again $\bar{X}_{it}$ are covariate values (potentially different variables to those used to model occurrence) and $\bar{f}$ was a spatial Gaussian process. The dispersion parameter for the negative binomial process (assumed constant across time and space) was given a normal prior with mean 0.5 and standard deviation 1, truncated at 0 (i.e. the density of this non-negative parameter was proportional to the normal density with this mean and standard deviation for non-negative values and zero otherwise). All coefficients were also assigned normal priors with mean 0 and standard deviation 1. Penalised complexity priors were used for the range and scale of each Gaussian process as in the previous example.

The final component of the model was a second likelihood, representing the likelihood of the observed township case totals. Recall that these are the health facility case counts aggregated over administrative regions. These totals included facilities that could not be precisely located and therefore did not contribute to the previous likelihood from the hurdle model described above. These totals were less informative than the data from precisely located facilities, since these cases were not associated with an exact location. Nevertheless, these data can still give an indication of overall

case burden in each township, which may be particularly useful in townships where many health facilities are missing location or population information. This likelihood was calculated as follows. For any given regression parameters and random field values, the rate in pixel j of township i at time t, λ_{ijt}, was calculated as follows

$$\log \lambda_{ijt} = \beta_0 + \beta^T X_{ijt} + f(s_{ij})$$

where X_{ijt} were the covariate values in this location and s_{ij} was the location of this pixel. Given these parameters, the expected number of cases occurring in this pixel in this year (that would be seen in the healthcare system) is $\text{pop}_{ij}\lambda_{ijt}$, where pop_{ij} was the population in this pixel. The total number of cases expected to be observed in the township at time t, cases_{it}, was then assumed to be the sum of the number of cases in each pixel in that township,

$$\text{cases}_{it} = \sum_{j=1}^{N_i} \text{pop}_{ij}\lambda_{ijt}$$

where N_i is the number of pixels in township i. For simplicity, the observed number of cases in this township, Y_{it}^{township}, was assumed to have a normal sampling distribution with a standard deviation of 20,

$$Y_{it}^{\text{township}} \sim \text{N}(\text{cases}_{it}, 20).$$

A more rigorous treatment of these townships totals could have been to use a Poisson sampling distribution, as discussed in the next section. It may also have been useful, if using a normal distribution, to scale the standard deviation by some function of the population in the township (see Nandi et al. [68] for an example).

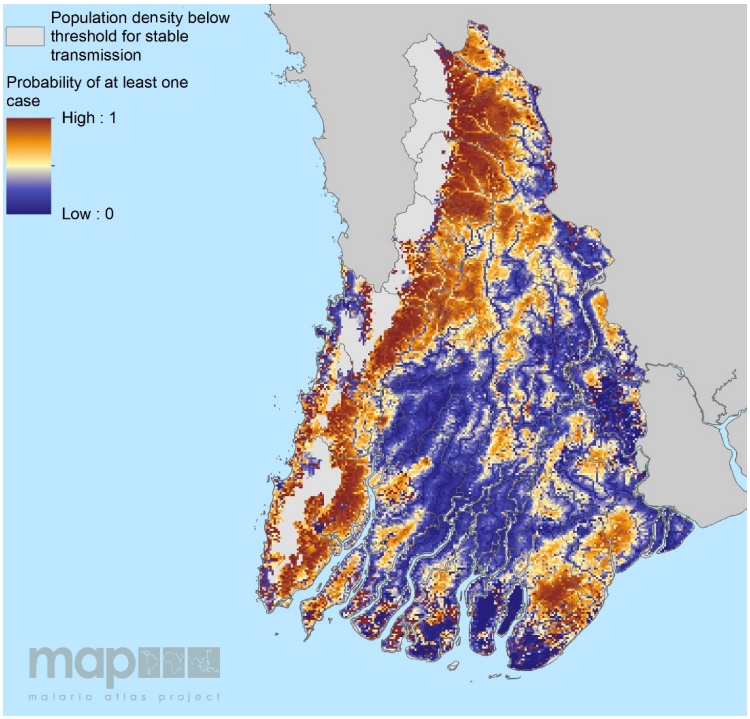
An internal covariate selection procedure was implemented during model building by fitting the model first under a very narrow prior on the coefficients for both the

occurrence and prevalence surfaces, with covariates selected for each if the posterior mean of these coefficients exceeded a threshold value. As a normal prior is equivalent to ridge regression, it may have been more effective to use a Laplace prior in this procedure. This is the Bayesian equivalent of a LASSO or L_1 penalty, which offers a pseudo-sparsity effect for Bayesian variable selection as small coefficients are shrunk more strongly towards zero [76].

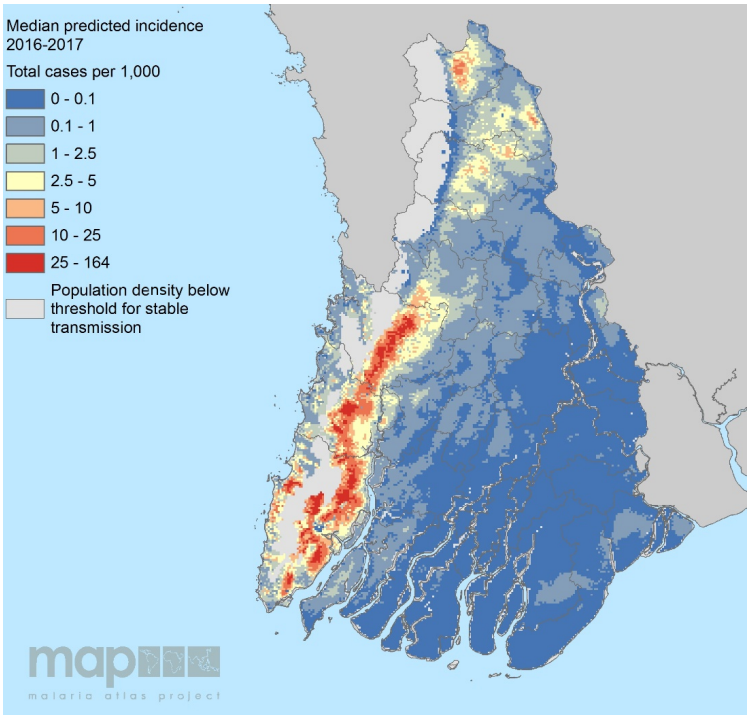
2.3.4 Results

The estimated probability of transmission occurring (averaged over both years) is shown in Figure 2.8(a). This probability is very high in the north and north-west, which is expected as cases were observed at many health facilities in this area. In contrast, much of the rest of the region has a very low probability of transmission, reflecting the large number of facilities observing no cases in this area, with pockets of higher probabilities (particularly in the south-east). While this was not the main output of interest for this analysis, this map is potentially useful for distinguishing areas of no transmission from those with low transmission, which could inform targeting of elimination efforts.

The predicted incidence rate, which is equal to the probability of occurrence multiplied by the negative binomial rate, is shown in Figure 2.8(b). Again this follows the pattern of the observed data, with the highest rates near the mountainous areas. Using both maps, we can identify areas (particularly in the southeast) where the predicted incidence rate is consistently low but the probability of occurrence is more variable. Here the probability of occurrence of transmission surface may in some sense capture the environmental suitability for transmission, even in areas where little or no transmission was observed. Tables 2.3 and 2.4 show the posterior mean and standard deviation of the coefficients of the covariates for the occurrence and negative binomial rate surfaces.



(a)



(b)

Figure 2.8: (a) Predicted probability of transmission and (b) predicted incidence rate (averaged over 2016 and 2017) in Ayeyarwady. Figure made by Katherine Battle.

Again, it is important to consider the assumptions underlying this model. Many of these assumptions are similar to those in the prevalence model in the previous example or come from the two-step data generating process. As with the prevalence model, all observations are treated as points, imposing an implicit assumption that every case observed at a health facility occurred within the same pixel that contains the facility. This was a pragmatic decision made to facilitate modelling and computation that, although almost certainly false at a granular level, was considered acceptable in this setting where there is a very high spatial coverage of the population by local health facilities. In Chapter 5, health facility case counts are instead modelled as an aggregation over a catchment area. The health facility catchment populations and use of township totals also assume a high level of treatment seeking. If a large proportion of the population is unlikely to seek care at health facilities, this model will systematically underestimate incidence in the community. Estimates of treatment seeking for fever in Myanmar are generally high [77], although this varies by region [78, 79] and I am unaware of any literature of treatment seeking behaviour in Ayeyarwady specifically. The analyses in Chapters 4 and 5 show one method for incorporating treatment seeking behaviour into this modelling framework.

2.4 Mapping incidence from aggregated data in four states in India

In this work I estimated annual incidence in four states in India—Maharashtra, Madhya Pradesh, Gujarat and Karnataka—as part of another collaboration with the Clinton Health Access Initiative. Unlike in the previous example, this incidence data was not connected to point locations but instead aggregated over 'blocks' (administrative regions two levels below states). The aim of this work was to predict incidence rates at a fine scale (at a similar resolution to the Madagascar and Ayeyarwady maps).

Table 2.3: Mean and standard deviation for the coefficients for the occurrence of transmission surface in the model of incidence in Ayeyarwady.

Covariate	Mean	SD
Aridity index [80]	0.91	0.41
Tree cover fraction [81]	0.71	0.72
Distance to water [58]	-0.01	0.26
Landcover – cropland [82]	-0.63	0.74
Landcover – forest [82]	0.95	0.96
Landcover – grassland [82]	1.60	0.55
Landcover – shrubland [82]	0.90	0.62
Landcover – water [82]	-0.81	0.84
Landcover – wetland [82]	1.38	0.92
Slope [56]	0.49	0.92
Night lights [83]	-0.51	0.38
Topographic wetness index [56]	-0.85	0.60
Land surface temperature - day [84]	1.34	0.45
Land surface temperature - night [84]	-0.30	0.64
Tasselled cap brightness [85]	-0.79	0.52

Table 2.4: Mean and standard deviation for the coefficients for the negative binomial rate surface in the model of incidence in Ayeyarwady.

Covariate	Mean	SD
Tree cover fraction [81]	0.33	0.23
Distance to water [58]	-0.37	0.10
Elevation [56]	0.33	0.67
Landcover – grassland [82]	0.00	0.08
Landcover – shrubland [82]	-0.29	0.26
Landcover – urban/built-up [82]	-0.18	0.25
Landcover – wetland [82]	-0.03	0.17
Night lights [83]	-0.28	0.23
Topographic wetness index [56]	0.45	0.26
Enhanced vegetation index [86]	0.19	0.27
Land surface temperature - day [84]	1.49	0.27
Land surface temperature - night [84]	0.06	0.32
Tasselled cap brightness [85]	-0.26	0.28

There are two major challenges when making predictions at a higher resolution than that of the observed data. The first is the 'ecological fallacy' or 'ecological bias', where relationships learned between variables on one spatial scale may not hold on another [87, 88, 89]. This can be caused by aggregating non-linear individual-level relationships, confounding variables, and 'contextual effects' (where the effect of an explanatory variable on the response at a fine-scale depends on both the value of the variable at that location and in nearby locations [89]). The second issue is the 'modifiable areal unit problem', where fine-scale predictions based on aggregated data may change based on how the data is aggregated geographically [90]. In fact, issues can arise even when predicting at the aggregate level, as common autoregressive models for aggregated data can produce counterintuitive and implausible correlation structures when the aggregation areas are irregular [91].

For these reasons, I used disaggregation regression [92, 93, 94] to model this data, which has become a common method for modelling aggregate data in spatial ecology and epidemiology [92, 93, 27]. Disaggregation regression attempts to reduce exposure to the ecological fallacy and modifiable areal unit problem by modelling the response variable on the same fine-scale at which predictions will be made. In a large-scale simulation study carried out by myself and colleagues, we found that disaggregation regression made useful fine-scale predictions of incidence even under some model misspecification, provided that the areas that the response data was aggregated over were not too large [95]. Most relevant for this project is that administrative level 3 units in India, which correspond to the blocks in these four states, were small enough to make accurate fine-scale predictions.

2.4.1 Background

Malaria is a major public health challenge in India, with an estimated 9.6 million cases occurring in 2019. A large majority of the population is at risk, with 162.5 million

people living in high transmission areas and a further 1.1 billion in low transmission areas (making up 12% and 81% of the population respectively). Both *P. falciparum* and *P. vivax* are endemic in the country, with the former accounting for approximately 57% of cases observed between 2010 and 2018 [96]. Recent control efforts have had a substantial impact in reducing malaria burden, with a 48% reduction in reported cases between 2000 and 2017 [75]. The National Strategic Plan 2017-2022 set out an aim to achieve elimination in all Category 1 and 2 districts (defined as districts with an annual incidence of less than 1 case per 1,000 people for Category 1 and between 1 and 2 cases per 1,000 people for Category 2) by 2022 and an overall goal of elimination throughout the country by 2027 [97, 98]. The highest burden areas are in the north and northeast, as shown in the estimates in Figure 2.9. The four states in this analysis were chosen to represent a range of epidemiological, demographic and socioeconomic settings. In 2015, Madhya Pradesh was classed as Category 3 state, while Gujarat, Maharashtra and Karnataka were Category 2.

2.4.2 Data

The response data was the number of observed cases due to *P. falciparum* and *P. vivax* in each block in 2015. The observed case rates (based on block populations) are shown in Figure 2.10. Incidence was generally higher for *P. vivax* than *P. falciparum* and for both species incidence rates were higher in the north and northeast. As with the Ayeyarwady analysis, the covariates used to inform this model were a mixture of environmental and socioeconomic variables, which are listed in Table 2.5.

2.4.3 Model specification

The incidence rate for each species was modelled separately. Let $Y_1, ..., Y_N$ be the number of cases observed in each block. The incidence rate in pixel j of the block i,

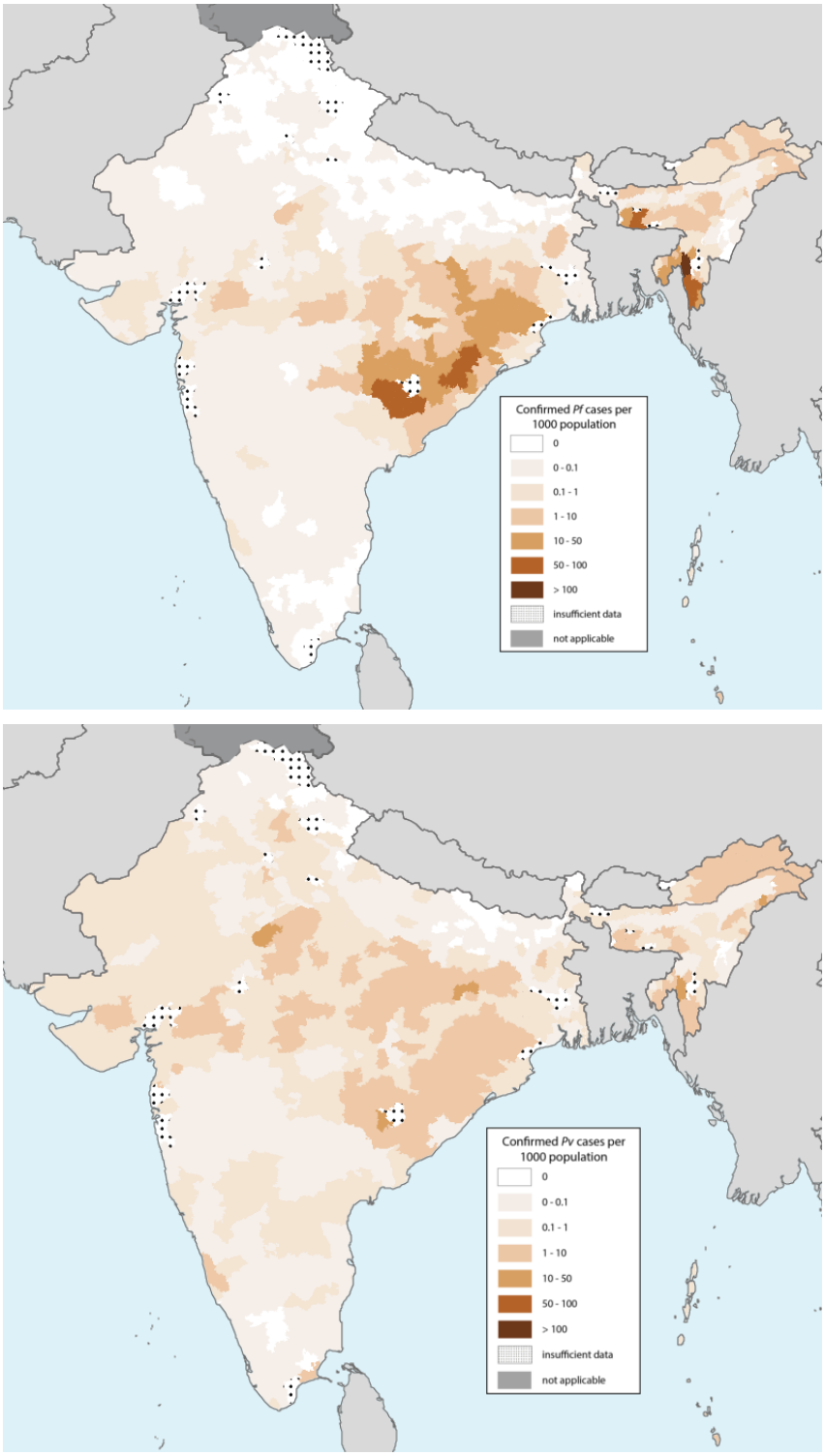


Figure 2.9: MAP estimates of incidence of *P. falciparum* (top) and *P. vivax* (bottom) malaria in India in 2015 [29, 27, 74]. Figure made by Jennifer Rozier.

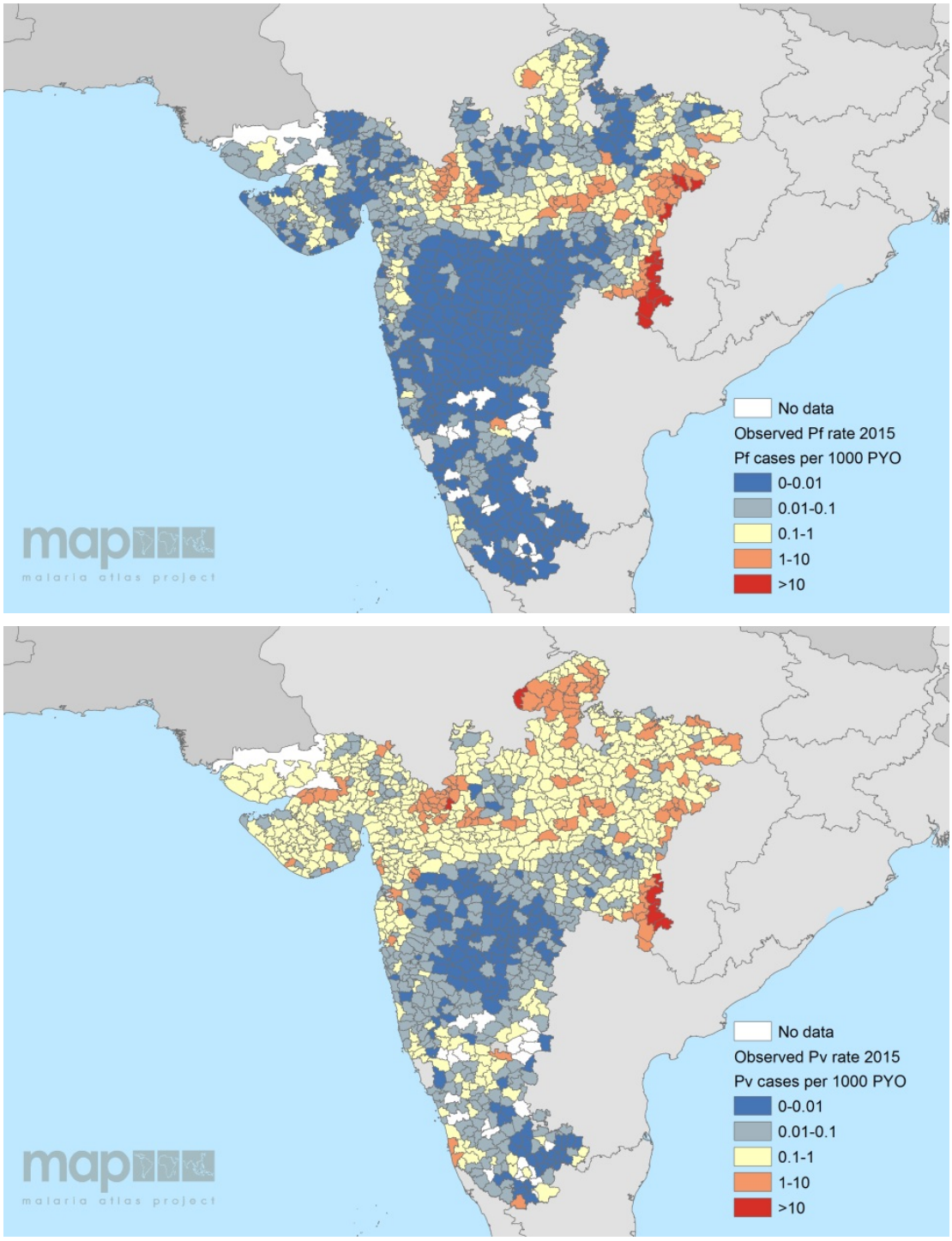


Figure 2.10: Observed incidence rates for *P. falciparum* (top) and *P. vivax* (bottom) in the four states in the study area in 2015. Figure made by Katherine Battle.

λ_{ij}, was modelled as

$$\log \lambda_{ij} = \beta_0 + \beta^T X_{ij} + f(s_{ij})$$

where s_{ij} was the location of this pixel, X_{ij} were covariate values and f was a spatial Gaussian process with Mátern covariance structure. As in previous models, normal priors were used for the coefficients and penalised complexity priors were applied to the Gaussian process hyperparameters. The number of cases occurring in each pixel over the year, Y_{ij}, was then modelled with a Poisson distribution,

$$Y_{ij} \sim \text{Pois}(\lambda_{ij} \times \text{pop}_{ij})$$

where pop_{ij} was the pixel population. The number of cases in block i was given by the sum of cases in each pixel in that block. Under the assumption that the Poisson distributions in each pixel were independent conditional on the underlying rate λ_{ij}, this sum is another Poisson distribution,

$$Y_{i1} + Y_{i2} + ... + Y_{iN_i} \sim \text{Pois}\left(\sum_{j=1}^{N_i} \lambda_{ij} \text{pop}_{ij}\right)$$

which completes the model by providing the sampling distribution of the observed aggregated case data.

2.4.4 Results

Figure 2.11 shows the predicted fine-scale incidence rates for *P. falciparum* and *P. vivax* and the coefficients of the covariates in the model are listed in Tables 2.5 and 2.6. As expected, the general spatial pattern of the predictions agrees with the observed data. The model produced high resolution estimates from the aggregated data and we were also able to make predictions for the areas with no data reported. Predictions in blocks with no data were based on the covariate values and the spatial

trends in nearby blocks. High estimates of *P. falciparum* incidence were noted in districts where more than 25% of the population identifies as tribal (according to the 2011 census [99]), which may indicate increased risk in these sub-populations. In the *P. vivax* estimates, there were slightly higher levels of incidence around the urban centres of Mumbai and Mangalore compared to the surrounding areas, suggesting significant levels of urban malaria.

Table 2.5: Mean and standard deviation of the coefficients in the *P. vivax* model for the four states in India.

Covariate	Mean	SD
Accessibility	0.41	0.06
Distance to water	-0.26	0.06
Elevation	-1.00	0.10
EVI (SD)	0.24	0.07
TCB (SD)	0.39	0.12
Night lights	0.15	0.03
EVI (mean)	-0.19	0.12
LST day (mean)	0.00	0.13
TCB (mean)	-1.94	0.27
TCW (mean)	-0.73	0.26

Table 2.6: Mean and standard deviation of the coefficients in the *P. falciparum* model for the four states in India.

Covariate	Mean	SD
Accessibility	0.55	0.05
Distance to water	0.05	0.05
Elevation	-0.78	0.09
EVI (SD)	0.21	0.09
TCB (SD)	0.14	0.17
Night lights	0.16	0.04
EVI (mean)	-0.10	0.16
LST day (mean)	-0.83	0.15
TCB (mean)	-1.83	0.30
TCW (mean)	-1.71	0.17

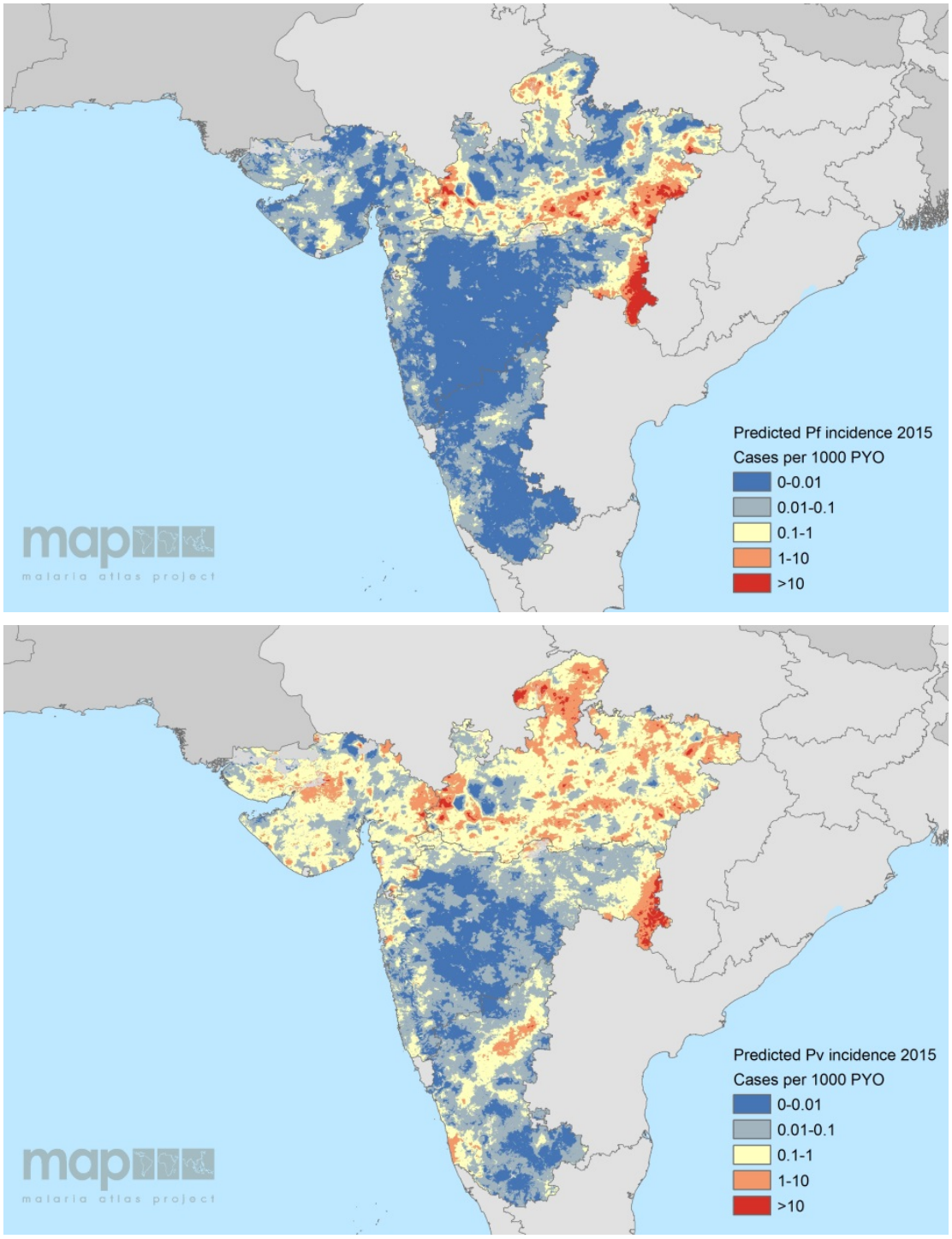


Figure 2.11: Predicted incidence rates for *P. falciparum* (top) and *P. vivax* (bottom) in the four states in the study area in 2015. Figure made by Katherine Battle.

2.5 Conclusion

In this chapter, I described three Bayesian models of malaria risk. Several concepts that were used in these models are used again throughout the rest of this thesis. Firstly, we saw a repeated strategy of modelling prevalence or incidence in each pixel as a function of covariate values and the value of a Gaussian process term in this pixel. The observed data was then connected to pixel-level risk either directly, by associating an observation with a specific pixel, or as an aggregation over many pixels. We also saw how the sampling distribution and model structure can change based on the nature of the observed data. These ideas are used again in the geospatial models in Chapters 4, 5 and 6. In Chapter 3, I also return to the idea of associating a single observation of a response variable (such as incidence in a region) with multiple covariate observations (such as rainfall in each pixel in that region), in the context of independence tests. Finally, we discussed the differences between incidence and prevalence data. In Chapter 4, I develop a model that makes use of both these sources of data, taking into account the strengths and limitations of each.

Chapter 3

Causal discovery with spatiotemporal data

The aim of this chapter is to propose a robust method for performing causal discovery—i.e. learning the causal structure of a set of variables based on joint observations—using spatiotemporal data. The majority of this chapter is devoted to introducing the necessary theory on causal discovery algorithms and kernel embeddings. At the end of the chapter I describe how these concepts can be used together to define a causal discovery procedure that can be applied to arbitrary spatiotemporal data, including cases where variables are observed at different spatial scales. In an epidemiological context, the results of this process may be of interest in their own right, as the causal relationships between environmental and socioeconomic factors, disease control methods, and disease risk are not necessarily known. Where the causal mechanisms *are* known or presumed (for example, rainfall may be confidently assumed to be a cause of malaria due to its impact on availability of vector breeding sites), causal discovery can help to identify the likely time lags between cause and effect. Causal discovery can also be used an as initial or intermediate step in other analyses; indeed its main application in this thesis is as a feature selection method (see Chapters 4 and 5) where the feature set is the set of direct causes of the response variable. In Chapter 6, I discuss the use of causal discovery for identifying confounding variables when assessing intervention effectiveness. Similar applications of this

method are likely to be found in other fields with a strong interest in spatial data, such as ecology or demography.

A key component of this causal discovery procedure is the repeated application of a constraint-based causal discovery algorithm (I use the PC algorithm [43, 40] but other constraint-based algorithms could be used). The algorithm is repeatedly applied to subsamples of the data in order to improve the stability of the final results and allow the confidence of each relationship to be quantified. The variables of interest are core epidemiological metrics and ancillary covariate data, such as environmental and socioeconomic data. There are three main challenges when applying causal discovery algorithms in a spatial epidemiological context that the proposed procedure overcomes:

1. These variables generally do not satisfy the requirements of many common conditional independence tests. For example, some tests assume variables are jointly Gaussian or require the conditioning variable to be discrete. For this reason, I explored a variety of non-parametric independence tests based on the theory of embedding distributions into reproducing kernel Hilbert spaces.

2. The spatial or spatiotemporal structure of the data can cause independent variables to appear dependent, inflating the number of causal relationships found. Moreover, this spatial structure can violate the assumption of independent and identically distributed observations that is used in many kernel-based tests. To address this, I developed methods to remove spatial and spatiotemporal trends from the data ('pre-whitening') before applying the causal discovery algorithm.

3. The variables may be observed on different spatial scales. Malaria incidence data, in particular, is generally an aggregation of cases over some area, while environmental and socioeconomic variables are generally available at a higher resolution. Accordingly, I extended the theory of kernel-based independence

tests to work with this mix of aggregated and fine scale data.

I begin this chapter by introducing causal discovery algorithms and describing the particular algorithm I implemented. I then discuss the kernel-based independence tests used and the modifications necessary to test variables that are observed on different spatial scales. I then describe the prewhitening methods used to remove spatiotemporal trends and finally summarise the full causal discovery procedure.

3.1 Graphical models of causality

Causal structures between sets of variables are commonly described using graphs, typically directed acyclic graphs (DAGs). These graphs are a convenient and intuitive way to encode causal information and allow causal discovery algorithms and methods for estimating causal effects to be expressed using graphical language [44, 40, 38, 31]. DAGs have also become a common tool in epidemiology and the medical sciences to communicate the known or presumed causal relationships underlying analyses [100]. By causal 'structure' or 'relationships' here I refer to information about which variable causes which but not the precise nature of these relationships (for example, functional forms of these relationships, if they exist). While the causal structure of both deterministic and stochastic systems can be expressed graphically, in this work I only consider stochastic systems of random variables.

In the introduction (Chapter 1) we discussed different notions for what it means for one variable to cause another. To express causal relationships graphically, we also need to distinguish direct and indirect causes. An indirect cause of a variable X can be thought of as a variable that only affects X through its effect on other variables in the system. For example, low temperatures might be a cause of road traffic accidents in some location but only because of the increased likelihood of there being ice on the road. In this case temperature would be an indirect cause of accidents, while ice

is a direct cause. The direct cause 'screens off' the variable of interest X from the indirect cause, in the sense that once the value of the direct cause is known, the value of the indirect cause no longer provides any information about the value of X. In our example, if we know that there is ice on the road, knowing the temperature will not change the number of accidents we expect. An important point to note is that the concepts of direct and indirect causes are defined relative to the set of variables observed or considered. If the presence of ice was not a variable in our system, for example, then temperature would be considered a direct cause of accidents.

This example highlights a common problem when considering causal structures in real-world settings. Even if temperature largely only affects the number of traffic accidents through its action on road conditions, we can imagine scenarios where temperature further affects the likelihood of accidents in other less significant ways. For example, temperature could have a small effect on each driver's concentration. Moreover, particularly extreme temperatures could have large effects, for example particularly low temperatures resulting in large snowfalls could in fact reduce accidents by reducing travel. Extending these ideas to their logical conclusion, almost any variable could be considered a direct cause of any other (with some exceptions, such as variables that can be strictly ordered in time). In practice, only causal effects of non-trivial size are considered and variables are limited to a pragmatic range of likely (rather than all possible) values. However, this is generally done implicitly and can be difficult to formalise [40].

To formally define a direct cause, we follow a similar approach to that of Spirtes et al. [40]. Let $\mathbf{V}$ be a set of variables with $X, Y \in \mathbf{V}$. The variable X is said to be a *direct cause* of Y with respect to $\mathbf{V}$ if there exists a set $\mathbf{C} \subset \mathbf{V}$ containing X such that

direct cause

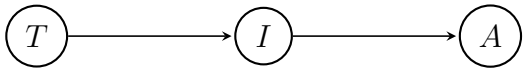
1. every element of $\mathbf{C}$ is a cause of Y,

2. for any fixed values of the variables in $\mathbf{C}$ the outcome of Y is unaffected by the

values of the variables in $\mathbf{V}\backslash\mathbf{C}$, and

3. no proper subset of $\mathbf{C}$ satisfies conditions 1 and 2.

This definition formalises the idea of the direct causes of Y screening off Y from all other variables. With the concept of direct and indirect causes defined, we can then describe the causal relationships between variables using a DAG, where each variable is represented by a node and an edge from variable X to variable Y exists if (and only if) X is a direct cause of Y. Our previous example would be expressed as



where T represents temperature, I represents ice, and A represents road traffic accidents. The fact that these graphs are acyclic follows from the axioms that causation is transitive and irreflexive (see Section 1.1), since if there did exist a cycle $X_1 \to X_2 \to ... \to X_m \to X_1$ in a causal structure then a repeated application of the transitive property would yield that X_1 is a cause of itself, contradicting the assumption of irreflexivity. In fact, by the same logic transitivity and irreflexivity also imply the third axiom, that causation is antisymmetric.

Now that we are able to describe causal relationships between sets of variables using DAGs, we can introduce some related terminology. The following definitions apply to both directed graphs and graphs containing both directed and undirected edges. Firstly, a variable X is a *parent* of variable Y in graph G if the edge $X \to Y$ is in G. The variable X is a *descendent* of Y in graph G if there exists any directed path $Y \to Z_1 \to ... \to Z_n \to X$. The set of parents of variable Y in graph G is denoted Parents(G, Y), often written as Parents(Y) when the graph is clear from context. Similarly, the set of descendants of Y can be written as Descendants(G, Y) or simply Descendants(Y). A variable X is *adjacent* to variable Y if there exists an edge (in any direction or undirected) between them. The set of variables that are adjacent to Y is written Adjacencies(G, Y) or Adjacencies(Y). For instance, in the example of

parent

descendent

adjacent

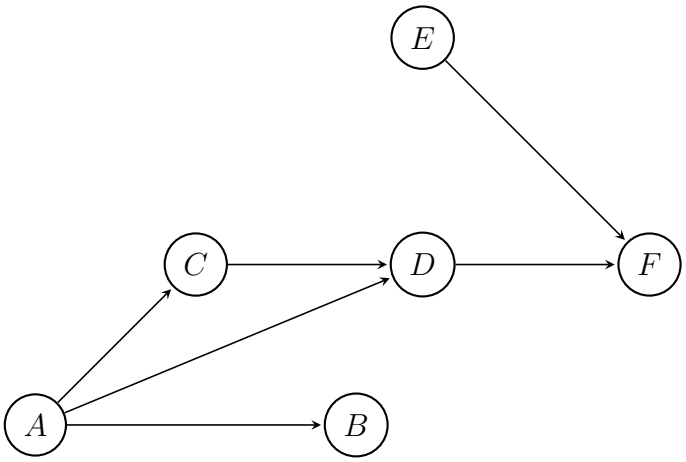


Figure 3.1: An example of a DAG.

the DAG shown in Figure 3.1,

$$\text{Parents}(D) = \{A, C\}$$
$$\text{Descendants}(A) = \{B, C, D, F\}$$
$$\text{Adjacencies}(D) = \{A, C, F\}.$$

A *path* is a sequence of edges where each edge starts at the end of the preceding edge. *path*

A *directed path* is a path in which all arrows point in the same direction, such as $A \to C \to D$ in Figure 3.1. *directed path*

We say a variable X is a *common cause* of variables Y and Z if X is a direct cause of Y and a direct cause of Z with respect to the set $\{X, Y, Z\}$. Note that if X, Y, Z are part of some larger set of variables $\mathbf{V}$, X can be a common cause of Y and Z without necessarily being a direct cause of either with respect to $\mathbf{V}$. Instead, there simply has to exist a directed path $X \to ... \to Y$ that does not contain Z and a directed path $X \to ... \to Z$ that does not contain Y. In Figure 3.1, A is a common cause of C and D but not of D and E. Finally, a set of variables $\mathbf{V}' \subset \mathbf{V}$ is *causally sufficient* if every common cause of any two variables in $\mathbf{V}'$ is also in $\mathbf{V}'$. *common cause* *causally sufficient*

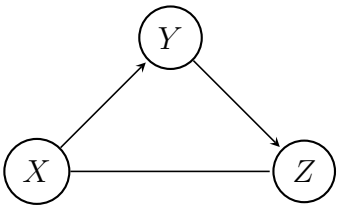


Figure 3.2: An example of a PDAG.

3.2 Causal discovery algorithms

Causal discovery algorithms aim to infer the causal structure of a set of variables using observational (or sometimes a mix of observational and experimental) data. These algorithms are typically described using graphical language and aim to infer an equivalence class of DAGs containing the DAG that describes the true causal structure. The output of these algorithms is an equivalence class, rather than a single DAG, because in many cases there is not enough information to identify the exact structure. For example, in a set $\mathbf{V} = \{X, Y\}$ of just two variables, algorithms that rely on conditional independence information are unable to distinguish between $X \to Y$ and $Y \to X$ (although with additional assumptions this may be possible [101]). These equivalence classes (defined in more detail in Section 3.2.2) can be represented by completed partially directed acyclic graphs. These graphs are defined as follows.

Definition 1. A *partially directed acyclic graph (PDAG)* is a graph in which edges are either directed or undirected and no directed path forms a cycle. A *completed PDAG (CPDAG)* is a PDAG in which every undirected edge can be individually oriented in either direction without creating a directed path that forms a cycle.

PDAG

CPDAG

In other words, a CPDAG is a PDAG in which every edge that can be oriented (based on the stipulation that the graph is acyclic) has been. For example, the graph shown in Figure 3.1 is a PDAG but not a CPDAG because orienting the edge between X and Z as $Z \to X$ would create a cycle.

For causal discovery from observational data to be possible, there must be a way to

connect the underlying causal structure of a set of variables to properties of their joint probability distribution. The most common way to do this is through the introduction of the following additional axioms—the causal Markov condition and faithfulness.

3.2.1 Causal Markov condition

Let $\mathbf{V}$ be a set of random variables and let G be a DAG over these variables (in other words, each node in G represents an element of $\mathbf{V}$). The joint distribution over the set of variables $\mathbf{V}$ is said to satisfy the *causal Markov condition* (often simply called the *Markov condition*) with respect to G if any variable $X \in \mathbf{V}$ is independent of any other variable in $\mathbf{V}$ that is not a descendant or parent of X when conditioning on the parents of X (the set Parents(X)) [40]. This definition can be simplified slightly by noting that X is trivially independent of any parent variable given its parents. Therefore the causal Markov condition holds if any variable is independent of any non-descendant given its parents.

causal Markov condition

The causal Markov condition is the fundamental assumption that relates DAGs to joint distributions. From any DAG G we can read off a list of independence relations that exist in all distributions which satisfy the causal Markov condition with respect to G. For example, in the DAG in Figure 3.1 the Markov condition implies the following relations:

$$B \perp\!\!\!\perp C, D, E, F \mid A$$

$$C \perp\!\!\!\perp B \mid A$$

$$D \perp\!\!\!\perp B \mid A, C$$

$$E \perp\!\!\!\perp A, B, C, D$$

$$F \perp\!\!\!\perp A, B, C \mid D, E$$

These statements were arrived at by choosing a variable (e.g. B) and then identifying

its non-descendants (C, D, E, F), which it must be independent of given its parents (A). Hence there is no conditional independence statement here for A, since it has no non-descendants. However, this procedure will not necessarily (and does not in this case) list every conditional independence relation implied by the causal Markov condition, as discussed further in the next section.

We can begin to see how this could be used for causal discovery; if Figure 3.1 was the true underlying DAG for these variables and we assume the joint distribution satisfies the Markov condition with respect to the true DAG, then we would expect these relationships to be reflected in the data. If we know that C and B are dependent when conditioning on A, for example, then we would know that Figure 3.1 cannot be the true causal structure. Of course, in practice we can never 'know' this statement about conditional dependence but we can evaluate the evidence for it in the data.

There are some philosophical disagreements on the validity of the causal Markov condition, in particular in indeterministic contexts (see [102, 103] for a detailed discussion and defence of this condition). On a more practical note, this condition can fail to hold if the data are a mixture of observations from sub-populations with different joint distributions (even if the condition holds for each distribution separately) [104, 40]. Nevertheless, this condition is assumed by many causal discovery algorithms [41].

3.2.2 Faithfulness

The causal Markov condition describes the independence relations that are implied by a graphical causal structure. The faithfulness condition acts in reverse, defining the dependence relationships expected in the joint distribution. It is defined as follows. Let $\mathbf{V}$ be a set of random variables with joint distribution P and let G be a DAG over the set $\mathbf{V}$. Then the distribution P and graph G satisfy the *faithfulness* condition if the only independence relations in P are those implied by the causal Markov condition

faithfulness

on G. Equivalently we can say that the distribution P is *faithful* to G.

Faithfulness ensures that the only independence relations that exist in the joint distribution are due to a variable being screened off by its direct causes. In the previous section we listed a number of independence relations implied by Figure 3.1 under the causal Markov condition. As discussed, this was not an exhaustive list as the causal Markov condition can imply other independence relations that are not immediately obvious from the definition. Clearly to understand the implications of the faithfulness condition it is necessary to know all the independence relations implied by the causal Markov condition. This is made easier using the concept of d-separation, which is described in the next section.

In the absence of faithfulness, a distribution in which variables X and Y are independent could be consistent (in terms of independence relations implied by the causal Markov condition) with, for example, X being a direct cause of Y. In the previous section, we saw that the causal Markov condition allows us to reject candidate causal structures that imply an independence relation that is not reflected in the data. With faithfulness, we can also reject a causal structure if it does not imply an independence relation that is found in the data.

The faithfulness condition also rules out the 'cancelling out' of effects. For example, it states that in Figure 3.1 under faithfulness the effects of A and C on B will not exactly cancel out to leave no dependence between A and B or A and C. This is often justified by arguing that such effects cancelling out could only result from highly specific and improbable coincidences. For example, if the effect of each variable on any other is linear it might be argued that the set of effect sizes that violate faithfulness is of measure zero [105]. Faithfulness is almost universally assumed in constraint-based causal discovery algorithms but some score-based algorithms assume only the causal Markov condition [41]. Having defined these two conditions, we can define the equivalence classes that causal discovery algorithms aim to infer.

Definition 2. Let $\mathbf{X}$ be a set of random variables and denote the set of DAGs over this set by $\mathbf{G}$. Define an equivalence relation on this set by saying G and H in $\mathbf{G}$ are equivalent if they imply the same conditional independence relations on the joint distribution over $\mathbf{X}$ under the causal Markov condition and faithfulness.

For discovery algorithms that do not assume faithfulness, this definition can be modified to depend on conditional independence relations implied only by the causal Markov condition. An equivalence class $\mathbf{G}' \subset \mathbf{G}$ can be represented by a CPDAG, which is constructed as follows. First construct a PDAG over $\mathbf{X}$ such that for any pair $X_i, X_j \in \mathbf{X}$ there exists

- a directed edge $X_i \to X_j$ if this directed edge exists in every $G \in \mathbf{G}'$,
- no edge between X_i and X_j if in every $G \in \mathbf{G}'$ there does not exist an edge between them, and
- an undirected edge between X_i and X_j otherwise.

The CPDAG is then generated by orienting any undirected edge in this PDAG in a given direction if the opposite direction would create a cycle.

3.2.3 d-separation

In Section 3.2.1 we listed a number of conditional independence relations that are implied by the DAG in Figure 3.1 under the causal Markov condition. However, there are other independence relations that seem plausible. For example, the causal Markov condition implies that F is independent of A, B and C (variables that are neither parents nor descendants of F) given D and E (the parents of F). Since E is a cause of F but otherwise unrelated to all other variables, we might intuitively expect that conditioning on D alone would be sufficient to screen off F from A, B or C and therefore that F is independent of A, B, C given D, and indeed this is also implied by

the causal Markov condition. Our reasoning here was fairly imprecise, however, and we might have trouble identifying more complex independence relations. d-separation, proposed by Pearl, is a graphical condition that characterises exactly the conditional independence relations implied by the Markov condition [106, 107, 108].

To motivate the definition of d-separation informally, consider a set of random variables $\mathbf{V}$ and variables $X, Y \in \mathbf{V}$. We are interested in understanding when X and Y are dependent, that is in which situations knowing the value of X influences our expectation for Y (or vice versa). In any DAG over the set $\mathbf{V}$, there are two possible kinds of paths between X and Y:

1. A path containing a *collider structure*, that is a structure $Z_1 \rightarrow Z_0 \leftarrow Z_2$,

$$X \rightarrow ... \rightarrow Z_1 \rightarrow Z_0 \leftarrow Z_2 \leftarrow ... \leftarrow Y.$$

   Note that we could have $Z_1 = X$ or $Z_2 = Y$, or both. In this structure, the variable Z_0 is said to be a *collider*. *collider*

2. A path that does not contain a collider. This can be either

   (a) a directed path $X \rightarrow ... \rightarrow Y$ or $Y \rightarrow ... \rightarrow X$, or

   (b) the union of directed paths from a common cause to both X and Y,

$$X \leftarrow ... \leftarrow Z \rightarrow ... \rightarrow Y.$$

We first focus on the latter—paths that do not contain a collider. Consider a directed path from X to Y. If no other variables are conditioned on then information can 'flow' in both directions—knowing the value of X informs what value we expect Y to take and vice versa. Returning to the example in Section 3.1, in which temperature caused ice on the road which in turn caused traffic accidents, if we know temperatures are low we would expect more accidents, and if we know there have been lots of accidents

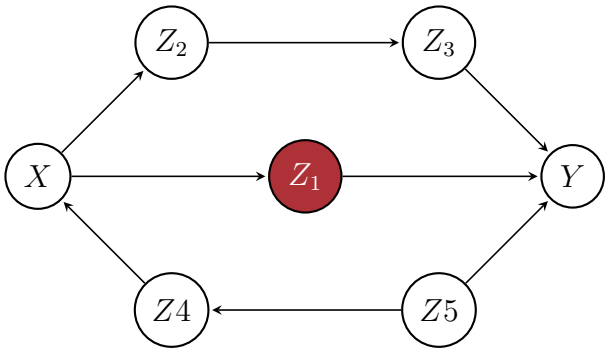


Figure 3.3: An example of a DAG with three paths between X and Y, none of which contain a collider. There are two directed paths between X and Y and one which contains a common cause, Z_5, of X and Y.

we might suspect that it was cold. However, knowing the value of any variable on this path blocks this flow of information, as knowing whether there was ice on the roads did in our example. Now consider a path of the type described in 2(b), containing a common cause of X and Y. If no variables are conditioned on, information can flow backwards from X to the common cause, Z, and then forwards to Y, and vice versa. However, if a variable on the path from Z to X is conditioned on then this stops the flow of information between X and Z and therefore also between X and Y. Conditioning on a variable on the path from Z to Y has a similar effect. Therefore, as with the directed paths, conditioning on any variable on this path is sufficient to stop the flow of information along this path. If there are multiple paths between X and Y that do not contain colliders (as in Figure 3.3) then conditioning on a variable on just one of these paths (such as Z_1 in Figure 3.3, shown in red) will still allow information to flow via the other paths. To stop the flow of information entirely we need to condition on at least one variable on each path of this kind.

Now we consider a path between X and Y containing a collider. If none of the variables in this path are conditioned on, then there is no flow of information between X and Y via this path. For example, in the path $X \rightarrow Z \leftarrow Y$ the existence of the common effect Z does not mean that knowing X provides any information about Y.

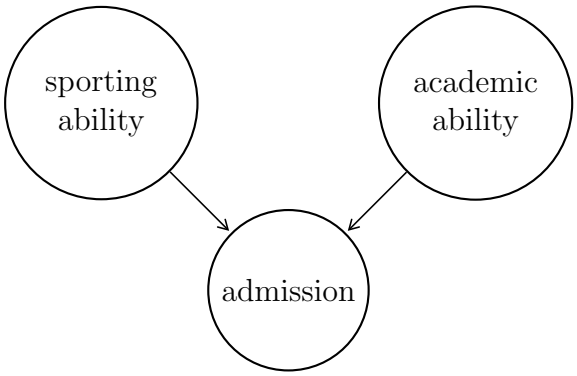


Figure 3.4: Corresponding DAG for collider bias example.

However, if the value of Z is known then further knowing the value of X *can* affect our expectation for Y. This unintuitive result, that conditioning on a common effect can induce dependence between two variables, is known variously as *collider bias*, *selection bias*, *Berkson's paradox* [109] or the *explaining away effect* [110]. A deterministic example of this phenomenon is a circuit containing a light and two switches connected in series (so that both switches must be on for the light to turn on). If we know that the light is off, then if one switch is on we can determine that the other must be off. For another example, suppose a selective school admits students who exceed some standard of sporting or academic ability (see Figure 3.4). If sporting and academic ability are independent in the population as a whole, when restricted to pupils at this school (equivalent to conditioning on a variable representing admission to the school) there will be an apparent negative correlation between sporting and academic ability. This is because the pupils who do not fulfil the academic requirement will necessarily be good at sports, and students who are not good enough at sports must meet the academic requirements (this is illustrated in Figure 3.5). Therefore while the presence of a collider on a path prevents the flow of information, conditioning on this collider (and similarly conditioning on any descendant of the collider) allows the flow of information again. If there are multiple colliders on a path then conditioning on every collider will allow the flow of information.

collider bias

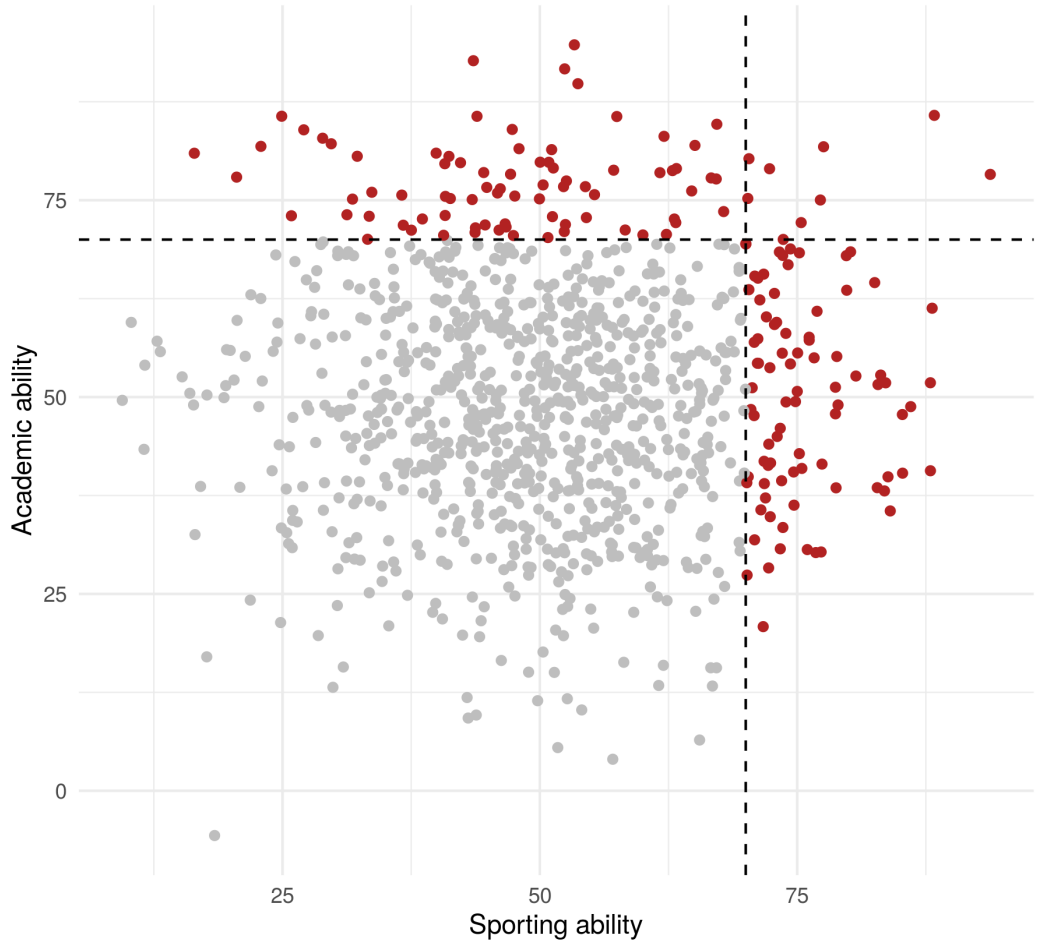


Figure 3.5: Illustration of the negative correlation induced by only considering the subset of the population (shown in red) with sporting or academic ability above certain thresholds. The rest of the population is shown in grey.

We can now present the formal definition of d-separation.

Definition 3. Let G be a DAG over vertex set $\mathbf{V}$. Let $X, Y \in \mathbf{V}$ be a pair of vertices and $\mathbf{W} \subset \mathbf{V}$ be a subset of vertices not containing X or Y. Let U be a path between X and Y. Then X and Y are *d-separated* given $\mathbf{W}$ with respect to path U if

d-separated

1. there is a collider on U not in $\mathbf{W}$ that does not have a descendant in $\mathbf{W}$, or
2. there is a variable on U in $\mathbf{W}$ that is not a collider.

If X and Y are d-separated given $\mathbf{W}$ with respect to every path between them then X and Y are d-separated given $\mathbf{W}$. If X and Y are not d-separated given $\mathbf{W}$ then they are *d-connected* given $\mathbf{W}$.

From our previous discussion, we can see that these two conditions are exactly the conditions needed in order for conditioning on $\mathbf{W}$ to stop the flow of information between X and Y on a given path. Therefore if X and Y are d-separated given $\mathbf{W}$ then we would expect conditioning on $\mathbf{W}$ to stop all flow of information, or in other words to make X and Y independent. In fact, this is exactly what is implied by the causal Markov condition, as stated in the following theorem (a rephrasing of Theorem 1.2.4, [38]).

Theorem 3.2.1. Let P be a distribution over variables $\mathbf{V}$ with causal structure described by DAG G and suppose the causal Markov condition holds. Let X, Y be variables in $\mathbf{V}$ not contained in $\mathbf{W} \subset \mathbf{V}$. Then if X and Y are d-separated given $\mathbf{W}$ in graph G, X and Y are independent conditional on $\mathbf{W}$. Furthermore, if X and Y are not d-separated given $\mathbf{W}$ then the causal Markov condition does not imply that X and Y are independent given $\mathbf{W}$.

This theorem tells us that the graphical condition of d-separation characterises exactly the conditional independence relations implied by the causal Markov condition. Note that under the causal Markov condition, if X and Y are not d-separated

we do not know whether X and Y are independent given $\mathbf{W}$. We require faithfulness to rule out independence in this case:

Corollary 3.2.1. Assume the conditions of Theorem 3.2.1 and additionally that P is faithful to G. Then X and Y are independent given $\mathbf{W}$ if and only if X and Y are d-separated by $\mathbf{W}$ in G .

This corollary follows trivially from Theorem 3.2.1 and the definition of faithfulness. Spirtes et al. [40] provide the following characterisation of faithfulness, which is useful for designing causal discovery algorithms.

Theorem 3.2.2 (Theorem 3.4, [40])**.** Let P be a distribution over variables $\mathbf{V}$ and let G be a DAG over this set of variables. Then P is faithful to G if and only if

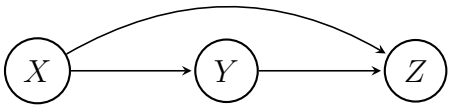
1. for any variables $X, Y \in \mathbf{V}$, X and Y are adjacent in G if and only if X and Y are dependent conditional on every subset $\mathbf{W} \subset \mathbf{V}$ that does not contain X or Y, and

2. for any $X, Y, Z \in \mathbf{V}$ such that X is adjacent to Y, Y is adjacent to Z, and X and Y are not adjacent, the subgraph containing these variables is $X \rightarrow Y \leftarrow Z$ if and only if X and Z are dependent conditional on any set containing Y and not X or Z.

Both statements in this theorem follow from d-separation and Corollary 3.2.1. To show the first statement, we can simply note that if there exists a path $X \rightarrow Y$ or $Y \rightarrow X$ then no set which does not include X or Y can d-separate X and Y with respect to this path. For the second statement, note that if the true subgraph is $X \rightarrow Y \leftarrow Z$ then X and Z cannot be d-separated with respect to this path by any set containing Y but not X or Z.

3.2.4 Score-based algorithms

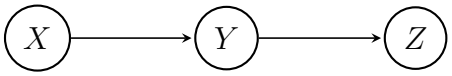
While I do not make use of score-based causal discovery algorithms, I introduce them briefly here for completeness as they are one of the two broad classes of discovery algorithms. Score-based algorithms associate any DAG over a set of random variables with a model structure based on the independence relationships given by the causal Markov condition. For example, for a DAG



the corresponding functional relationship could be written as

$$Y = f(X)$$
$$Z = g(X, Y)$$

whereas for the DAG



the model would simply be

$$Y = f(X)$$
$$Z = g(Y).$$

To evaluate how well each DAG fits with the observed data, additional assumptions are typically made about the functional forms of these relationships (a common approach is to model these relationships as a linear system) and the resulting models are then fit to the available data. The DAG is then scored with some metric of goodness-of-fit of the model, often the Bayesian Information Criterion [111, 112]. The problem is then how to efficiently search for the model with the highest score. Chickering [42]

showed that there exists an efficient greedy algorithm, termed the *greedy equivalence search*, which identifies the correct equivalence class of DAGs in the large sample limit if the true generative model is in this class. This algorithm starts with an empty graph and adds or removes a single edge at each step.

greedy equivalence search

There are two main limitations to score-based methods. The first is that the parametric assumptions in these models may not be appropriate for spatiotemporal epidemiological data. Secondly, no score-based algorithms have been developed to deal with the presence of unobserved confounding variables (unobserved common causes of observed variables) [41]. For these reasons, I instead focused on using constraint-based algorithms.

3.2.5 Constraint-based algorithms

Constraint-based causal discovery algorithms are formulated directly in terms of conditional independence relations, with those observed in the data constraining the set of compatible DAGs under the assumptions of the causal Markov condition and faithfulness. In theory these algorithms are more flexible than score-based methods, avoiding the assumptions required for model formulation, but in practice many commonly used conditional independence tests do require strong parametric assumptions. Later in this chapter, I discuss non-parametric independence tests based on the theory of kernel embeddings. Using these tests within constraint-based algorithms, non-parametric causal discovery can be performed.

In this section, I describe the causal discovery algorithm that I use in my work, the PC algorithm (named after its inventors Peter Spirtes and Clark Glymour) [43, 44]. The PC algorithm is one of the most commonly used constraint-based discovery algorithms. Given joint observations of a set of random variables, the PC algorithm aims to identify the equivalence class of DAGs that contains the DAG representing the true causal structure between these variables. Given consistent conditional independence

tests this algorithm is itself consistent and if the true underlying DAG is sparse then it runs in polynomial time in terms of the number of variables. I begin by describing the simpler SGS algorithm [40], which is similar to the PC algorithm but less efficient. Both the SGS and PC algorithms assume the following.

1. Causal sufficiency of the set of observed variables.
2. The causal relations between variables are the same for every individual in the population.
3. Perfect information about conditional independence relations between variables.

Under these assumptions, both algorithms provably determine the correct equivalence class.

3.2.5.1 The SGS algorithm

The SGS algorithm is another constraint-based causal discovery algorithm that uses joint observations of a set of variables to infer the equivalence class of DAGs containing the true DAG structure. This algorithm is less efficient than the PC algorithm, running in exponential time in the number of variables, but is conceptually simpler. Once we have described the SGS algorithm, the PC algorithm can be understood as a modification of this algorithm.

Before describing this algorithm, it is useful to introduce two new definitions. Firstly, the *skeleton* of a DAG is the graph with all directed edges replaced by undirected ones. For example, the skeleton of the graph $X \to Y \to Z$ is X—Y—Z. Secondly, three vertices X, Y, Z in an undirected graph G are said to be an *unshielded triple* if X, Y and Y, Z are adjacent but X, Z are not.

skeleton

unshielded triple

We can now describe the SGS algorithm. Let $\mathbf{V}$ be the set of variables. The algorithm begins with a complete undirected graph H (where every vertex is connected to every other) on the set $\mathbf{V}$. The first step of the algorithm aims to infer the skeleton

of the true graph. The second and third steps then aim to orient as many of these edges as possible. These steps are as follows:

1. For each pair of vertices X, Y, test X and Y for conditional independence given each subset $\mathbf{S} \subset \mathbf{V}$ that does not contain X or Y. If any test fails to reject independence, remove the edge between X and Y from H.

2. For any unshielded triple X, Y, Z in the skeleton inferred in step one, orient these edges as $X \rightarrow Y \leftarrow Z$ (a collider) if and only if there are no subsets $\mathbf{S} \subset \mathbf{V}$ that contain Y and do not contain X or Z such that X and Z are independent given $\mathbf{S}$.

3. For any edge X, Y, orient this edge as $X \rightarrow Y$ if the orientation $X \leftarrow Y$ would produce either a cycle or a collider. Repeat until no more edges can be oriented.

The first step of this algorithm makes use of the first part of Theorem 3.2.2, which states that an edge between X and Y exists in the true graph if and only if X and Y are dependent conditional on any set of variables not including X or Y. It follows that if any such subset $\mathbf{S}$ is found such that X and Y are independent given $\mathbf{S}$, then there cannot be an edge between X and Y. Similarly, the second step uses the second part of this theorem, the fact that an unshielded triple X, Y, Z is a collider if and only if X, Z are dependent conditional on any set containing Y and not X or Z. Finally, step three relies on the assumption that the true structure is acyclic and the fact that if a collider exists it would have been discovered in the second step.

The result of this algorithm is a CPDAG. This represents the equivalence class of DAGs that can be produced by orienting the undirected edges (although not all combinations of orientations produce valid DAGs).

When using consistent conditional independence tests, the SGS algorithm is consistent in the following sense. Let $\hat{G}_n$ be the equivalence class that is the result of

applying the algorithm with n independent and identically distributed joint observations and let G be the equivalence class containing the true DAG. Then,

$$\lim_{n\to\infty} \mathrm{P}(\hat{G}_n \neq G) = 0.$$

In other words, the algorithm converges in probability to the correct equivalence class [40]. However, this algorithm is inefficient, as it has the potential to perform many conditional independence tests that, due to the results of previous tests, are unnecessary. For example, suppose that part way through the first step, all edges involving a variable X have been tested and removed. Suppose that Y, Z is a subsequent edge that has not yet been tested. Since we have inferred that X is not connected to any other variable in the true underlying DAG, by considering the definition of d-separation we can see that if Y and Z are d-separated by some set $\mathbf{W}$ that contains X, they must also be d-separated by $\mathbf{W} \setminus \{X\}$. Therefore, when testing the edge Y, Z, there is no need to condition on any set containing X. Similar redundant tests can occur in the second step. The PC algorithm improves on this by avoiding some of this redundancy.

3.2.5.2 The PC algorithm

The PC algorithm makes fairly minor modifications to the SGS algorithm, which significantly improve efficiency. The algorithm is described in detail in Algorithm 1 below.

The first modification concerns which conditioning subsets $\mathbf{S}$ are used in the first step. When testing an edge between variables X and Y, the PC algorithm tests X and Y for independence conditional on subsets $\mathbf{S} \subset \mathbf{V}$ that are subsets of the adjacencies of X (in the current graph G') that do not include Y or subsets of the adjacencies of Y that do not include X. In other words $\mathbf{S}$ must be a subset of Adjacencies$(G', X)\backslash\{Y\}$

or Adjacencies$(G', Y)\backslash\{X\}$. The claim here is that if two variables X and Y are d-separated by some subset $\mathbf{S}_{XY} \subset \mathbf{V}\backslash\{X, Y\}$ in the true DAG G, then at any stage in the algorithm they must be conditionally independent given some subset of the Adjacencies$(G', X)\backslash\{Y\}$ or Adjacencies$(G', Y)\backslash\{X\}$. To see that this is true, assume without loss of generality that Y is not a descendant of X in the true DAG G (if not, swap X and Y in this argument). We know that Y is also not a parent of X because if it was there could not exist a subset $\mathbf{S}_{XY}$ that d-separates X and Y. Therefore, by the causal Markov condition, X is independent of Y given Parents(G, X). Finally, we can note that at any given point in the first step, the set Parents(G, X) is contained in Adjacencies$(G', X)\backslash\{Y\}$.

The first step of the PC algorithm therefore only tests each edge X, Y for independence given subsets of Adjacencies$(G', X)\backslash\{Y\}$ or Adjacencies$(G', Y)\backslash\{X\}$. The size of the conditioning subsets starts at zero, so each edge is tested for unconditional independence (removing an edge if independence is found). The conditioning subset size is then increased to 1, and each edge X, Y that remains is tested for independence conditional on all subsets of Adjacencies$(G', X)\backslash\{Y\}$ or Adjacencies$(G', Y)\backslash\{X\}$ of size 1. The conditioning subset size is increased again and this process repeats until each edge has either been removed or has been tested against all subsets of Adjacencies$(G', X)\backslash\{Y\}$ or Adjacencies$(G', Y)\backslash\{X\}$. Whenever an edge X, Y is removed the graph G' is updated and the conditioning subset $\mathbf{S}$ that resulted in this edge being removed is recorded as Sepset(X, Y) and Sepset(Y, X), which is used in the second step.

The second step is also modified from the SGS algorithm, in order to require no conditional independence information beyond that discovered in step one. In this step, an unshielded triple X, Y, Z is oriented as a collider, $X \rightarrow Y \leftarrow Z$, if Y is not in Sepset(X, Z). Note that Sepset(X, Z) must exist as there is no edge between X and Z. The claim here is that Y is a collider in the true DAG if and only if Y is not in

Sepset(X, Z). We can consider each direction in this double implication separately:

- By the second part of Theorem 3.2.2, Y is a collider in the true DAG only if X and Z are dependent conditional on every subset containing Y and not X and Z. Since X and Z are independent conditional on Sepset(X, Z), this means that Y is a collider *only if* Y is not in Sepset(X, Z).

- On the other hand, suppose Y is not a collider. The other possibilities for this path are $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$ and $X \leftarrow Y \rightarrow Z$. For any of these options, X and Z cannot be d-separated with respect to this path by any subset that does not contain Y. Therefore if Y is not a collider then Y must be in Sepset(X, Z), which is the same as saying Y is a collider *if* Y is not in Sepset(X, Z).

This double implication ensures that not only will this second step correctly identify colliders but that it will identify all colliders. The third step is identical to that of the SGS algorithm, orienting any undirected edges if the opposite direction would lead to a cycle or collider.

The PC algorithm requires fewer conditional independence tests than the SGS algorithm. The scale of this improvement depends on how much smaller the sets Adjacencies$(G', X)\backslash\{Y\}$ or Adjacencies$(G', Y)\backslash\{X\}$ are than $\mathbf{V}\backslash\{X, Y\}$ for each edge X, Y at the relevant step in the algorithm. It can be shown that the complexity of the PC algorithm is bounded by $\mathcal{O}(n^q)$ where n is the number of variables and q is the maximum number of adjacencies of any variable in the true DAG [113]. Where the underlying structure is sparse (and therefore q does not grow or grows slowly with the number of variables) this algorithm is therefore of polynomial complexity in the number of variables.

The exact algorithm, as defined by Spirtes et al. [40], is as follows. First, define CondSet(X, Y) to be the union of all subsets of Adjacencies$(G', X)\backslash\{Y\}$ and subsets

of Adjacencies$(G', Y)\backslash\{X\}$. Also define maxCondSetSize(X, Y) to be the maximum size of any of these sets. Then the algorithm can be written as in Algorithm 1.

Initialisation
Form complete undirected graph G;

Step 1
$n = 0$;
while *there exists an edge* X, Y *with* maxCondSetSize$(X, Y) \geq n$ **do**
  for *each such edge* X, Y **do**
    for *each* $S \in$ CondSet(X, Y) *of size* n **do**
      if X, Y *are independent conditional on* S **then**
        remove edge X, Y;
        record S as Sepset(X, Y);
      end
    end
  end
  $n = n + 1$
end

Step 2
for *each unshielded triple* X, Y, Z **do**
  if Y *is in* Sepset(X, Z) **then**
    orient as $X \rightarrow Y \leftarrow Z$;
  end
end

Step 3
orient edges where other direction would generate a cycle or a collider, repeat until no more edges can be oriented;

Algorithm 1: The PC algorithm

When implementing this algorithm with a finite sample, conditional independence relations can be incorrectly inferred and the effect of these errors can propagate through the rest of the algorithm. In theory each step is unstable, although in practice the first step is much more reliable than the second or third [40]. Therefore we may have more confidence in the existence or non-existence of direct causal rela-

tionships between variables and less certainty about their direction. In some cases, directions can be inferred from contextual information (such as the timing of events), and ambiguities or errors in the output equivalence class thereby "manually" resolved.

Many additional algorithms for causal discovery have been developed which modify the PC algorithm for different purposes. Spirtes et al. [40] describe the PC* algorithm, which requires even fewer tests by noting that when testing an edge between X and Y, we only need to consider conditioning on variables that lie on a path between X and Y (rather than all variables adjacent to either). With perfect information about conditional independence relations, the PC and PC* algorithms produce the same output, but with fallible tests we would expect the PC* algorithm to make fewer errors (as it requires fewer tests). However, the additional computational overhead required by the PC* algorithm is usually impractical or not worth the lower error rate. With imperfect information the result of the PC algorithm often depends on the order in which independence tests are performed but order-independent variants have been proposed [114, 115]. In my proposed causal discovery procedure (see Section 3.5), the PC algorithm is run many times and ordering effects are overcome by changing the order of variables in each repeat.

Another important modification is the fast causal inference (FCI) algorithm, which is similar to the PC algorithm but allows for the possibility of unmeasured common causes [40]. While this may seem like a significant improvement over the PC algorithm, the outputs of the FCI and PC algorithms are usually similar. In fact, the skeleton of the output of each algorithm will be the same but the FCI may direct fewer edges or identify edges where a common cause may exist. In my implementation of the causal discovery procedure I used the PC algorithm throughout for simplicity. However, the FCI algorithm could be used in its place with no other modification to the method. Variants for the FCI algorithm have also been proposed, such as the RFCI algorithm [116] which is faster but less informative.

3.2.6 Two-step discovery algorithm

In my applications, I use the PC algorithm twice, first to infer the causal structure between environmental and socioeconomic covariates and then subsequently to identify the interactions between these variables and malaria risk. The PC algorithm could simply be used once with all variables but in this specific context this two-step discovery method allows for greater flexibility regarding which locations and time points are used for observations of environmental and socioeconomic variables.

In this scenario, we have joint observations of some measure of malaria risk (typically prevalence or incidence information), M, and socioeconomic and environmental variables $X_1, ..., X_n$. Let $\mathbf{V} = \{M, X_1, ..., X_n\}$ and G be the DAG representing the true causal structure. The socioeconomic and environmental variables are available at a high spatial and temporal resolution, as these variables are measured on a fine grid of pixels across the world, monthly or every 8 days for variables that vary over time. On the other hand, the malaria incidence or prevalence information is limited by the data set available, for example prevalence information may only be available at a small number of survey locations in the time frame of a few months.

I assume that malaria risk is not a cause of any other variable in this model. This means that if two variables X_i and X_j are d-separated by some subset $\mathbf{S} \in \mathbf{V} \backslash \{X_i, X_j\}$ then they are also d-separated by a subset $\mathbf{S}' \in \mathbf{V} \backslash \{M, X_i, X_j\}$. Let G' be the DAG created by removing the variable M from G. The consequence of this is that the equivalence class containing G' can be discovered by applying the PC algorithm to the set $\mathbf{V}' := \{X_1, ..., X_n\}$. Note that this is only true because M cannot be a common cause of any other two variables.

Furthermore, I assume that a variable cannot be a cause of a variable that occurs earlier in time. This is most common when considering dynamic variables at different time lags, for example I assume that a dynamic variable (such as rainfall) with a one month timelag (i.e. rainfall one month prior to the time of the observation) cannot

be a cause of any other dynamic variable (such as rainfall or temperature) at a two month timelag. Additionally, I assume that dynamic variables cannot cause static variables (such as elevation).

I therefore first apply the PC algorithm to the set $\mathbf{V}'$ to discover the causal structure between these variables. To include the assumptions in the previous paragraph, the initial complete graph over $\mathbf{V}'$ contains directed edges (instead of undirected ones) where the causal direction is assumed. The advantage of this first application of the PC algorithm is that number and location of observations is not limited by the malaria incidence or prevalence data. Therefore, large sample sizes with good coverage of the area of interest can be used, producing more reliable results.

Let G_1 be the CPDAG representing the result of this step. This graph is then augmented by adding a node representing M (the measure of malaria risk) and directed edges from each variable X_i to M. I then use a modified version of the PC algorithm that only tests an edge if it contains M and uses the causal structure G_1 to define adjacency sets. Note that since we assume the direction of all these edges, only the first step of the PC algorithm is needed in this case. The result of this second application of the PC algorithm is the final result.

3.3 Kernel-based independence tests

In Section 1.2, we saw that kernel functions can be used to map data points into higher dimensional Hilbert spaces. These mappings can encode relevant characteristics of the data in a way that is more convenient. In the kernel support vector machine example in Section 1.2, the image of the feature map included the distance of each data point from the origin as the third coordinate, allowing the linear classifier to make use of this information. In this example, the feature space (the Hilbert space that each data point is embedded into) was three-dimensional. This relatively low-dimensional space

was sufficient because of the simple structure of the problem (the true classification was based solely on distance from the origin) and the careful choice of feature map—it required us to notice the pattern in the data and choose a feature map that included the relevant information. More commonly high or infinite-dimensional feature spaces are chosen, which result in more flexible algorithms that require less prior knowledge.

The main application of these embeddings in this work is in non-parametric independence tests, which involve embedding entire probability distributions into Hilbert spaces. Embedding probability distributions in this way allows us to define distances between distributions without the need for density estimation and to define various covariance and cross-covariance operators on these spaces that encode information about dependencies. Many results from statistical learning theory can be applied to inform the convergence of estimators. In this section, I introduce the theory of reproducing kernel Hilbert spaces and kernel mean embeddings and then describe two kernel-based tests of independence, the Hilbert Schmidt independence criterion and the randomized conditional independence test.

3.3.1 Reproducing kernel Hilbert spaces

So far we have discussed mapping data points and probability distributions to points in high dimensional Hilbert spaces. In fact, kernel methods involve a specific class of Hilbert spaces, reproducing kernel Hilbert spaces (RKHSs). This is a wide class and we will see that there is a simple condition for a Hilbert space to be a RKHS (that every evaluation functional is bounded) and for a kernel function to correspond to a RKHS (that the kernel is positive definite). I begin by defining a RKHS in the conventional way, in terms of bounded evaluation functionals. After this I describe how to construct the corresponding RKHS from a chosen kernel function.

Hilbert space

Definition 4. A vector space $\mathcal{H}$ is a *Hilbert space* if it is a complete inner product space. That is, every Cauchy sequence in $\mathcal{H}$ converges in $\mathcal{H}$ with respect to the norm

induced by the inner product. This norm is defined as

$$\|x\|_{\mathcal{H}} = \sqrt{\langle x, x \rangle_{\mathcal{H}}}$$

for any $x \in \mathcal{H}$. All Hilbert spaces discussed here are assumed to be real (rather than complex) vector spaces.

Definition 5. Let $\mathcal{H}$ be a Hilbert space of functions from some set $\mathcal{X}$ to $\mathbb{R}$. The *evaluation functional* associated with some point $x \in \mathcal{X}$, δ_x, is the functional that evaluates functions in $\mathcal{H}$ at that point. In other words $\delta_x(f) = f(x)$ for any $f \in \mathcal{H}$.

evaluation functional

Definition 6 (Reproducing kernel Hilbert space [117])**.** A Hilbert space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ is a *reproducing kernel Hilbert space* if for all $x \in \mathcal{X}$ the evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ is bounded. That is, for each $x \in \mathcal{X}$ there exists some $C_x \in \mathbb{R}_{>0}$ such that

$$|\delta_x(f)| = |f(x)| \leq C_x ||f||_{\mathcal{H}}.$$

reproducing kernel Hilbert space

This is equivalent to a condition of continuity of the evaluation functional, so $f_n \xrightarrow{\mathcal{H}} f$ implies $\delta_x(f_n) \to \delta_x(f)$, i.e. if two functions converge in norm in a RKHS then they also converge pointwise. One interpretation of these spaces is that they are Hilbert spaces in which the functions are 'well-behaved'.

Given a RKHS, $\mathcal{H}$, as defined above, we can define the corresponding kernel function, k, to be the function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ such that

1. for any $x \in \mathcal{X}$, $k(x, \cdot)$ is an element of $\mathcal{H}$, and
2. for any $f \in \mathcal{H}$ and $x \in \mathcal{X}$ we have the *reproducing property*:

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x).$$

reproducing property

The reproducing property tells us that function evaluation of any $f \in \mathcal{H}$ at any point

$x \in \mathcal{X}$ is equivalent to the inner product with a specific element of $\mathcal{H}$. This also implies that $\langle k(x,\cdot),\, k(y,\cdot)\rangle_{\mathcal{H}} = k(x,y)$ and $\|k(x,\cdot)\|^2_{\mathcal{H}} = k(x,x)$ for all $x, y \in \mathcal{H}$. That such a kernel function exists follows from the Riesz representation theorem.

Theorem 3.3.1 (Riesz representation theorem)**.** Let $\mathcal{H}$ be a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ and T be a bounded linear functional on this space. Then there exists some $g \in \mathcal{H}$ such that for any $f \in \mathcal{H}$,

$$T(f) = \langle f, g\rangle_{\mathcal{H}}.$$

A kernel k satisfying the above properties can therefore be constructed as follows. First define k_x for each $x \in \mathcal{X}$ to be the element of $\mathcal{H}$ such that

$$\langle f, k_x\rangle_{\mathcal{H}} = f(x) = \delta_x(f)$$

for any $f \in \mathcal{H}$. Since the evaluation functional is bounded, this element exists by the Riesz representation theorem. The required kernel function can then be defined as $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$k(x,y) = k_x(y).$$

In addition to satisfying the above properties, this kernel is also positive definite, in the following sense.

Definition 7. A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *positive definite kernel* if

positive definite kernel

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{3.1}$$

for any $n \in \mathbb{N}$, $c_1, ..., c_n \in \mathbb{R}$ and $x_1, ..., x_n \in \mathcal{X}$.

To see that the kernel k as defined is indeed positive definite, note that for any

$n \in \mathbb{N}$, $c_1, ..., c_n \in \mathbb{R}$ and $x_1, ..., x_n \in \mathcal{X}$, the function

$$\sum_{i=1}^{n} c_i k(x_i, \cdot)$$

is an element of $\mathcal{H}$ with norm equal to the left hand side of (3.1). This norm is non-negative and therefore (3.1) holds.

We can additionally define the *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ associated to this RKHS to be

feature map

$$\phi(x) = k(x, \cdot)$$

for any $x \in \mathcal{X}$ and we may refer to elements $\phi(x)$ as *feature vectors*. By the reproducing property we have the kernel trick (as discussed in Section 1.2) of computing inner products between feature vectors using the kernel,

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

This allows any algorithm that only requires the computations of inner products (such as support vector machines or ridge regression) to be applied in feature space using kernel evaluations (rather than first computing the feature vectors and then inner products). Similarly note that the reproducing property then implies that for any $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}. \tag{3.2}$$

In practice, it is common to directly select the kernel function used in an algorithm, rather than deriving it from a chosen RKHS. Therefore, it is natural to ask which kernel functions correspond to RKHSs. We have seen that such a kernel must be positive definite. In fact this is a sufficient condition and there is a one-to-one correspondence between positive definite kernels and RKHSs. This is the Moore-Aronszajn theorem:

Theorem 3.3.2 (Moore-Aronszajn theorem [117])**.** For any set $\mathcal{X}$, a function k :

$\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if and only if it is the kernel of some RKHS.

Given a positive definite kernel k on some (non-empty) set $\mathcal{X}$, the construction of the corresponding RKHS, $\mathcal{H}$, can be outlined as follows. Define $\tilde{\mathcal{H}}$ to be the vector space linearly spanned by the functions $\{k(x, \cdot) : x \in \mathcal{X}\}$ equipped with inner product

$$\left\langle \sum_i c_i k(x_i, \cdot), \sum_j c_j k(x_j, \cdot) \right\rangle_{\tilde{\mathcal{H}}} := \sum_{i,j} c_i c_j k(x_i, x_j).$$

The space $\mathcal{H}$ is then generated by completing $\tilde{\mathcal{H}}$ with respect to the norm induced by this inner product.

A trivial example of a kernel function on $\mathcal{X} = \mathbb{R}^n$ is the *linear kernel*,

$$k(x, y) = x^T y$$

in which case the corresponding RKHS space can be identified with $\mathbb{R}^n$ (the feature map being the identity). More common examples of kernels on $\mathbb{R}^n$ include:

1. *Polynomial kernel*

   $$k(x, y) = (x^T y + c)^d$$

   where $c \geq 0$ and $d \in \mathbb{N}_{>0}$.

2. *Radial basis function kernel* (also known as the *Gaussian* or *squared exponential kernel*),

   $$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

   where $\sigma > 0$.

3. *Laplace kernel*

   $$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$

   where $\sigma > 0$.

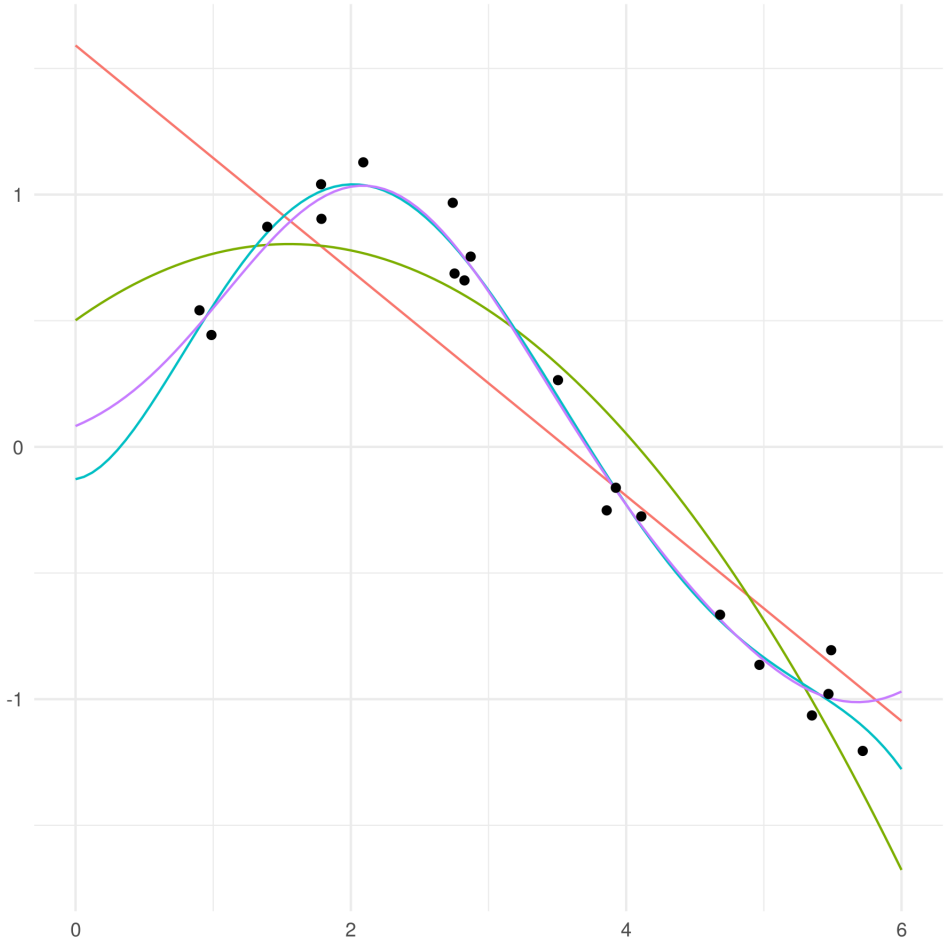


Figure 3.6: Mock data (black points) fit using kernel ridge regression with a linear kernel (red), polynomial kernel with $c = 2, d = 2$ (green), polynomial kernel with $c = 1, d = 5$ (blue) and radial basis function kernel with $\sigma = 1$ (purple).

The RKHS corresponding to the polynomial kernel is of dimension on the order n^d (with n and d as above), while for the Gaussian and Laplace kernel these spaces are infinite dimensional. Figure 3.6 shows mock data fit with kernel ridge regression using different kernels. We can see that in general the regression model appears to be more flexible as the dimension of the RKHS increases, although this does depend on the specific kernels (including the values of parameters such as c and σ). While I have only used RKHSs defined on subsets of $\mathbb{R}^n$ in this work, a notable use of kernel embeddings is to define distance measures on sets $\mathcal{X}$ which lack natural metrics. Examples include kernel functions on graphs or text [118, 119, 120].

3.3.2 Bochner's theorem

The kernel trick allows kernel methods to be implemented without explicitly computing the embeddings of data points or distributions into the high (possibly infinite) dimensional Hilbert space and it is this property that makes these algorithms computationally tractable (and in many cases, computationally efficient). However, many of these algorithms still scale poorly with the number of data points, limiting their utility when using large datasets. An important result for approximating these methods towards improved computational efficiency is Bochner's theorem. This result underlies random Fourier feature (RFF) methods for approximating kernel-based techniques with more scalable algorithms [121, 122].

Theorem 3.3.3 (Bochner's theorem [123])**.** Let $\bar{k}$ be a continuous translation-invariant kernel on $\mathbb{R}^d$. Let $k : \mathbb{R}^d \to \mathbb{R}$ be this kernel written as a function of one variable, so

$$\bar{k}(x, y) = k(x - y)$$

which is possible because $\bar{k}$ is translation-invariant. Then $\bar{k}$ is positive definite if and only if k is the Fourier transform of a non-negative measure. That is,

$$k(\delta) = \int_{\mathbb{R}^d} \exp(iw^T\delta)\, \mathrm{d}\Gamma(w)$$

where Γ is a non-negative measure.

Without loss of generality we can assume that the total measure of Γ is one and so it is a probability measure. We can use this result to construct a low dimensional approximation to the RKHS, $\mathcal{H}$, induced by kernel k as follows. By Bochner's theorem

we have,

$$\begin{aligned} k(x-y) &= \int_{\mathbb{R}^d} \exp(iw^T(x-y))\,\mathrm{d}\Gamma(w) \\ &= \int_{\mathbb{R}^d} \cos(w^T(x-y)) + i\sin(w^T(x-y))\,\mathrm{d}\Gamma(w). \end{aligned}$$

If we are only considering real valued kernels, this second term must be zero and so we have,

$$\begin{aligned} k(x-y) &= \int_{\mathbb{R}^d} \cos(w^T(x-y))\,\mathrm{d}\Gamma(w) \\ &= \int_{\mathbb{R}^d} \cos(w^Tx)\cos(w^Ty) + \sin(w^Tx)\sin(w^Ty)\,\mathrm{d}\Gamma(w). \end{aligned}$$

This integral can be approximated by sampling weights $w_1, ..., w_D \overset{i.i.d.}{\sim} \Gamma$ (these weights are in $\mathbb{R}^d$), as

$$k(x-y) \approx \frac{1}{D}\sum_{i=1}^{D} \cos(w_i^Tx)\cos(w_i^Ty) + \sin(w_i^Tx)\sin(w_i^Ty).$$

This can be written as

$$k(x-y) \approx z(x)^Tz(y)$$

where

$$z : x \mapsto \frac{1}{\sqrt{D}}\left(\cos(w_1^Tx),\, \sin(w_1^Tx),\, ...,\, \cos(w_D^Tx),\, \sin(w_D^Tx)\right). \tag{3.3}$$

We can therefore use the map z to approximate the kernel k. Moreover, we can use this map to define a RKHS that is $2D$-dimensional and can be used to approximate $\mathcal{H}$ (the RKHS induced by k), as follows. Define the kernel function k' on $\mathcal{X} = \mathbb{R}^d$,

$$k'(x,y) := z(x)^Tz(y).$$

This kernel is positive definite, since for any $n \in \mathbb{N}, x_1, ..., x_n \in \mathbb{R}^d, c_1, ..., c_n \in \mathbb{R}$ we

have

$$\sum_{i,j=1}^{n} c_i c_j k'(x_i, x_j) = \sum_{i,j=1}^{n} c_i c_j z(x_i)^T z(x_j)$$
$$= \left(\sum_{i=1} c_i z(x_i) \right)^T \left(\sum_{i=j} c_j z(x_j) \right) \geq 0.$$

Therefore by the Moore-Aronszajn theorem this kernel corresponds to a RKHS of functions from $\mathbb{R}^d$ to $\mathbb{R}$, which we denote $\mathcal{H}'$. The corresponding feature map, $\phi_{\mathcal{H}'}$, sends any $x \in \mathbb{R}^d$ to $k'(x, \cdot)$, which can be written

$$\begin{aligned} \phi_{\mathcal{H}'} : \mathbb{R}^d &\longrightarrow \mathcal{H}' \\ x &\longmapsto (\cdot \mapsto z(x)^T z(\cdot)). \end{aligned}$$

The space $\mathcal{H}'$ can be written,

$$\mathcal{H}' = \{ f : \mathbb{R}^d \to \mathbb{R} \mid f(\cdot) = z(\cdot)^T a, a \in \mathbb{R}^{2D} \}$$

and functions in this space can be used to approximate elements of $\mathcal{H}$. This space is $2D$-dimensional, indeed we can define a bijection $\psi : \mathbb{R}^{2D} \to \mathcal{H}'$ by

$$\begin{aligned} \psi : \mathbb{R}^{2D} &\longrightarrow \mathcal{H}' \\ a &\longmapsto (\cdot \mapsto z(\cdot)^T a). \end{aligned}$$

The map z is sometimes referred to as the *random feature map*, with elements under its image known as *random feature vectors*. The basis elements of the feature spaces are the *random Fourier features*. Techniques based on these approximations are known as random Fourier feature methods.

random Fourier features

In fact, strictly speaking $\mathcal{H}'$ is *at most* $2D$-dimensional, as for some values of

the weights w_i this space is of lower dimension and the map ψ is a surjection but not a bijection—for example if one or more weights equal to zero. However, for the commonly used kernels (such as the radial basis function, Laplace and Mátern kernels) and associated probability measures, $\mathcal{H}'$ is almost surely $2D$-dimensional.

Bochner's theorem ensures that every continuous translation-invariant kernel on $\mathbb{R}^d$ corresponds to a probability distribution on $\mathbb{R}^d$ that can be sampled from to construct the map z. Examples of distributions, Γ, corresponding to specific kernels are the normal distribution for the radial basis function kernel and the t-distribution for a Mátern kernel (see e.g. Ton et al. [124], Li et al. [125]). The parameters of these distributions depend on the parameters of the kernel. An example of such an approximation is shown in Figure 3.7. Here $\mathcal{X} = \mathbb{R}$ and $\mathcal{H}$ is the RKHS of functions from $\mathbb{R}$ to $\mathbb{R}$ induced by radial basis function with $\sigma = 1$. The corresponding feature map, $\phi_{\mathcal{H}}$, sends $x \in \mathbb{R}$ to the function

$$y \longmapsto \exp\left(-\frac{(x-y)^2}{2}\right).$$

To approximate this space, we draw weights $w_1, ..., w_D$ from a standard normal distribution and use these to construct the space $\mathcal{H}'$ as described above. Recall that the feature map in this space, $\phi_{\mathcal{H}'}$, sends x to the function

$$y \longmapsto z(x)^T z(y)$$

which can be written

$$y \longmapsto \frac{1}{D}\sum_{i=1}^{D} \cos(w_i(x-y)).$$

Note that we write $w_i(x-y)$ rather than $w_i^T(x-y)$ in this case because $w_i, x, y \in \mathbb{R}$. Figure 3.7 shows this approximation for $x = 0$ compared to the true function for different samples of weights and for samples of different sizes. The quality of the

approximation is lower and more variable when smaller numbers of weights are used and improves as the number increases.

This approximation is used in the next section to define scalable independence and conditional independence criteria. Here I describe another example, approximating kernel ridge regression or Gaussian process regression. First, we review how Gaussian process regression can be thought of as Bayesian linear regression in feature space (see for example Williams and Rasmussen [126] for more details). In the standard linear regression setting, we have observations of a response variable, $y_1, ..., y_n$ and corresponding observations of explanatory variables (or covariates) $\mathbf{x}_1, ..., \mathbf{x}_n$. It is assumed that

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

where $\mathbf{w}$ are coefficients to be learned and $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_n^2)$. Let $\mathbf{y}$ be the vector of response values and $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be the matrix of covariate values. In a Bayesian context, we may place a Gaussian prior on these weights,

$$\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \sigma_p^2 I).$$

In this case, the prior and likelihood are conjugate and the posterior distribution of the coefficients can be shown to be

$$\mathbf{w} \,|\, \mathbf{y}, X \sim \mathrm{N}\left(\frac{1}{\sigma_n^2} A^{-1} X \mathbf{y},\, A^{-1}\right)$$

where $A = \sigma_n^{-2} X X^T + \sigma_p^2 I$. The posterior mean is also the solution to the corresponding (linear) ridge regression problem, i.e. minimising the penalised error expression

$$(y - \mathbf{x}^T \mathbf{w})^T (y - \mathbf{x}^T \mathbf{w}) + \sigma_n^{-2} \|\mathbf{w}\|_2^2.$$

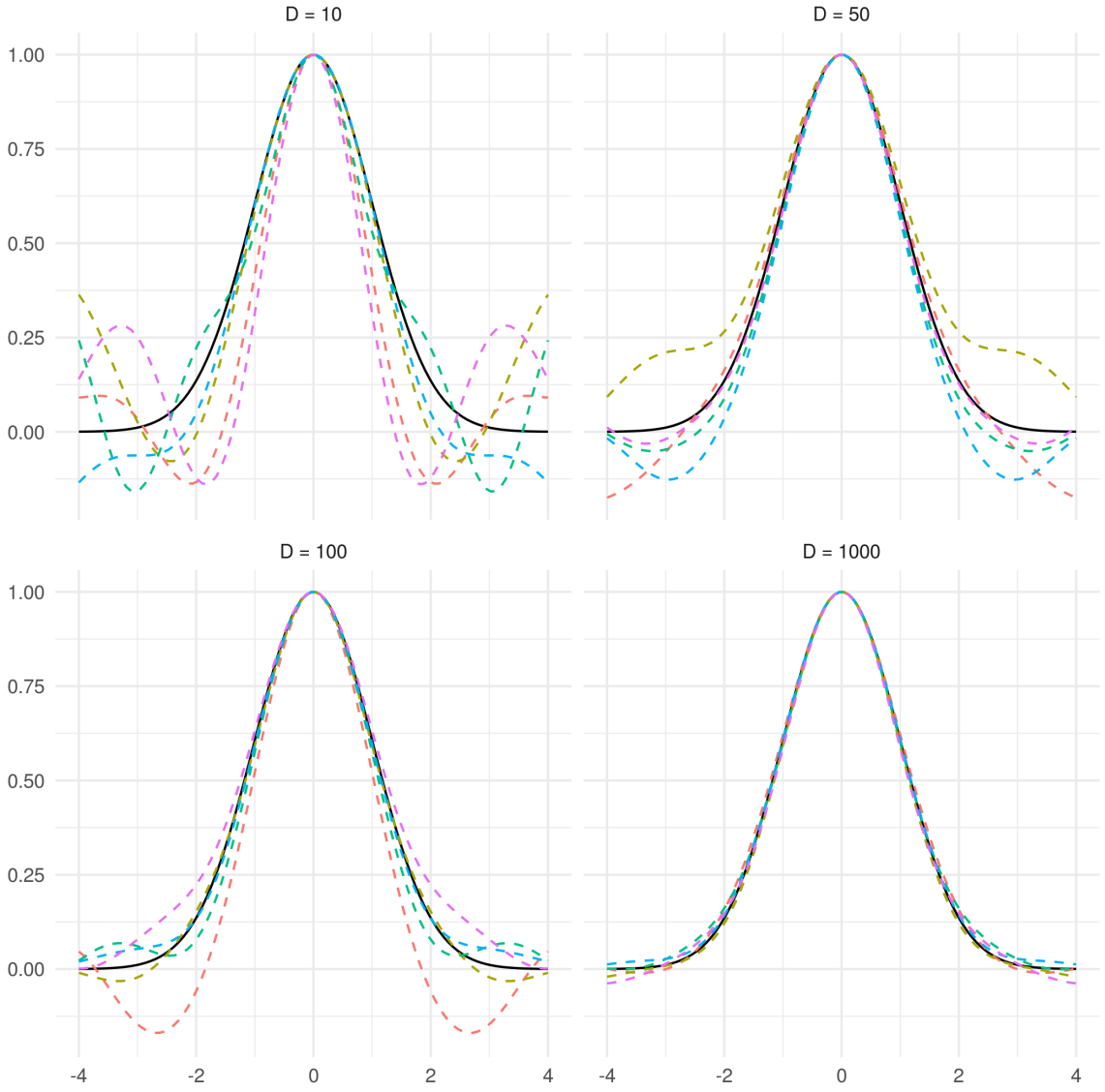


Figure 3.7: The embedding of $x = 0$ when using a radial basis function kernel ($\sigma = 1$) is shown by the black line. The dashed lines show approximations to this function when using different numbers of weights (different values of D), with the different colours representing different random samples of weights.

The posterior predictive distribution for a point with covariate values $\mathbf{x}_*$ is

$$\mathrm{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}_*^T A^{-1} X\mathbf{y},\, \mathbf{x}_*^T A^{-1}\mathbf{x}_*\right). \tag{3.4}$$

Gaussian process regression can be derived using the same regression model with the explanatory variables mapped into a RKHS. Suppose $\mathbf{x} \in \mathbb{R}^d$, let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with corresponding RKHS $\mathcal{H}$ and feature map ϕ. If $\mathcal{H}$ is of finite dimension, $n_{\mathcal{H}}$, the problem above can modified to be

$$y_i = \phi(\mathbf{x}_i)^T\mathbf{w} + \epsilon_i$$

now with $\mathbf{w} \in \mathbb{R}^{n_{\mathcal{H}}}$. The posterior predictive distribution can be written in a similar way to (3.4),

$$\mathrm{N}\left(\frac{1}{\sigma_n^2}\phi(\mathbf{x}_*)^T A^{-1}\Phi\mathbf{y},\, \phi(\mathbf{x}_*)^T A^{-1}\phi(\mathbf{x}_*)\right). \tag{3.5}$$

where $A = \sigma_n^{-2}\Phi\Phi^T + \sigma_p^2 I \in \mathbb{R}^{n_{\mathcal{H}}} \times \mathbb{R}^{n_{\mathcal{H}}}$ and $\Phi = (\phi(\mathbf{x}_1), ..., \phi(\mathbf{x}_n))$ is the matrix of observed data projected into feature space. For high dimensional spaces $\mathcal{H}$, inverting the matrix A becomes impractical. However, the posterior mean can be rewritten in terms of $n \times n$ matrices (where n is the number of observations),

$$\sigma_p^2\phi(\mathbf{x}_*)^T\Phi(\sigma_p^2 K + \sigma_n^2 I)^{-1}\mathbf{y} \tag{3.6}$$

as can the covariance,

$$\sigma_p^2[\phi(\mathbf{x}_*)^T\phi(\mathbf{x}_*) - \sigma_p^2\phi(\mathbf{x}_*)^T\Phi(\sigma_p^2 K + \sigma_n^2 I)^{-1}\Phi^T\phi(\mathbf{x}_*)] \tag{3.7}$$

where $K = \Phi^T\Phi$. These computations only involve inverting $n \times n$ matrices and therefore remain computationally feasible for high dimensional RKHS. Furthermore, all these calculations can be done using the kernel function directly without computing

the features themselves (this is the 'kernel trick' again), as

$$\begin{aligned} K_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j), \\ \phi(\mathbf{x}_*)^T \Phi &= (k(\mathbf{x}_1, \mathbf{x}_*), ..., k(\mathbf{x}_n, \mathbf{x}_*)), \text{and} \\ \phi(\mathbf{x}_*)^T \phi(\mathbf{x}_*) &= k(\mathbf{x}_*, \mathbf{x}_*). \end{aligned}$$

Therefore infinite dimensional RKHSs can also be used. The usual formulation for kernel ridge regression and Gaussian process regression is (3.6) and, for Gaussian process regression only, (3.7), although often σ_p is incorporated in the kernel itself.

These computations scale poorly with the number of observations, however, due to the matrix inversion necessary. If k is continuous and translation invariant, then this problem can be solved by applying the RFF theory described above. If we sample weights from the distribution Γ corresponding to k, defined as in Bochner's theorem, then we can use the map z, defined in (3.3), to find a low dimensional feature space which approximates $\mathcal{H}$. The dimension of this space depends on the number of weights sampled. We can then use the formulation in (3.5) with these low-rank features, i.e. replacing ϕ with z. A similar approach can be used for more complex models that involve Gaussian processes (such as those described in Chapter 6), with the Gaussian process being replaced by a linear combination of feature vectors wherever it occurs in the model.

3.3.3 Kernel mean embeddings of distributions

Having defined reproducing kernel Hilbert spaces, we can turn to the topic of embedding distributions, rather than just data points, into these spaces. For more details, see Smola et al. [127] for a short introduction and several examples of applications and Muandet et al. [128] for a more comprehensive review. Kernel mean embeddings provide a convenient framework for comparing distributions. By associating proba-

bility distributions with points in a Hilbert space, the distance between distributions can be defined by the distance between their embeddings. Moreover, these distances can be estimated from a sample by estimating these embeddings. This is the basis for a number of non-parametric two-sample tests [129, 130, 131]. The independence tests that we focus on in this chapter are instead derived by defining linear operators on RKHSs which encode cross-covariance information [45, 132, 133, 47], although these tests can also be interpreted as estimating the distance between embeddings of joint distributions and the product of marginal distributions [134]. Unlike other non-parametric methods for comparing distributions, kernel mean embedding techniques do not require kernel density estimation as an intermediate step (which is known to be more prone to the curse of dimensionality than RKHS methods, particularly when estimating conditional distributions [127, 135]).

The mean embedding of a distribution is defined as follows. Let X be a random variable with co-domain $\mathcal{X}$ and probability distribution P and let $\mathcal{H}$ be a RKHS of real-valued functions on $\mathcal{X}$ with kernel k. Then the *mean embedding* of this distribution, μ_P, is defined to be the element of $\mathcal{H}$ such that

mean embedding

$$\langle f, \mu_P \rangle_{\mathcal{H}} = \mathrm{E}_X[f(X)] \tag{3.8}$$

for all $f \in \mathcal{H}$. This element exists if $\mathrm{E}_X[k(X,X)^{1/2}] < \infty$. To see this, define the linear operator $T : \mathcal{H} \to \mathbb{R}$ as

$$T(f) = \mathrm{E}_X[f(X)]$$

and note that, by the Riesz representation theorem, an element μ_P satisfying (3.8)

exists if this operator is bounded. And indeed, we have

$$|T(f)| = |\mathrm{E}_X [f(X)]| \tag{3.9}$$

$$= |\mathrm{E}_X [\langle k(X, \cdot), f\rangle_{\mathcal{H}}]| \tag{3.10}$$

$$\leq \mathrm{E}_X [\|k(X, \cdot)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}] = \mathrm{E}_X [\|k(X, \cdot)\|_{\mathcal{H}}] \|f\|_{\mathcal{H}} = \mathrm{E}_X \left[k(X, X)^{1/2}\right] \|f\|_{\mathcal{H}} \tag{3.11}$$

which shows that this operator is bounded as required.

The mean embedding can also be written as

$$\mu_P = \int_{x \in \mathcal{X}} k(x, \cdot) \, \mathrm{d}P(x).$$

Evaluating the inner product with this element computes the expectation of a function, in a similar way to how the reproducing property allows function evaluation via inner products. Given a sample $\mathbf{x} = x_1, ..., x_n \overset{i.i.d.}{\sim} P$, we can also define the *empirical mean embedding*,

$$\mu[\mathbf{x}] := \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot)$$

which can be used as an estimate for μ_P and to compute sample means, since

$$\langle f, \mu[\mathbf{x}]\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{m} f(x_i)$$

which follows from the reproducing property. Figure 3.8 shows an example of these estimates for different samples drawn from a standard normal distribution compared to the true mean embedding when using a radial basis function kernel and Figure 3.9 shows the corresponding estimates when using a Laplace kernel.

A natural question is, when is the mapping of probability distributions to mean embeddings injective? This is relevant if we wish to characterise the distance between distributions by the distance between their embeddings, since we would like a distance

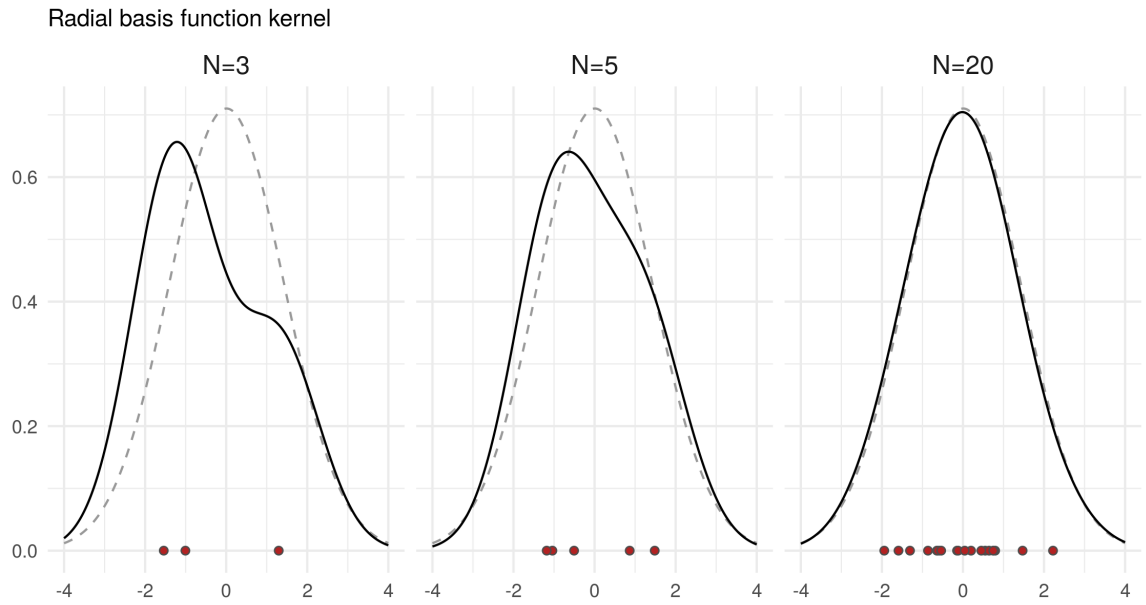


Figure 3.8: Estimates $\mu[\mathbf{x}]$ (black line) compared to the true function μ_P (grey dashed line) samples of different sizes, N, from a standard normal distribution when using a radial basis function kernel ($\sigma = 1$). The sample values are shown as red points.

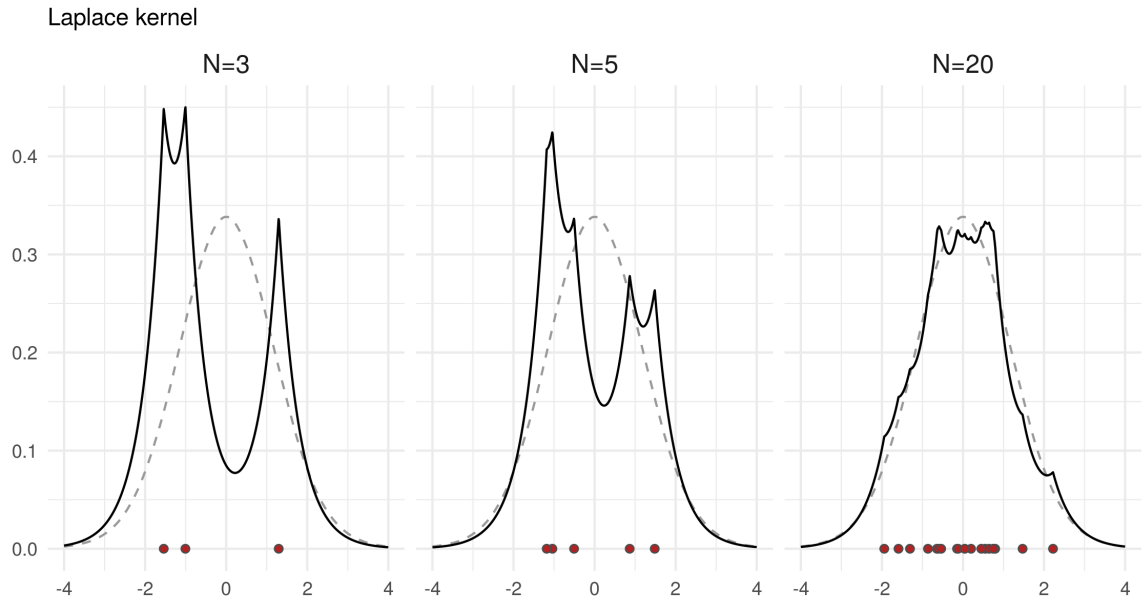


Figure 3.9: Estimates $\mu[\mathbf{x}]$ (black line) compared to the true function μ_P (grey dashed line) samples of different sizes, N, from a standard normal distribution when using a Laplace kernel ($\sigma = 1$). The sample values are shown as red points.

of zero between the mean embeddings of two distributions to imply that the two distributions are the same. Equivalently, we would like to know when the embedding μ_P preserves all the information about the distribution P. This is not the case for every kernel—under the linear kernel, for example, each distribution is mapped to its mean. Similarly, polynomial kernels preserve only a finite number of moments. On the other hand, the radial basis function kernel on a compact space does preserve the information of all moments and so the mean embedding map induced by this kernel is injective [136, 137]. The kernels for which this mapping is injective are known as *characteristic kernels*.

characteristic kernels

A strict subset of the characteristic kernels are the universal kernels, defined as follows.

Definition 8 (Universal kernel [136])**.** Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive definite kernel. Then the kernel k is *universal* if the corresponding RKHS is dense in $\mathcal{C}(\mathcal{X})$, the space of continuous bounded functions on $\mathcal{X}$.

universal kernel

Gretton et al. [137] prove the following theorem (a rephrasing of Theorem 5 [137]):

Theorem 3.3.4. Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a universal kernel with corresponding RKHS $\mathcal{H}$. Let P and Q be probability measures on this space with mean embeddings μ_P and μ_Q, respectively, with respect to the kernel k. Then $\|\mu_P - \mu_Q\|_{\mathcal{H}} = 0$ if and only if $P = Q$.

This theorem tells us exactly that every universal kernel on a compact space is characteristic. The definition of a universal kernel connects this aspect of kernel theory to the analysis of the well-studied Banach space, $\mathcal{C}(\mathcal{X})$. This allows related theoretical results to be employed to verify the characteristic nature of many commonly used kernels. One such result is the Stone-Weierstrass approximation theorem:

Theorem 3.3.5 (Stone-Weierstrass theorem [138])**.** Let $\mathcal{X}$ be a compact metric space and let A be a vector subspace of $\mathcal{C}(\mathcal{X})$. Then A is dense in $\mathcal{C}(\mathcal{X})$ if

1. for any $x \in \mathcal{X}$ there exists some $f \in A$ such that $f(x) \neq 0$, and
2. for any $x, y \in \mathcal{X}$ with $x \neq y$, there exists some $f \in A$ such that $f(x) \neq f(y)$.

These conditions can be checked to confirm that the radial basis function, infinite polynomial kernel and Laplace kernel are all universal, and therefore characteristic, on every compact subset of $\mathbb{R}^n$ [136]. Some other sufficient conditions for a kernel to be characteristic are discussed by Fukumizu et al. [139, 140, 141].

In the rest of this section, I will describe two kernel-based independence tests which I implemented to perform causal discovery, the Hilbert-Schmidt independence criterion and the randomized conditional independence test.

3.3.4 Hilbert-Schmidt independence criterion

The Hilbert-Schmidt independence criterion (HSIC) [45, 129] is one of the most commonly used kernel-based dependence measures and is used to test two variables for unconditional independence. The task is to determine whether variables X and Y are independent given joint observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. We can test this hypothesis by using an estimate of the HSIC as the test statistic in a permutation test, i.e. comparing the estimate computed using the observed data to an approximation of the distribution under the null hypothesis of independence generated by repeatedly permuting the observations of one variable and recomputing this statistic. Alternatively, the null distribution can be approximated using asymptotic results. A high value of the HSIC represents stronger dependence, so the null hypothesis will be rejected if the estimate of the HSIC is above a certain value, determined by the null distribution and the desired significance level.

This independence criterion is derived from a covariance operator that encodes information about the dependence of two variables. To understand this operator we first need to define a Hilbert-Schmidt operator.

Definition 9. Let $\mathcal{F}, \mathcal{G}$ be separable Hilbert spaces (that is, each has an orthonormal basis) and let $A : \mathcal{F} \to \mathcal{G}$ be a linear operator. The *Hilbert-Schmidt norm* of this operator is

$$\|A\|_{\mathrm{HS}}^2 := \sum_{i \in I} \sum_{j \in J} \langle v_i, A u_j \rangle_{\mathcal{G}}$$

where $\{v_i : i \in I\}$ and $\{u_j : j \in J\}$ are orthonormal bases for $\mathcal{G}$ and $\mathcal{F}$, respectively. If $\|A\|_{\mathrm{HS}} < \infty$ then A is a *Hilbert-Schmidt operator*. An inner product can be defined on the set of Hilbert-Schmidt operators from $\mathcal{F}$ to $\mathcal{G}$, $\mathrm{HS}(\mathcal{F}, \mathcal{G})$, by

Hilbert-Schmidt operator

$$\langle A, B \rangle_{\mathrm{HS}} = \sum_{i,j} \langle v_i, A u_j \rangle_{\mathcal{G}} \langle v_i, B u_j \rangle_{\mathcal{G}}$$

and moreover with this inner product $\mathrm{HS}(\mathcal{F}, \mathcal{G})$ is a separable Hilbert space.

Now let X and Y be random variables taking on values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{F}$ be a separable RKHS of functions on $\mathcal{X}$ with kernel k and $\mathcal{G}$ be a separable RKHS of functions on $\mathcal{Y}$ with kernel l. Suppose that $\mathrm{E}_X[k(X, X)] < \infty$ and $\mathrm{E}_Y[l(Y, Y)] < \infty$. Then define the *cross-covariance operator* [142], $\Sigma_{YX} : \mathcal{F} \to \mathcal{G}$, to be the operator such that

cross-covariance operator

$$\begin{aligned} \langle g, \Sigma_{YX} f \rangle_{\mathcal{G}} &= \mathrm{Cov}[f(X), g(Y)] \\ &= \mathrm{E}_{XY}[f(X) g(Y)] - \mathrm{E}_X[f(X)] \mathrm{E}_Y[g(Y)]. \end{aligned}$$

for any $f \in \mathcal{F}, g \in \mathcal{G}$. To show that such an operator exists, it suffices to show that for any $f \in \mathcal{F}$ there exists an element $g_f \in \mathcal{G}$ such that

$$\langle g, g_f \rangle_{\mathcal{G}} = \mathrm{Cov}[f(X), g(Y)]. \tag{3.12}$$

for all $g \in \mathcal{G}$. Indeed, it can be shown that (see Appendix A in [143] for details)

$$\begin{aligned}|\mathrm{Cov}[f(X), g(Y)]| &= |\mathrm{E}_{XY}[f(X)g(Y)] - \mathrm{E}_X[f(X)]\mathrm{E}_Y[g(Y)]| \\ &\leq 2\mathrm{E}_X[k(X,X)]^{1/2}\mathrm{E}_Y[l(Y,Y)]^{1/2}\|f\|_{\mathcal{F}}\|g\|_{\mathcal{G}}\end{aligned}$$

and therefore the operator $T_f : g \mapsto \mathrm{Cov}[f(X), g(Y)]$ is a bounded linear operator. It follows that an element $g_f \in \mathcal{G}$ satisfying (3.12) exists by the Riesz representation theorem. Finally, we can then define Σ_{YX} to be the operator that sends f to g_f.

The Hilbert-Schmidt norm of this operator is

$$\|\Sigma_{YX}\|^2_{\mathrm{HS}} = \sum_{i\in I}\sum_{j\in J}\langle v_i, Au_j\rangle_{\mathcal{G}} \tag{3.13}$$

$$= \sum_{i\in I}\sum_{j\in J}\mathrm{Cov}[u_j(X), v_i(Y)] \tag{3.14}$$

for bases $\{v_i : i \in I\}$ and $\{v_i : i \in I\}$ of $\mathcal{G}$ and $\mathcal{F}$. The cross-covariance operator can be thought of as similar to the cross-covariance matrix of two vectors of random variables $\mathbf{X}$ and $\mathbf{Y}$, $K_{\mathbf{XY}}$, which is defined as

$$(K_{\mathbf{XY}})_{ij} = \mathrm{Cov}[\mathbf{X}_i, \mathbf{Y}_j]$$

where $(K_{\mathbf{XY}})_{ij}$ is the (i,j)-entry of the matrix. This matrix can be used to compute the covariance between linear transformations of $\mathbf{X}$ and $\mathbf{Y}$ (as the cross-covariance operator does for functions in the RKHSs),

$$\mathrm{Cov}[A\mathbf{X}, B\mathbf{Y}] = AK_{\mathbf{XY}}B^T.$$

Just as the mean embedding of a distribution allows the expectation of any function to be computed using the inner product, the cross-covariance operator computes the cross-covariance between functions in $\mathcal{F}$ and $\mathcal{G}$ via an inner product. We may expect

that for sufficiently rich classes of functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$ this operator will capture all the dependence between the two variables, again in an analogous way to the mean embedding preserving all information about a distribution. This is formalised in the following theorem by Gretton et al. [45, 46].

Theorem 3.3.6. Let X, Y be random variables with compact co-domains $\mathcal{X}, \mathcal{Y}$ and let $\mathcal{F}, \mathcal{G}$ be separable RKHSs on $\mathcal{X}, \mathcal{Y}$ with universal kernels. Then $\|\Sigma_{YX}\|_{\text{HS}} = 0$ if and only if X and Y are independent.

This result provides the motivation for their independence criterion [45]:

Definition 10. Let X, Y be random variables with joint measure P_{XY} and let $\mathcal{F}, \mathcal{G}$ be corresponding separable RKHSs. The *Hilbert-Schmidt Independence Criterion (HSIC)* is defined as the squared norm of the cross-covariance operator, HSIC

$$\text{HSIC}(P_{XY}, \mathcal{F}, \mathcal{G}) := \|\Sigma_{YX}\|^2_{\text{HS}}.$$

Under the conditions of Theorem 3.3.6, testing the hypothesis that two variables are independent,

$$H_0 : X \perp\!\!\!\perp Y$$
$$H_1 : X \not\perp\!\!\!\perp Y$$

can be equivalently formulated using this criterion,

$$H_0 : \text{HSIC}(P_{XY}, \mathcal{F}, \mathcal{G}) = 0$$
$$H_1 : \text{HSIC}(P_{XY}, \mathcal{F}, \mathcal{G}) > 0.$$

Gretton et al. [45] show that this criterion can be written in terms of the kernels k

and l for RKHSs $\mathcal{F}$ and $\mathcal{G}$,

$$\begin{aligned}\mathrm{HSIC}(P_{XY}, \mathcal{F}, \mathcal{G}) =& \mathrm{E}_{X,X',Y,Y'}[k(X,X')l(Y,Y')] + \mathrm{E}_{X,X'}[k(X,X')]\mathrm{E}_{Y,Y'}[l(Y,Y')] \\ &- 2\mathrm{E}_{X,Y}[\mathrm{E}_{X'}[k(X,X')]\mathrm{E}_{Y'}[l(Y,Y')]]\end{aligned}$$

which motivates the following estimator. Let $\mathcal{Z} = \{(x_1, y_1), ..., (x_n, y_n)\}$ be independent joint observations of X, Y, then an estimator of the HSIC is given by

$$\widehat{\mathrm{HSIC}}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) := (n-1)^{-2} \operatorname{trace}(KHLH)$$

where H, K, L are n-by-n matrices with $K_{ij} := k(x_i, x_j)$, $L_{ij} := l(y_i, y_j)$ and $H_{ij} := \delta_{ij} - n^{-1}$. K and L are the Gram matrices of kernels k and l with respect to the observations of X and Y, respectively, and L is a centering matrix. This is a biased estimator but can be computed more efficiently (in $\mathcal{O}(n^2)$ time) than the unbiased equivalent by taking advantage of the structure of these matrices. It is shown in [45] that this estimator converges to the HSIC at a speed $1/\sqrt{n}$.

The distribution of this test statistic under the null hypothesis of independence can be obtained by repeatedly permuting the observations of one variable (resulting in joint observations $(x_{\sigma(1)}, y_1), ..., (x_{\sigma(n)}, y_n)$ where $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ is a bijection) and recomputing this statistic. Alternatively, Gretton et al. [137] derive the asymptotic distribution of this test statistic.

The $\mathcal{O}(n^2)$ complexity of computing the HSIC can limit the utility of this criterion when working with large data sets. A number of approximations have been proposed to reduce this complexity [133, 122]. If the kernels k and l are both continuous and translation-invariant, one approach is to approximate $\mathcal{F}$ and $\mathcal{G}$ (which are high, often infinite, dimensional) using RFFs (as introduced in Section 3.3.2). Let Γ_k and Γ_l be the probability measures associated with kernels k and l as in Bochner's theorem. In other words, k is the Fourier transform of Γ_k and l is the Fourier transform of Γ_l. Let

$w_1, ..., w_D \overset{i.i.d.}{\sim} \Gamma_k$ and $v_1, ..., v_{D'} \overset{i.i.d.}{\sim} \Gamma_l$ be randomly sampled weights and define the feature maps $z : \mathcal{X} \to \mathbb{R}^{2D}$ and $z' : \mathcal{Y} \to \mathbb{R}^{2D'}$ by

$$z(x) = \frac{1}{\sqrt{D}} \left(\cos(w_1^T x), \sin(w_1^T x), \ldots, \cos(w_D^T x), \sin(w_D^T x)\right) \tag{3.15}$$

$$z'(y) = \frac{1}{\sqrt{D'}} \left(\cos(v_1^T y), \sin(v_1^T y), \ldots, \cos(v_{D'}^T y), \sin(v_{D'}^T y)\right). \tag{3.16}$$

These maps send elements of $\mathcal{X}$ and $\mathcal{Y}$ to points in $\mathbb{R}^{2D}$ and $\mathbb{R}^{2D'}$, respectively. Let $\mathcal{F}'$ and $\mathcal{G}'$ be the spaces of functions $\mathcal{X} \to \mathbb{R}$ and $\mathcal{Y} \to \mathbb{R}$ corresponding to $\mathbb{R}^{2D}$ and $\mathbb{R}^{2D'}$ (see Section 3.3.2 for details on this equivalence). Since these feature spaces are of finite dimension, we can derive an approximation for the HSIC by working directly with (3.13), the expression for the norm of the cross-covariance operator in terms of basis elements. Recall from Section 3.3.2 that the function on $\mathcal{X}$ represented by a point $a \in \mathbb{R}^{2D}$ is $f_a : x \mapsto z(x)^T a$. Therefore the basis functions of $\mathcal{F}'$ corresponding to the standard basis vectors $e_1, ..., e_{2D} \in \mathbb{R}^{2D}$ are

$$\begin{aligned}
f_{e_1}(x) &= z(x)^T e_1 = D^{-\frac{1}{2}} \cos(w_1^T x) \\
f_{e_2}(x) &= D^{-\frac{1}{2}} \sin(w_1^T x) \\
&\ldots \\
f_{e_{2D-1}}(x) &= D^{-\frac{1}{2}} \cos(w_D^T x) \\
f_{e_{2D}}(x) &= D^{-\frac{1}{2}} \sin(w_D^T x).
\end{aligned}$$

For convenience we write these functions as simply $f_1, ..., f_{2D}$ and let the basis functions of $\mathcal{G}'$ be $g_1, ..., g_{2D'}$. Using these bases in (3.13) gives an approximation to the HSIC,

$$\text{HSIC}_{\text{RFF}} := \sum_{i=1}^{2D'} \sum_{j=1}^{2D} \text{Cov}[f_i(X), g_j(Y)]$$

and given observations, $\mathcal{Z}$, as above we have a plug-in estimator

$$\widehat{\text{HSIC}}_{\text{RFF}}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) := \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{2D'} \sum_{j=1}^{2D} (f_i(x_k) - \bar{f}_i)(g_j(y_k) - \bar{g}_j)$$

where $\bar{f}_i = \frac{1}{n} \sum_{r=1}^{n} f_i(x_r)$ and $\bar{g}_j = \frac{1}{n} \sum_{r=1}^{n} g_j(y_r)$. More details on these approximations can be found in [122] but importantly we can note that the time taken to compute this estimator scales linearly with the number of observations.

3.3.5 Randomized Conditional Independence Test

Constraint-based causal discovery algorithms, including the PC algorithm, require both unconditional and conditional independence tests. One method for performing conditional independence tests is *regression with subsequent independence test (RESIT)* [144, 101], a non-linear extension to testing for independence of partial correlations. An attractive feature of this method is that it uses non-linear regression to reduce the problem to one of unconditional independence testing (where we can apply the HSIC or other unconditional tests) and so it is conceptually simpler than many other kernel-based conditional independence tests. To test variables X, Y for independence given some set of conditioning variables Z, both X and Y are first regressed with respect to the conditioning variables to obtain residuals,

RESIT

$$r_{XZ} = X - \hat{E}[X|Z]$$
$$r_{YZ} = Y - \hat{E}[Y|Z].$$

These residuals are then tested for unconditional independence, with conditional independence rejected if the test between residuals rejects independence. I implemented this method using Gaussian process regression and the HSIC and used this extensively in early experiments with causal discovery. The assumption underlying RESIT

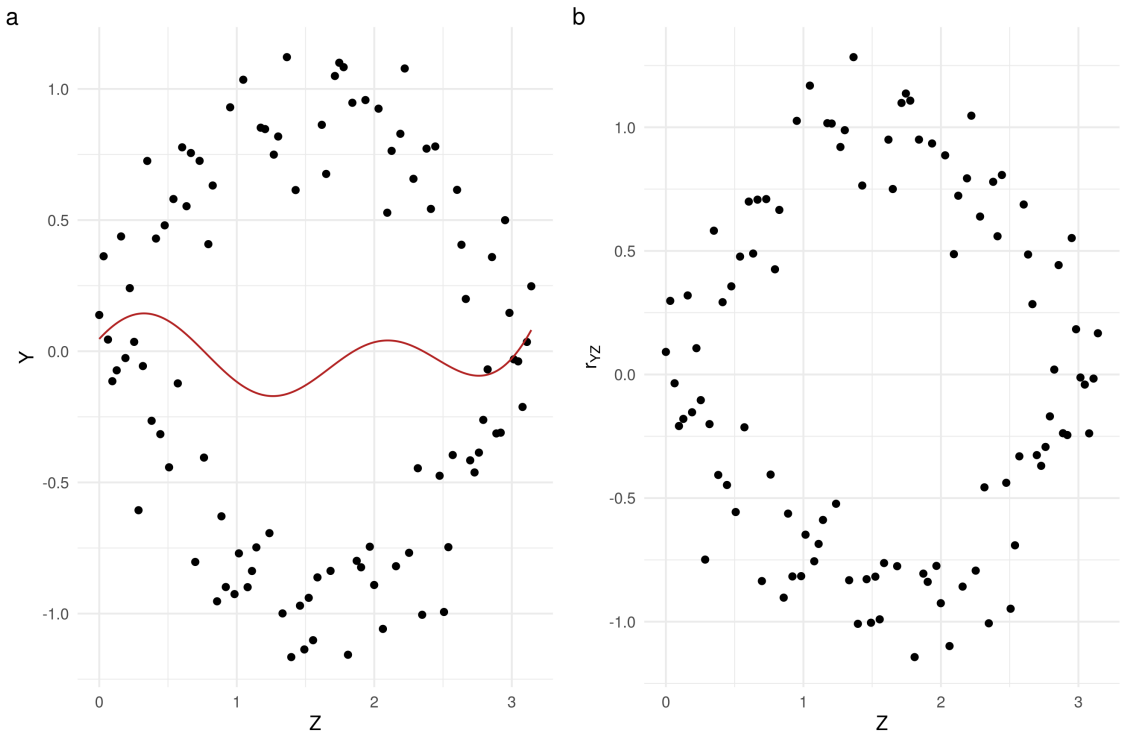


Figure 3.10: (a) Joint observations of variables Z and Y (black points) and Gaussian process regression of Y on Z (red line). (b) Residuals r_{YZ} from the Gaussian process regression of Y against Z, plotted against Z.

is that there is indeed a functional relationship between X and Z (and similarly Y and Z) that can be identified using non-linear regression. However, this is not necessarily the case. For example, consider the relationship illustrated in Figure 3.10(a) between variables Y and Z. It is trivially true that $Y \perp\!\!\!\perp Z|Z$ but the relationship between Y and Z cannot be described using a function and RESIT fails—the residuals r_{YZ} still have a dependence on Z, as shown in Figure 3.10(b). Other methods have been proposed for reducing conditional independence tests to unconditional ones based on discretising Z and using a series of unconditional tests [145, 146]. While these tests do not depend on functional relationships like RESIT, they suffer from loss of information due to discretisation and the curse of dimensionality as the size of the conditioning set grows. This is problematic for constraint-based causal discovery algorithms, in which tests conditioning on multiple variables are likely to be required.

In the final implementation of the causal discovery procedure, I therefore focused on kernel-based tests of conditional independence that addressed the conditional independence problem directly within the RKHS framework. The test I implemented was the Randomized Conditional Independence Test (RCIT) [47] which is a RFF-based approximation of another kernel-based test, the Kernel-based Conditional Independence Test (KCIT) [132]. These tests have been shown empirically to work well, even in high dimensional settings [132, 47].

The RCIT and the KCIT are both based on a conditional cross-covariance operator, which can be thought of as an extension of the cross-covariance operator that underlies the HSIC. Much of this theory was developed by Fukumizu et al. [143]. I summarise the construction of this operator here. Let X, Y, Z be random variables with corresponding RKHSs $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ and kernels k_1, k_2, k_3. Define the covariance operator $\Sigma_{XX} : \mathcal{H}_1 \to \mathcal{H}_1$ in a similar way to the cross-covariance operator, that is such that

$$\langle f_1, \Sigma_{XX} f_2 \rangle_{\mathcal{H}_1} = \mathrm{E}_X[f_1(X) f_2(X)] - \mathrm{E}_X[f_1(X)]\mathrm{E}_X[f_2(X)]$$

for any $f_1, f_2 \in \mathcal{H}_1$. Then we have the following important result (Theorem 2 [143]), which allows us to begin to connect covariance operators and conditional distributions.

Theorem 3.3.7. Let $X, Y, \mathcal{H}_1, \mathcal{H}_2, k_1, k_2$ be as above. Suppose that $\mathrm{E}_X[k_1(X, X)] < \infty$ and $\mathrm{E}_Y[k_2(Y, Y)] < \infty$, and that $\mathrm{E}_{Y|X}[g(Y)] \in \mathcal{H}_1$ for all $g \in \mathcal{H}_2$. Then

$$\Sigma_{XX} \mathrm{E}_{Y|X}[g(Y)] = \Sigma_{XY} g$$

for all $g \in \mathcal{H}_2$, where Σ_{XY} is the cross-covariance operator.

To show that this is true, it suffices to show that

$$\langle f, \Sigma_{XX} \mathrm{E}_{Y|X}[g(Y)] \rangle_{\mathcal{H}_1} = \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_1}$$

for all $f \in \mathcal{H}_1$. By definition, the right hand side is equal to $\mathrm{Cov}[f(X), g(Y)]$, and the left hand side is

$$\begin{aligned}
&\langle f, \Sigma_{XX} \mathrm{E}_{Y|X}[g(Y)] \rangle_{\mathcal{H}_1} \\
&\qquad = \mathrm{E}_X[f(X)\mathrm{E}_{Y|X}[g(Y)]] - \mathrm{E}_X[f(X)]\mathrm{E}_X[\mathrm{E}_{Y|X}[g(Y)]] \\
&\qquad = \mathrm{E}_{XY}[f(X)g(Y)] - \mathrm{E}_X[f(X)]\mathrm{E}_Y[g(Y)] \\
&\qquad = \mathrm{Cov}[f(X), g(Y)]
\end{aligned}$$

as required. Additionally, it can be shown that a sufficient condition for $\mathrm{E}_{Y|X}[g(Y)|X]$ to be an element of $\mathcal{H}_1$ for all $g \in \mathcal{H}_2$ is that the ranges of X and Y are bounded [147, 143]. It also follows from Theorem 3.3.7 that

$$\Sigma_{XX}^{-1} \Sigma_{XY} g = \mathrm{E}_{Y|X}[g(Y)]$$

when Σ_{XX} is invertible.

We are now ready to define the conditional cross-covariance operator that the independence tests of interest are based on. With the set up as above, the *conditional cross-covariance operator of* (X, Y) *given* Z is

conditional cross-covariance operator

$$\Sigma_{YX|Z} := \Sigma_{YX} - \Sigma_{YZ} \tilde{\Sigma}_{ZZ}^{-1} \Sigma_{ZX}$$

where $\tilde{\Sigma}_{ZZ}^{-1}$ is the inverse of Σ_{ZZ} or the right inverse if it is not invertible. Now suppose that $\mathrm{E}_X[k_1(X, X)]$, $\mathrm{E}_Y[k_2(Y, Y)]$ and $\mathrm{E}_Z[k_3(Z, Z)]$ are finite and that $\mathrm{E}_{X|Z}[f(X)|Z]$ and $\mathrm{E}_{Y|Z}[g(Y)|Z]$ are in $\mathcal{H}_3$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Then this operator has the property (Proposition 5 [143]) that

$$\begin{aligned}
\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_2} &= \mathrm{E}_{XY}[f(X)g(Y)] - \mathrm{E}_Z[\mathrm{E}_{X|Z}[f(X)]\mathrm{E}_{Y|Z}[g(Y)]] && (3.17) \\
&= \mathrm{E}_Z[\mathrm{Cov}_{XY|Z}(f(X), g(Y))]. && (3.18)
\end{aligned}$$

This follows from Theorem 3.3.7. We can see the parallel between this property and the definition of the (unconditional) cross-covariance operator. In the case of the cross-covariance operator, we saw that if $\mathcal{X}$ and $\mathcal{Y}$ are compact and the kernels on these spaces are universal then $\|\Sigma_{YX}\|_{\text{HS}} = 0$ if and only if X and Y are independent. However, it is *not* the case that, under similar conditions, $\|\Sigma_{YX|Z}\|_{\text{HS}} = 0$ implies that X and Y are independent given Z. For example, consider a joint distribution in which X and Y are independent given Z except at a set of values of Z with measure zero. In this case, $\|\Sigma_{YX|Z}\|_{\text{HS}} = 0$ but X, Y are not independent given Z. However, the norm of this operator being zero does imply conditional independence if either X or Y are extended to include Z, as follows [143, 139].

Theorem 3.3.8. Let X, Y, Z and associated RKHSs be as above. Let $\ddot{X} = (X, Z)$ and $k_{\ddot{\mathcal{X}}} = k_1 k_3$ and let $\mathcal{H}_{\ddot{\mathcal{X}}}$ be the corresponding RKHS. Additionally, assume that $k_{\ddot{\mathcal{X}}} k_2$ is characteristic on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, and that $\mathcal{H}_3 + \mathbb{R}$ is dense in $L^2_{\mathcal{Z}}$. Then

$$\Sigma_{\ddot{X}Y|Z} = 0 \iff X \perp\!\!\!\perp Y|Z.$$

With this result, if the above conditions are satisfied the hypothesis test of conditional independence

$$H_0 : X \perp\!\!\!\perp Y|Z$$
$$H_1 : X \not\perp\!\!\!\perp Y|Z$$

can be equivalently formulated,

$$H_0 : \|\Sigma_{\ddot{X}Y|Z}\|^2_{\text{HS}} = 0$$
$$H_1 : \|\Sigma_{\ddot{X}Y|Z}\|^2_{\text{HS}} > 0.$$

in the same way that the HSIC (the norm of the operator Σ_{YX}) was used for uncon-

ditional independence tests. To derive the estimate of this norm that is used in the RCIT, first note that by (3.17) the Hilbert-Schmidt norm of this operator is

$$\|\Sigma_{\ddot{X}Y|Z}\|^2_{\text{HS}} = \sum_{i,j} \langle v_i, \Sigma_{\ddot{X}Y|Z} u_j \rangle_{\mathcal{H}_{\ddot{\mathcal{X}}}} \tag{3.19}$$

$$= \sum_{i,j} \text{E}_Z \left[\text{Cov}_{\ddot{X}Y|Z} \left[u_j(\ddot{X}), v_i(Y) \right] \right] \tag{3.20}$$

$$= \sum_{i,j} \text{E}_Z \left[\text{E}_{\ddot{X}Y|Z} \left[\left(u_j(\ddot{X}) - \text{E}_{\ddot{X}|Z}[u_j(\ddot{X})] \right) \left(v_i(Y) - \text{E}_{Y|Z}[v_i(Y)] \right) \right] \right] \tag{3.21}$$

$$= \sum_{i,j} \text{E}_{\ddot{X}YZ} \left[\left(u_j(\ddot{X}) - \text{E}_{\ddot{X}|Z}[u_j(\ddot{X})] \right) \left(v_i(Y) - \text{E}_{Y|Z}[v_i(Y)] \right) \right] \tag{3.22}$$

where $\{v_i\}$ is a basis for $\mathcal{H}_{\ddot{\mathcal{X}}}$ and $\{u_j\}$ is a basis for $\mathcal{H}_{\mathcal{Y}}$. In general we cannot estimate this quantity by replacing the expectation with observed values because the size of these bases could be high or infinite. However, we can use the theory of random Fourier features to obtain bases for finite dimensional spaces that approximate $\mathcal{H}_{\ddot{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{Y}}$. This is very similar to the approximation of the HSIC discussed at the end of Section 3.3.4.

Let Γ_1 and Γ_2 be probability measures such that $k_{\ddot{\mathcal{X}}}$ is the Fourier transform of Γ_1 and k_2 is the Fourier transform of Γ_2. Let $w_1, ..., w_D \overset{i.i.d.}{\sim} \Gamma_1$ and $v_1, ..., v_{D'} \overset{i.i.d.}{\sim} \Gamma_2$ be randomly sampled weights which define feature maps z_1 and z_2 and feature spaces $\mathcal{H}'_{\ddot{\mathcal{X}}}$ and $\mathcal{H}'_{\mathcal{Y}}$. These spaces approximate $\mathcal{H}_{\ddot{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{Y}}$ respectively. Finally let $f_1, ..., f_{2D}$ and $g_1, ..., g_{2D'}$ be the finite bases associated with $\mathcal{H}'_{\ddot{\mathcal{X}}}$ and $\mathcal{H}'_{\mathcal{Y}}$. These basis functions are defined exactly as in the end of Section 3.3.4. Using these basis functions in (3.22), we have

$$\|\Sigma_{\ddot{X}Y|Z}\|^2_{\text{HS}} \approx \sum_{i=1}^{2D'} \sum_{j=1}^{2D} \text{E}_{\ddot{X}YZ} \left[\left(f_j(\ddot{X}) - \text{E}_{\ddot{X}|Z}[f_j(\ddot{X})] \right) \left(g_i(Y) - \text{E}_{Y|Z}[g_i(Y)] \right) \right]$$

and an empirical estimate, S, is given by

$$S := \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{2D'}\sum_{j=1}^{2D}\left[\left(f_j(\ddot{x}_k) - \widehat{\mathrm{E}}_{\ddot{X}|Z}[f_j(\ddot{X})|Z = z_k]\right)\left(g_i(y_k) - \widehat{\mathrm{E}}_{Y|Z}[g_i(Y)|Z = z_k]\right)\right].$$

Strobl et al. [47] propose using kernel ridge regression with RFFs (see the end of Section 3.3.2) to fit the regression functions $\widehat{\mathrm{E}}_{\ddot{X}|Z}[f_i(\ddot{X})]$ and $\widehat{\mathrm{E}}_{Y|Z}[g_i(Y)]$. This ensures the complexity of calculating S is $\mathcal{O}(n)$, where n is the number of data points. This is the test statistic used in the RCIT.

To summarise, let X, Y, Z be random variables with associated kernels k_1, k_2, k_3 that satisfy the assumptions in Theorem 3.3.7. Then the test statistic, S, can be calculated as follows.

1. Join observations X and Z to produce $\ddot{X} = (X, Z)$ with kernel $k_{\ddot{\mathcal{X}}} := k_1 k_3$.

2. Sample random weights from the distributions corresponding to $k_{\ddot{\mathcal{X}}}$ and k_2 to get the corresponding bases $f_1, ...f_{2D}$ and $g_1, ...g_{2D'}$, where D and D' are fixed integers that are pre-specified.

3. Sample random weights from the distribution corresponding to k_3 and use these for RFF kernel ridge regression to fit functions $\widehat{\mathrm{E}}_{\ddot{X}|Z}[f_j(\ddot{X})]$ and $\widehat{\mathrm{E}}_{Y|Z}[g_i(Y)]$ for $j = 1, ..., 2D, i = 1, ..., 2D'$ using the observed data. Define the residual functions:

$$R_{\mathcal{Z}} f_j(\ddot{x}, z) := f_j(\ddot{x}) - \widehat{\mathrm{E}}_{\ddot{X}|Z}[f_j(\ddot{X})|Z = z]$$
$$R_{\mathcal{Z}} g_i(y, z) := g_i(y) - \widehat{\mathrm{E}}_{Y|Z}[g_i(Y)|Z = z]$$

4. Compute S using (3.3.5),

$$S = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{2D'}\sum_{j=1}^{2D} R_{\mathcal{Z}} f_j(\ddot{x}_k, z_k) R_{\mathcal{Z}} g_i(y_k, z_k).$$

The asymptotic distribution of this test statistic under the null hypothesis is derived by Strobl et al. [47] but, as with the HSIC, this statistic can also be used in a permutation test which is the approach I used.

3.3.6 Independence tests with aggregated data

In Chapter 2 we saw that we cannot always treat malariometric information as being associated with a specific point in space. Incidence data, in particular, is often collected at health facilities and does not include information about where the individual lives (or where the case occurred). Therefore even when case counts are available at each facility, these are aggregations over their (often unknown) catchment areas. It is also common for incidence data to be in the form of case counts over administrative regions. In contrast, many of the other variables of interest, such as environmental or socioeconomic covariates, are available at a higher resolution, typically on a grid of approximately 1km-by-1km or 5km-by-5km pixels.

This presents the problem for independence testing (and the causal discovery algorithms that depend on these tests) of how to compare variables that are measured on different spatial scales. As mentioned when discussing disaggregation regression models in Section 2.4, relationships between variables observed on one spatial scale may not hold on another (known as the ecological fallacy [88]). In the context of causal discovery, we wish to understand the mechanisms that occur on the individual level and so should aim to infer relationships between variables as close to this individual level as possible. Therefore aggregating all variables to the resolution of the incidence observations (or more generally to the lowest resolution of all variables considered) may not be the most efficient use of available information.

Instead, we would like to somehow associate multiple observations from the higher resolution variables with each observation of the aggregated variable. For example, if we were interested in the relationship between incidence, observed over administra-

tive regions, and rainfall, we would like to associate the incidence in a region with the values of rainfall in all pixels in that region. To do this, I propose below a simple extension to the theory of kernel-based independence testing that closely resembles distribution regression, which addresses the similar problem of performing regression with aggregated response data and finer scale explanatory variables [148, 149, 150]. This method allows us to perform independence and conditional independence testing using variables that are observed at different levels of aggregation. There are two features that make this method particularly attractive. The first is that it is not specific to a particular independence or conditional independence test, but instead modifies the kernel function and therefore can be used with any kernel-based test. Secondly, where random Fourier features are used in the original tests, similar approximate feature mappings can easily be computed, thereby preserving the computational efficiency of these tests.

For simplicity, I start by describing this method as applied to unconditional independence testing. Let X and Y be variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ that we wish to test for unconditional independence. Suppose that we have aggregated observations of Y, $y_1, ..., y_n$, and corresponding groups of observations of X, $\{x_1^i\}_{i=1}^{m_1}, ..., \{x_n^i\}_{i=1}^{m_n}$. Let k_X be a kernel on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}_X$ and feature map ϕ_X. We can use ϕ_X to define a mapping ψ of each group of observations of X to a single point in this space,

$$\psi : \{x_j^i\}_{i=1}^{m_j} \longmapsto \hat{\mu}_j := \frac{1}{m_j} \sum_{i=1}^{m_j} \phi_X(x_j^i). \tag{3.23}$$

As the notation suggests, $\hat{\mu}_j$ could be thought of as an empirical estimate of the mean embedding of some distribution associated with the variable X in the aggregation area. Unlike many other transformations of the grouped observations of X (such as the mean), with a suitable choice of kernel this embedding preserves all information from the individual observations. The problem is then re-framed as testing for independence between Y and this aggregate random variable, which we denote X_μ,

using joint observations $(y_1, \hat{\mu}_1), ..., (y_n, \hat{\mu}_n)$. Any kernel-based independence test can be applied to this problem. A number of kernels can be used for X_μ (see Szabó et al. [149]) but as the values of X have already been embedded in a Hilbert space, I use the simplest choice, the linear kernel

$$k_\mu(\hat{\mu}_i, \hat{\mu}_j) := \langle \hat{\mu}_i, \hat{\mu}_j \rangle_{\mathcal{H}_X}$$

which can be written in terms of the kernel k_X,

$$k_\mu(\hat{\mu}_i, \hat{\mu}_j) = \frac{1}{m_i m_j} \sum_{r=1}^{m_i} \sum_{s=1}^{m_j} k_X(x_i^r, x_j^s).$$

With these new observations and kernel function, a kernel-based independence test (such as the HSIC) can be applied in the standard way.

Using a characteristic kernel ensures that the map in (3.23) is injective (up to reordering of observations $\{x_j^i\}_{i=1}^{m_j}$). To see this, first we can define a map between groups of observations of X and linear combinations of Dirac measures as follows,

$$\{x_j^i\}_{i=1}^{m_j} \longmapsto \frac{1}{m_j} \sum_{i=1}^{m_j} \delta_{x_j^i}$$

where for any $x \in \mathcal{X}$ the measure δ_x is defined as

$$\delta_x(A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

for any A in the σ-algebra of subsets of $\mathcal{X}$. This map is a bijection (again up to reordering of observations). Then note that the map

$$\frac{1}{m_j} \sum_{i=1}^{m_j} \delta_{x_j^i} \longmapsto \hat{\mu}_j$$

is the mean embedding of these probability measures with respect to kernel k. Therefore, if the kernel is characteristic this map is injective. The composition of these two maps,

$$\{x_j^i\}_{i=1}^{m_j} \longmapsto \frac{1}{m_j}\sum_{i=1}^{m_j}\delta_{x_j^i} \longmapsto \hat{\mu}_j$$

is equivalent to (3.23) and must also be injective.

In our applications, Y could be the incidence rate and X an environmental variable. The observations y_j might then be the incidence rate in administrative region j and $\{x_j^i\}_{i=j}^{m_j}$ the values of the environmental variable in each pixel in this region. Often we would like to weight each pixel observation by the population in that pixel. We can effectively think of this as having observations of the environmental variable for each individual. In this case, let our pixel-level observations for region j be $\{(x_j^i, w_j^i)\}_{i=j}^{m_j}$, where the weights w_j^i are strictly positive and normalised, so $\sum_i w_j^i = 1$. The group of observations is mapped as follows,

$$\psi : \left\{(x_j^i, w_j^i)\right\}_{i=1}^{m_j} \longmapsto \hat{\mu}_j := \sum_{i=1}^{m_j} w_j^i \phi_X(x_j^i) \tag{3.24}$$

and the kernel function is

$$k_\mu(\hat{\mu}_i, \hat{\mu}_j) = \sum_{r=1}^{m_i}\sum_{s=1}^{m_j} w_i^r w_j^s k_X(x_i^r, x_j^s).$$

Conditional independence testing can be approached in a similar way, although here there are more possibilities depending on which variables are aggregated. In our applications, the only aggregated (i.e., lower resolution) variable used was malaria incidence and this variable was never conditioned on (see Section 3.2.6 for more details). Therefore the only independence tests involving variables at different levels of aggregation that occur are testing X and Y for independence given $Z = (Z_1, ., , , Z_n)$, where X or Y (but not both) represents incidence which is aggregated and all other

variables are observed at a high resolution. In this case, for every variable apart from the aggregated one, the map (3.23) or (3.24) is applied to grouped observations and this reduces the problem to testing between variables on the same spatial scale. More generally, in a situation when an arbitrary subset of $X, Y, Z_1, ..., Z_n$ may be observed on the aggregate level, the conditional independence test could proceed as follows. If either X or Y is aggregated, for any other variable that is not aggregated use the map (3.23) or (3.24) to map each group of observations to a single point, then proceed with the test as usual. On the other hand, if X and Y are not aggregated but some of $Z_1, ..., Z_n$ are, then do not use the mapping described above. Instead, implement the test on the fine scale, with any aggregate values being repeated for all the relevant higher resolution observations. For example, suppose X and Y are being tested for independence given Z, where X and Y are measured at a pixel level and the value of Z is available for each administrative region. Let x_{ij} and y_{ij} be the values of X and Y, respectively, in pixel j of region i and let z_i be the value of Z in region i. Then the value of Z in a region is assigned to each pixel in that region, so the joint observations are $(x_{11}, y_{11}, z_1), (x_{12}, y_{12}, z_1), ..., (x_{1m_1}, y_{1m_1}, z_1), (x_{21}, y_{21}, z_2), ..., (x_{nm_n}, y_{nm_n}, z_n)$.

These ideas could also be extended to variables observed at different nested levels of aggregation, both in time or space. While the rules on when to embed variables and which spaces to embed them in may become more complex with these different levels, the fundamental process is unchanged.

3.3.6.1 Random Fourier approximations with aggregated data

In Section 3.3.2, we saw how RFFs can be used in scalable approximations to kernel methods. This was used in the RCIT and the RFF approximation for the HSIC. In this section, I show how similar approximations can be found when using the method described in the previous section for using aggregated data.

Returning to the example of unconditional independence testing in the previous

section, we saw that given joint observations of variables X (which is not aggregated) and Y (which is aggregated),

$$\left(\{x_1^i\}_{i=1}^{m_1}, y_1\right), \left(\{x_2^i\}_{i=1}^{m_1}, y_2\right), , ..., \left(\{x_n^i\}_{i=1}^{m_n}, y_n\right)$$

we can map the grouped observations of X to single points in a RKHS,

$$\{x_j^i\}_{i=1}^{m_j} \longmapsto \hat{\mu}_j.$$

The problem is then testing variables Y and X_μ for independence given observations $(y_1, \hat{\mu}_1), ..., (y_n, \hat{\mu}_n)$. The values $\hat{\mu}_i$ lie in a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ and we can define the kernel k_μ on this space simply as the linear kernel. In general, this function space cannot be identified with $\mathbb{R}^m$ for some m. Therefore we cannot apply Bochner's theorem and the RFF method in this space. However, we can use these ideas to find a similar approximation to the kernel k_μ.

Let k_X be the kernel on $\mathcal{X} = \mathbb{R}^d$ and assume this kernel is characteristic, continuous and translation-invariant. Just as we can find a low-rank map z which allows us to approximate $k_X(x, y)$ with $z(x)^T z(y)$ using RFFs, we would like to find a low-rank map to approximate the kernel k_μ associated with X_μ. Let Γ be the probability measure on $\mathcal{X}$ such that k_X is the Fourier transform of Γ. As described in Section 3.3.2, we can sample weights from this distribution, $v_1, ..., v_D$, and define the feature map $z : \mathcal{X} \to \mathbb{R}^{2D}$,

$$z : x \mapsto \frac{1}{\sqrt{D}} \left(\cos(v_1^T x), \sin(v_1^T x), ..., \cos(v_D^T x), \sin(v_D^T x)\right). \tag{3.25}$$

and we have that

$$k_X(x_1, x_2) \approx z(x_1)^T z(x_2).$$

To approximate k_μ, define the map $z_\mu : \text{Image}(\psi) \to \mathbb{R}^{2D}$ by

$$z_\mu : \hat{\mu}_j \overset{\psi^{-1}}{\longmapsto} \{x_j^i\}_{i=1}^{m_j} \longmapsto \frac{1}{m_j} \sum_{i=1}^{m_j} z(x_j^i)$$

where ψ is the map defined in (3.23). The inverse ψ^{-1} is well defined because ψ is injective when k_X is characteristic. The fact that this map is only defined on the image of this function is not a problem, as we only ever need to apply z_μ to elements in this space. We then have that

$$\begin{aligned} z_\mu\left(\psi(\{x_s^i\}_{i=1}^{m_s})\right)^T z_\mu\left(\psi(\{x_r^i\}_{i=1}^{m_r})\right) &= \left(\sum_{i=1}^{m_s} z(x_s^i)\right)^T \left(\sum_{k=1}^{m_r} z(x_r^k)\right) \\ &= \sum_{i=1}^{m_s}\sum_{k=1}^{m_r} z(x_s^i)^T z(x_r^k) \\ &\approx \sum_{i=1}^{m_s}\sum_{k=1}^{m_r} k_X(x_s^i, x_r^k) = k_\mu(\hat{\mu}_s, \hat{\mu}_r). \end{aligned}$$

We can therefore use the map z_μ to approximate the kernel k_μ. The map z_μ can then be used in RFF-based algorithms wherever the random feature map z would be used in the original algorithm. If the observations of X are weighted the equivalent map is

$$z_\mu : \hat{\mu}_j \overset{\psi^{-1}}{\longmapsto} \{(w_j^i, x_j^i)\}_{i=1}^{m_j} \longmapsto \sum_{i=1}^{m_j} w_j^i z(x_j^i).$$

To give a concrete example of this approximation, here I illustrate the application of RFFs via the map z to approximate kernel ridge regression and the analogous use of the map z_μ to approximate distribution regression. The former was discussed in more detail in Section 3.3.2. Firstly, suppose we have joint observations $(x_1, y_1), ..., (x_n, y_n)$ of variables X and Y where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. To perform standard ridge regression of Y against X, we would use the design matrix $\mathbf{X} = (x_1, ..., x_n)$. Let k be a continuous translation invariant kernel on $\mathbb{R}^n$. To approximate kernel ridge regression we could sample weights $v_1, ..., v_D$ from the corresponding probability distribution Γ. Then we

could perform ridge regression using the design matrix

$$\begin{aligned}\mathbf{X}' &= (z(x_1), ..., z(x_n)) \\ &= \frac{1}{\sqrt{D}} \begin{pmatrix} \cos(v_1^T x_1) & \cos(v_1^T x_2) & \cdots & \cos(v_1^T x_n) \\ \sin(v_1^T x_1) & \sin(v_1^T x_2) & \cdots & \sin(v_1^T x_n) \\ \vdots & \vdots & & \vdots \\ \cos(v_D^T x_1) & \cos(v_D^T x_2) & \cdots & \cos(v_D^T x_n) \\ \sin(v_D^T x_1) & \sin(v_D^T x_2) & \cdots & \sin(v_D^T x_n) \end{pmatrix}.\end{aligned}$$

Now suppose we have groups of observations of X and corresponding individual observations of Y,

$$\left(\{x_1^i\}_{i=1}^{m_1}, y_1\right), \left(\{x_2^i\}_{i=1}^{m_1}, y_2\right), , ..., \left(\{x_n^i\}_{i=1}^{m_n}, y_n\right).$$

Distribution regression proceeds by using the kernel k_μ, defined as above, in the place of the kernel k in kernel ridge regression (see e.g. Flaxman et al. [150] for more details). This allows for Y to be regressed against these grouped observations of X and predictions of Y to be made for new groups of observations of X. To approximate this using the map z_μ, we can simply use the design matrix

$$\begin{aligned}\mathbf{X}' &= (z_\mu(\psi(\{x_1^i\}_{i=1}^{m_1}\}), ..., z(\psi(\{x_n^i\}_{i=1}^{m_n}\})) \\ &= \frac{1}{\sqrt{D}} \begin{pmatrix} m_1^{-1}\sum_{i=1}^{m_1}\cos(v_1^T x_1^i) & \cdots & m_n^{-1}\sum_{i=1}^{m_n}\cos(v_1^T x_n^i) \\ m_1^{-1}\sum_{i=1}^{m_1}\sin(v_1^T x_1^i) & \cdots & m_n^{-1}\sum_{i=1}^{m_n}\sin(v_1^T x_n^i) \\ \vdots & & \vdots \\ m_1^{-1}\sum_{i=1}^{m_1}\cos(v_D^T x_1^i) & \cdots & m_n^{-1}\sum_{i=1}^{m_n}\cos(v_D^T x_n^i) \\ m_1^{-1}\sum_{i=1}^{m_1}\sin(v_D^T x_1^i) & \cdots & m_n^{-1}\sum_{i=1}^{m_n}\sin(v_D^T x_n^i) \end{pmatrix}\end{aligned}$$

within the standard ridge regression algorithm. This produces an approximation to

distribution regression that scales linearly with the number of data points.

3.4 Spatiotemporal prewhitening

So far we have addressed two challenges that occur when testing the independence of epidemiological, environmental or socioeconomic variables. The first was that the distributions of these variables or the relationships between them may not satisfy the assumptions that underpin many classical conditional independence tests. This is overcome by using kernel-based nonparametric tests. The second was that variables may be observed at different levels of aggregation, which was addressed by the method presented in the previous section.

One remaining issue is the fact that all variables are expected to have spatial or spatiotemporal structure. This has two effects, firstly it violates the assumption of almost all kernel-based independence tests that the joint observations are independent and identically distributed, and secondly it can lead to variables appearing dependent when they are not. The former is a less significant problem because spatiotemporal dependence does not invalidate any of the kernel-based tests we have discussed so far, although it may impact the convergence rates of estimators and the asymptotic null distributions (which I do not use). The latter is likely to have more serious effects on the results, since if dependence is inferred within a causal discovery algorithm where no dependence truly exists, edges that should have been removed may remain in the final output. It is worth noting that in many applications of causal discovery (such as feature selection or identifying confounding variables), while the presence of an edge that does not exist in the true DAG is undesirable from an efficiency perspective, it is unlikely to significantly bias the results. This is in contrast to, for example, missing an edge in the output that does exist in the true DAG. Nonetheless, we would like to avoid these mistakes if possible.

To see an example of the effect of spatial structure on measures of dependence, consider randomly sampling the values of a zero-mean Gaussian process with a radial basis function kernel at locations $x = 0, 0.1, ..., 10$ (which can be thought of as locations in space). Repeating this process, we can obtain joint observations at each value of x (each location) of different variables that are, by design, independent but each have spatial structure. We can also generate samples from similar variables with no spatial structure by repeatedly sampling 101 values (the number of locations) from a standard normal distribution. Figure 3.11 shows the density of pairwise correlations between variables where both have spatial structure, one has spatial structure and one does not, or neither do. We can see correlation between variables that both have spatial structure is usually higher than when just one or neither have this spatial autocorrelation, even though all these variables are independent. This is relevant when we use the permutation method to create the null distribution of the test statistic in our kernel-based tests. In this case, our test statistic will be based on two variables with spatial (or spatiotemporal) structure but when permuting one of these variables to create the null distribution this spatial structure will be removed. Therefore, our test statistic is likely to be higher than many values in the null distribution even when the variables are independent and so the type I error (rejecting the null hypothesis of independence when it is true) rate will be higher than the chosen level. When these tests are used within a causal discovery algorithm, this may lead to edges being retained incorrectly.

To account for this effect, I followed the suggestion by Flaxman et al. [151] of 'prewhitening' all variables—removing spatiotemporal structure—before testing for independence (or applying a causal discovery algorithm which uses independence tests). For variables $X_1, ..., X_n$, a spatiotemporal (or spatial for variables that do not vary over time) regression is carried out and residuals are calculated. Let X_{ij} be observation j of variable i which is at space-time location S_j. Then we calculate the

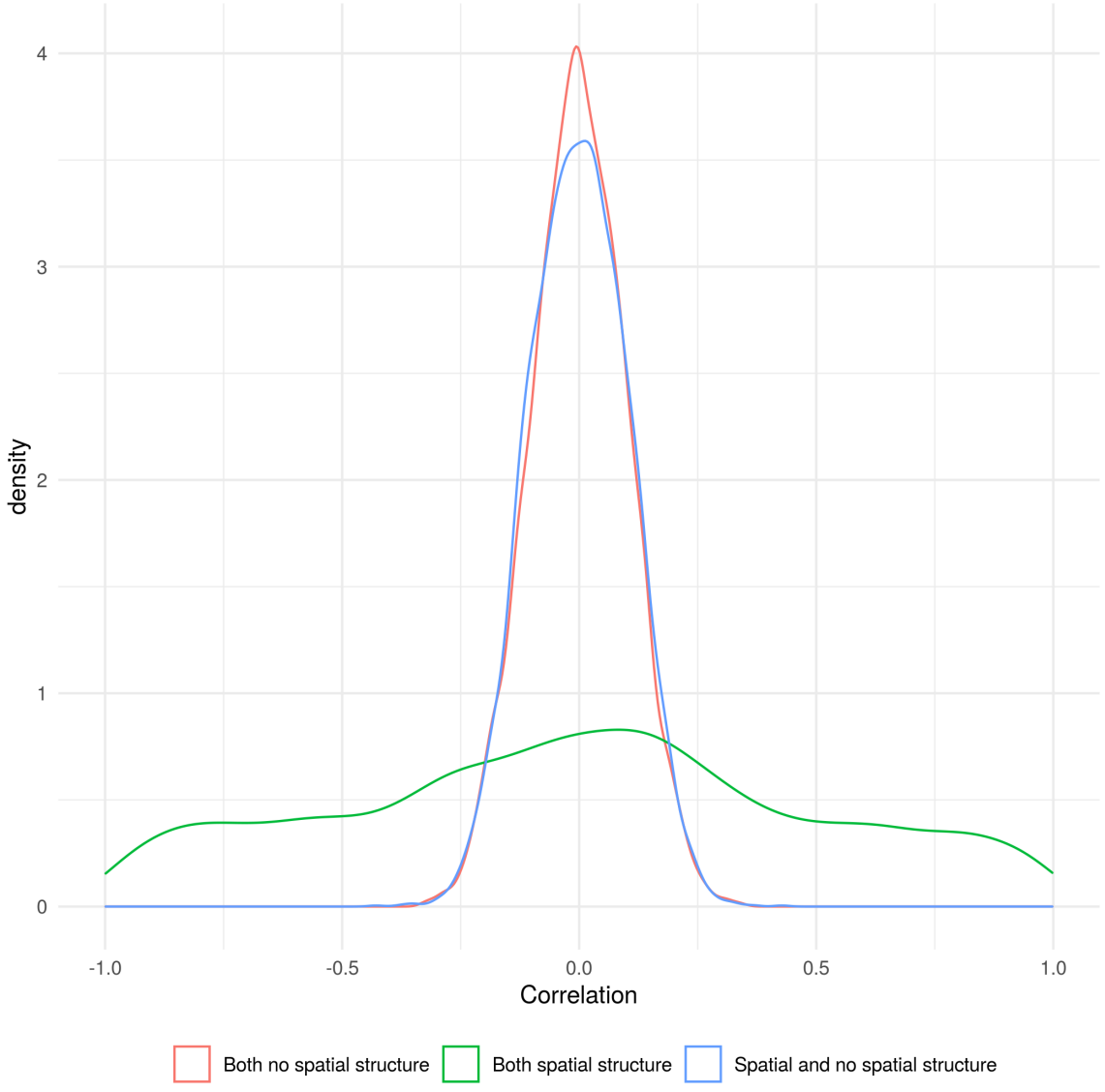


Figure 3.11: Density of correlations between independent variables when both have spatial structure (green), one has spatial structure and the other does not (blue) or neither have spatial structure (red).

residual

$$X'_{ij} = X_{ij} - \hat{\mathrm{E}}[X_i|S_j]$$

and the observed values of random variable X'_i, representing the variable X_i with the spatiotemporal trend removed, are $X'_{i1}, ..., X'_{im}$. Any independence tests or causal discovery algorithms are then performed using these detrended variables $X'_1, ..., X'_n$.

The argument here is similar to the idea behind the RESIT for conditional independence testing discussed in Section 3.3.5 but here we can think of space or time as the confounding variable. As in the RESIT method, this does depend on the existence of functional dependencies with space-time that can be identified and removed.

For environmental and socioeconomic covariates, the spatiotemporal regression is performed using a spatiotemporal Gaussian process, with the parameters of the kernel function learned within a Bayesian model or fixed. For incidence or prevalence information, I use a geostatistical model (with no covariates) to take into account the data generating process. Further details on the specific models used are given when discussing each application.

3.5 Causal discovery procedure

Having described the different components, I now describe the causal discovery procedure, which can be applied to any spatiotemporal dataset. Assume we have joint observations of a set of variables $\mathbf{V} = \{X_1, ..., X_n\}$. Causal discovery can be performed as follows:

1. Prewhiten each variable to remove spatial or spatiotemporal trends. Let X'_i be the detrended version of X_i for $i = 1, ..., n$.

2. Apply the PC algorithm or the two-step discovery method (as described in Section 3.2.6) to the joint observations of $X'_1, ..., X'_n$.

3. Within the PC algorithm use the RCIT test for conditional independence testing. Where aggregated variables are involved, map groups of observations to single points in a RKHS as described in Section 3.3.6.1.

4. Repeatedly subsample observations with replacement and apply the previous steps.

The only step here that we have not yet discussed in detail is subsampling the data and repeatedly running the discovery algorithm. This increases the stability of the results and gives some idea of the certainty of each edge, quantified by the proportion of repeats in which that edge occurs. In practice, the size of this sample is context-specific, generally chosen to be as large as computationally feasible. The output of the procedure is a set of CPDAGs $\mathbf{G} = \{G_1, ..., G_N\}$, where G_i is the output of the discovery algorithm for repeat i and N is the total number of repeats. For any variables X_i, X_j, define the *certainty* of the edge $X_i \to X_j$, denoted $c(X_i \to X_j)$, to be the proportion of output graphs that contain this edge,

certainty

$$c(X_i \to X_j) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \to X_j \text{ is in } G_i).$$

The certainty of an undirected edge can be defined in the same way. This is a measure of the confidence that this edge is in the true DAG. It is important to note that this measure is meaningful relative to the certainty of other edges but less so in absolute sense (for example, it is not a probability of the edge being including in the true DAG) because the result of the causal discovery algorithm depends on the significance level chosen for the conditional independence test. With lower significance levels, fewer edges will be retained (because the null hypothesis of independence is less likely to be rejected) and so each edge is likely to be retained in fewer of the outputs $G_1, ..., G_N$.

The problem of choosing this significance level in a principled way is not straightforward. Attempting to control the error rate for the entire algorithm is difficult

because the number of tests performed is not fixed. Furthermore, it may be more natural to seek to control the probability of incorrectly removing an edge. This corresponds to the type II error rate which is difficult to control, even within a single test, and depends on sample size. Therefore, in practice for each application the range of significance levels that produced reasonable graphs (in particular did not remove too many edges) was found via trial and error. This significance level was then fixed or randomly sampled within this range for each repeat. In most cases, I found that the relative certainties of edges were robust within these fairly large ranges of significance levels. It is worth noting that the problem of choosing this test significance level occurs whenever constraint-based causal discovery methods are used, rather than being specific to this implementation.

Chapter 4

An updated map of prevalence in Madagascar informed by health survey and routine case data

In this chapter I present work on estimating the prevalence of malaria in Madagascar between 2013 and 2016. The modelling strategy used here, an extension to the kinds of models introduced in Chapter 2, is a novel method for including both incidence and prevalence data to inform prevalence estimates. Despite increasing interest in using routine surveillance data to better understand malaria burden in sub-Saharan Africa, there are a number of sources of bias (including variation in treatment-seeking rates and reporting incompleteness) that limit the utility of these data and there have been few attempts to use this data in a risk mapping context [17]. In this work, I both develop this new methodology and apply it to make maps of prevalence in Madagascar that could be used to inform public health policy. An important feature of this method is the robustness to sources of bias in the incidence data. I also use the causal discovery procedure developed in Chapter 3 as a form of non-parametric feature selection. Some of the work in this chapter was published in [2] and presented to members of the Ministry of Health and National Malaria Control Programme (NMCP) of Madagascar in November 2020. I led this work, which was in collaboration with the NMCP and other members of the Malaria Atlas Project. Where parts of this work were carried out by other people this is clearly noted.

4.1 Malaria in Madagascar

Madagascar is among the most economically deprived countries in the world, with an estimated 74% of the population living below the international poverty line in 2019 [152]. Its resource-limited health infrastructure faces a number of challenges, including infectious diseases and widespread chronic malnutrition [153]. Among these, malaria remains one of the most significant, with an estimated 2.16 million cases leading to more than 5,000 deaths in 2018. All of the country's 26 million inhabitants are considered at risk of malaria [96]. In 2017, malaria was the 7th highest cause of mortality [154].

Malaria transmission in Madagascar is highly heterogeneous, reflecting the diverse ecological landscape of the island. Transmission has historically been low in the highland regions, which range from 800m to 1800m in altitude, due to unfavourable environmental conditions for the *Anopheles* vector. The highlands, which are believed to have previously been malaria free, experienced outbreaks in the late 19th century, potentially linked to the development of rice cultivation in the area (providing increased mosquito breeding sites) and population movement between the highlands and the higher transmission coastal regions [155, 156]. Transmission has been much higher in the coastal regions, with the exception of the desert south [157, 155, 54].

Efforts to combat malaria in Madagascar have been characterised by periods of intense control measures and relative success, and periods of resurgence in endemicity. Beginning in 1949, a long term campaign of DDT spraying and mass chemoprophylaxis with chloroquine, targeted at schoolchildren, virtually eliminated malaria in the highland regions and reduced infant mortality due to severe malaria across the country [158, 159]. However, the 1970s saw a reduction in control efforts as the public health system deteriorated, including interruption to DDT spraying and an end to mass chemoprophylaxis due to limited supplies of chloroquine. There was also increasing evidence of resistance of *P. falciparum* to chloroquine and decreasing clinical

effectiveness of the drug, although it remained the most common choice for prevention and treatment until the early 2000s [160, 161, 162, 163]. By the 1980s, transmission was increasing across the country and malaria had reappeared in the central highlands [164, 165, 166, 156]. Outbreaks were widespread in the late 1980s and were particularly severe in the central highlands, likely due to the lack of immunity in the population [167, 168]. Estimates of mortality in these regions due to large outbreaks in 1986 and 1987 range from 30,000 to 100,000 deaths [167].

In response to these outbreaks, control efforts were intensified across the country, with mass use of chloroquine (recommended to treat fever throughout the country) [169, 159], surveillance as part of an early warning system for outbreaks, and DDT spraying [170, 171, 172]. Control measures were scaled up further in the early 2000s, with increased funding from outside sources such as the US President's Malaria Initiative [173]. These measures included mass distribution of insecticide-treated bednets (ITNs), targeted indoor residual spraying (IRS), intermittent preventive treatment of pregnant women (IPTp) and a reduction in chloroquine use in favour of more effective artemisinin-based combination therapies (ACTs) [159, 173]. These efforts achieved substantial reductions in transmission, with an estimated 54% and 25% decrease in clinical incidence and mortality per capita, respectively, between 2000 and 2011 [26, 27]. However, the 2009 political coup had severe economic and social effects in the country [174]. The healthcare system was badly affected due to budget cuts and withdrawal of foreign aid, resulting in a lack of healthcare professionals, particularly in remote areas [153]. Over the past decade there has been a resurgence in malaria endemicity, with more than twice as many estimated cases and attributable deaths in 2018 compared to 2011 (the population increased from 21.7 million to 26.2 million in this time period) [96]. Similar trends can be seen when comparing prevalence in the 2011 and 2016 Malaria Indicator Surveys (MISs) [53, 175, 55].

In recent years, transmission has been highest in the east and west coastal regions,

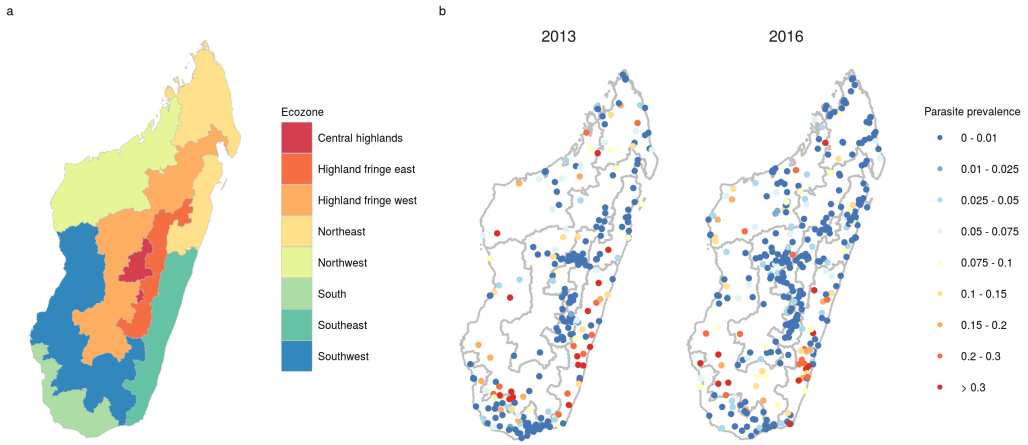


Figure 4.1: (a) Ecozones defined by Howes et al. [54] representing contiguous areas with distinct patterns of transmission. (b) Prevalence rates at survey sites in the 2013 and 2016 MIS.

where it follows a seasonal pattern with clinical incidence peaking between February and May depending on location. In the central highlands and the desert south, where transmission is lower, annual trends are less consistent and temporal variation appears to be driven by outbreak dynamics [176, 177, 54]. Howes et al. [54] identified eight contiguous ecozones representing distinct transmission settings (Central highlands, Highlands fringe west, Highlands fringe east, Northeast, Northwest, Southeast, Southwest, South) which are shown in Figure 4.1(a). The National Malaria Control Programme (NMCP) of Madagascar classifies 106 of the 114 districts in Madagascar as control areas, 3 pre-elimination and 5 elimination. Control strategies are currently stratified by risk level, with IPTp and mass distribution of ITNs in the 106 control districts and IRS targeted at high transmission districts in the southeast and southwest. In pre-elimination settings the focus is on outbreak and active case detection [16, 15].

4.2 Risk mapping

There are two main sources of data on malaria transmission in Madagascar, which have different strengths and limitations when it comes to understanding of spatiotem-

poral variation in risk. Routine malaria case data are collected through the Health Management Information System (HMIS) from reports from health facilities. These data are collected monthly and have a high spatial coverage (see Figure 4.3) but also suffer from a number of potential biases. This passive case detection is unlikely to capture all cases of malaria in the community, missing those that do not seek care or do so from informal or private providers, which likely represents a large fraction of the population (treatment-seeking rates in the public sector in 2013 were estimated to be around 45%) [178, 77, 55]. Furthermore, cases seen at health facilities may not be diagnosed or reported to the central system due to resource constraints (such as malaria rapid diagnostic test (RDT) stock-outs) or weak communication infrastructure [54]. Nevertheless, while these data may not represent all malaria incidence, they are an important source of information on trends in transmission due to their high temporal and spatial coverage [179, 180, 54, 181, 177].

Malaria Indicator Surveys provide another source of data for understanding the spatiotemporal patterns of malaria endemicity. The most recent surveys took place in 2011, 2013 and 2016 [175, 178, 55], with the planned 2020 survey ongoing [182]. These cross-sectional surveys are designed to be nationally representative and collect data on a number of indicators, including prevalence of *P. falciparum* infection in individuals between 6 and 59 months of age. They are conducted with a standardised methodology applied over all sites and surveys and, unlike routine case data, are not affected by reporting incompleteness, treatment-seeking behaviour or standards for clinical diagnosis. For these reasons, prevalence information from national health surveys has traditionally been the primary source of data for mapping malaria risk in sub-Saharan Africa [26, 53, 29, 27]. Kang et al. [53] used a Bayesian hierarchical model to map prevalence in Madagascar in 2011, 2013 and 2016 using parasite prevalence data from the 2011, 2013 and 2016 MIS reports [175, 178, 55]. However, these cross-sectional surveys, which are timed to take place during the high transmission

season, provide limited insight into seasonal patterns of transmission or transmission in years when no survey took place. Variation in the timing of surveys between years may also make it difficult to distinguish changes in prevalence between survey years from seasonal variation. Moreover, survey data is less informative in low burden areas, where sample sizes are likely to be inadequate to accurately assess changes in transmission due to low rates of detectable parasitaemia [18]. Operational planning for malaria control in Madagascar is predominantly based on raw incidence data from the HMIS, rather than prevalence survey data or the prevalence maps made by Kang et al. [53], because this data provides more recent and more frequent information on transmission [15, 183, 16].

In this work, routine case data and prevalence survey data were combined within a formal modelling framework, taking advantage of their relative strengths, in order to provide a more complete understanding of the spatiotemporal heterogeneity of transmission between 2013 and 2016. One method for combining prevalence and incidence data is to use a joint model, a single model which includes the likelihoods of both metrics and the relationship between them [184, 185, 186]. These models allow for sharing of information between metrics and produce predictions of both incidence and prevalence. However, balancing this sharing of information can be challenging, particularly in situations where one response variable has many more observations (and therefore a much greater likelihood contribution) than the other [187, 184]. Moreover, there is limited scope to learn the relationship between response variables without introducing too many degrees of freedom and therefore this relationship is often fixed or highly constrained. Lucas et al. [186], for example, modelled malaria prevalence and incidence jointly using a fixed previously estimated relationship [50]. To allow for learning the relationship between incidence and prevalence and to control the relative effect of each metric on the final results, I used a two-step process in which spatially smooth monthly incidence surfaces were produced, which were then used as

a covariate in a Bayesian geostatistical model of prevalence. This method is conceptually simpler than a joint model and allows the incidence-to-prevalence relationship to be learned within the prevalence model. An important benefit of learning this relationship is that systematic biases in routine case data are implicitly accounted for. Modelling the incidence and prevalence processes separately is equivalent to making a 'cut' between the incidence and prevalence processes in the model. Modularising the inference in this way prevents over-reliance on the less reliable incidence data in the final prevalence estimates [187, 184]. This approach is similar to the use of modelled surfaces, such as temperature suitability [188] or accessibility [57], as inputs when mapping malaria risk [26, 53, 29, 27]. It is also similar to the work of Lucas et al. [189] who used modelled prevalence surfaces as inputs to an incidence model. However, their prevalence model used environmental and socioeconomic covariates and therefore overall this approach was equivalent to a non-linear model of incidence using these covariates. In contrast, the incidence surfaces in this model are a direct smoothing of treatment-seeking adjusted observed incidence rates and therefore represent malaria risk more directly.

This two-step model was used to incorporate routine case data in order to update the estimates of prevalence for 2013 and 2016 made by Kang et al. [53] and to produce estimates for 2014 and 2015. With monthly prevalence maps over all four years, both long term and seasonal trends in prevalence across the country could be assessed. The eight ecozones identified by Howes et al. [54] were used as a basis for assessing how these trends varied spatially.

4.3 Methods

4.3.1 Study data

Parasite prevalence data was available from Malaria Indicator Surveys that took place in Madagascar in 2013 and 2016. These data consist of geo-located clusters where the prevalence of *P. falciparum* infection was measured, determined by microscopy, in individuals between 6 and 59 months of age ($Pf\mathrm{PR}_{6-59\mathrm{mo}}$). These surveys were conducted with a standard protocol and survey sites were selected to produce nationally representative estimates of prevalence in this age group. The number of positive individuals and the total number tested was recorded at each site. In the 2013 and 2016 surveys, 6323 and 6927 individuals were screened across 274 and 358 sites, respectively. The survey sites and proportions of positive individuals are shown in Figure 4.1(b) and full details can be found in the original reports [178, 55].

Monthly health facility case data between January 2013 and December 2016 were provided by the NMCP. These data come from HMIS data reports and represent clinical cases of malaria confirmed by an RDT for all ages, irrespective of parasite species. Data were available from 3342 health facilities in Madagascar, of which 2801 were geo-located and verified using a separate dataset of geo-located health facilities from the Institut Pasteur de Madagascar (as described in [177]). This data cleaning and verification was carried out by Suzanne Keddie and Emma Collins.

A number of covariates were used to inform the prevalence model, which are detailed in Table 4.1. Most of these variables were environmental, influencing vector abundance, parasite behaviour and environmental suitability. Two of these covariates (accessibility to cities [57] and night lights [83]) relate to the development and urbanicity of a location and therefore are related to both vector abundance and socioeconomic factors, such as access to healthcare. These covariates were processed by Jennifer Rozier, Camilo Vargas-Ruiz and Joseph Harris. In total there were 26

potential features for the prevalence model—8 static covariates and 6 dynamic covariates, each considered at 0, 1 and 2 month time lags. The features included in the final model were selected using the causal discovery procedure described in Chapter 3. This selection method is described in more detail in Section 4.3.5.

4.3.2 Catchment population

In order to calculate incidence rates at each health facility, estimates for catchment populations (the number of people likely to seek treatment at each facility) were needed. A catchment model was used to estimate these (treatment-seeking) catchment populations, based on travel time to health facilities. This model was designed by myself, Ewan Cameron, Suzanne Keddie and Emma Collins. The travel times were calculated by Suzanne Keddie and Katherine Twohig. The functional relationship between travel times and treatment-seeking behaviour was designed by Katherine Twohig and myself. I implemented the rest of the model.

The country was divided into a grid of approximately 5km-by-5km pixels. For each pixel, the travel time to each health facility was calculated using a friction surface (defining travel time through each pixel) developed by Weiss et al. [57] and a least cost algorithm [191]. Given that an individual in pixel i seeks treatment, the probability they seek treatment in health facility j, $p(\text{pixel}_i \rightarrow \text{HF}_j)$, was modelled as inversely proportional to square of the travel time to that health facility. That is,

$$p(\text{pixel}_i \rightarrow \text{HF}_j) = \frac{t(\text{pixel}_i \rightarrow \text{HF}_j)^{-2}}{\sum_{j=1}^{N_{\text{HF}}} t(\text{pixel}_i \rightarrow \text{HF}_j)^{-2}}$$

where $t(\text{pixel}_i \rightarrow \text{HF}_j)$ represents the travel time from pixel i to health facility j and N_{HF} was the total number of health facilities. The catchment population of health

Table 4.1: List of covariates.

Covariate	Description	Type	In any causal feature set	In final feature set
Rainfall [190]	Rainfall estimates from rain gauge and satellite observations	Dynamic	Lag 0	Lag 0
LST day [84]	Daytime land surface temperature	Dynamic	No	No
LST night [84]	Nighttime land surface temperature	Dynamic	Lag 1	No
TCB [85]	Tasselled cap brightness; spectral measure of land reflectance, associated with ground cover type	Dynamic	Lag 2	No
EVI [86]	Enhanced vegetation index	Dynamic	No	No
TSI Pf [188]	Temperature suitability index for *P. falciparum*	Dynamic	Lag 2	No
Accessibility [57]	Distance to cities with population > 50,000	Static	Yes	Yes
AI [80]	Aridity index; ratio of rainfall to evapotranspiration rate, calculated from historical (1970-2000) climatic data	Static	Yes	Yes
Elevation [56]	Elevation as measured by the Shuttle Radar Topography Mission	Static	Yes	No
PET [80]	Potential Evapotranspiration; estimated capacity of the landscape to move water from the surface and soil to the air, calculated from historical (1970-2000) climatic data	Static	Yes	No
Slope [56]	Average slope of the pixel, determined from the elevation of pixels in its neighbourhood	Static	Yes	No
Night lights [83]	Index that measures the ambient luminosity of towns, cities and other sites with persistent lighting	Static	No	No
Distance to water [58]	GIS-derived surface that measures distance to permanent and semi-permanent water based on presence of lakes, wetlands, rivers and streams; accounting for slope and precipitation	Static	Yes	Yes
TWI [56]	Topographic wetness index; slope-based estimate of local drainage capacity	Static	No	No

facility j was then calculated as

$$\sum_{i=1}^{N_{\text{pixel}}} \text{pop}_i \times p(\text{pixel}_i \rightarrow \text{HF}_j)$$

where pop_i was the treatment seeking-adjusted population of pixel i and N_{pixel} was the total number of pixels. The proportion of the population in each pixel who would seek treatment at any health facility was also modelled as a function of travel time, which has been identified as an important factor in treatment-seeking behaviour for fever in Madagascar [192]. This proportion was modelled as a logistic function (similar to the functional forms considered by Alegana et al. [193])

$$\frac{\alpha}{1 + \exp(\sigma t)} + \beta$$

where t was the travel time to the nearest health facility in minutes. The parameters values $\alpha = 0.6$, $\sigma = 0.00916$, $\beta = 0.15$ were chosen such that the maximum and minimum possible treatment-seeking proportions were 0.6 and 0.15, and the treatment-seeking proportion at $t = 120$ minutes was 0.3. These parameter values were chosen to produce a similar relationship between treatment-seeking and travel time as that observed in the 2013 and 2016 MISs [178, 55] (see Figure 4.2) and match overall estimated treatment-seeking rates [77]. Catchment populations were calculated for each year between 2013 and 2016 using hybrid population surfaces from the Gridded Population of the World v4 [194] and WorldPop [195].

In order to assess the sensitivity of the final estimates to these modelling assumptions, three alternative sets of populations were considered: catchment populations supplied by the NMCP (based on the 1993 national census adjusted by a fixed annual growth rate [196] with no adjustment for treatment-seeking behaviour) and two sets of catchment populations generated by the catchment model under different treatment-seeking parameters (assuming lower and higher rates of treatment-seeking). In the

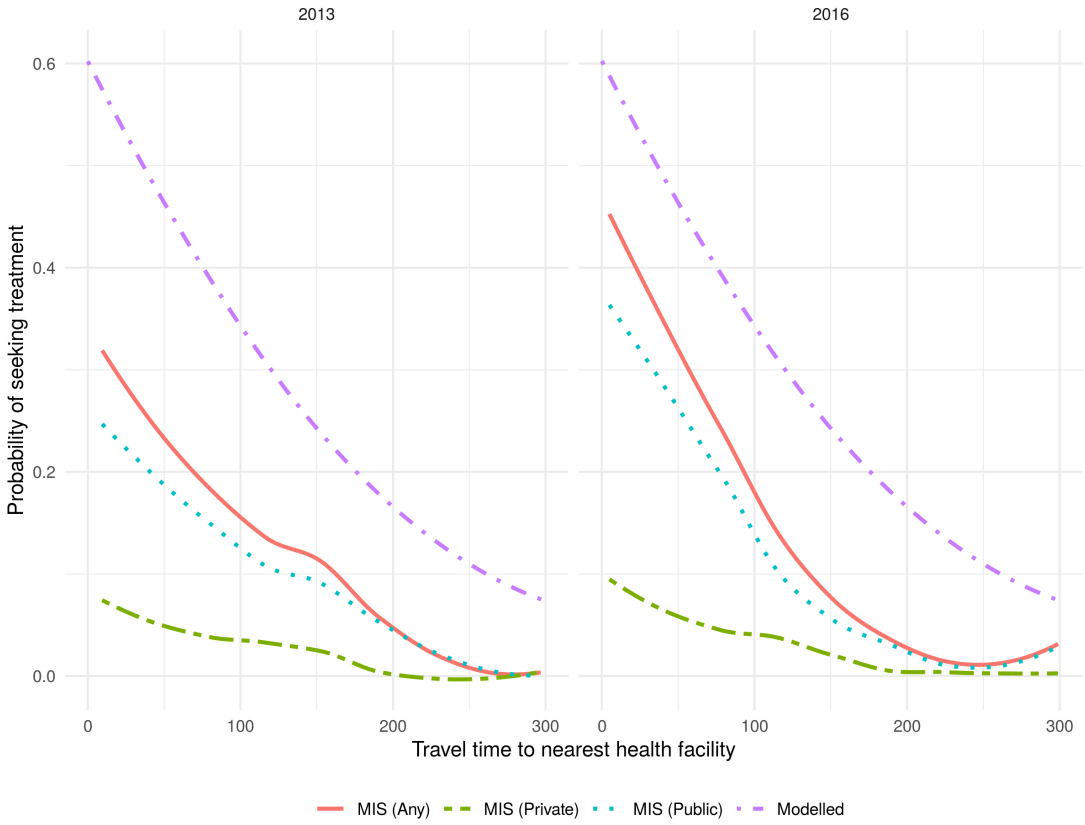


Figure 4.2: Locally smoothed estimates of treatment-seeking rates (at public, private and any facility) against travel time to the nearest health facility from the 2013 and 2016 MISs and the relationship used in the catchment model.

lower treatment-seeking setting, the parameters were chosen such that maximum and minimum treatment-seeking proportions of 0.6 and 0 and a proportion of 0.1 at $t = 90$ minutes. In the higher treatment-seeking setting, maximum and minimum treatment-seeking proportions were 0.6 and 0.25, with a proportion of 0.3 at $t = 200$ minutes.

4.3.3 Incidence model

Routine case data from health facilities were modelled using a Bayesian geospatial model to produce monthly incidence surfaces which were then used as inputs to the prevalence model. Let C_{it} be the number of cases observed in month t (t =1, ..., 48) at health facility i (i =1, ..., N), which is at location s_i and has a treatment-seeking catchment population E_i. The number of cases observed was modelled as a Poisson process

$$C_{it} \sim \text{Pois}(E_i \times \lambda_{it})$$

with mean equal to the product of the catchment population and the underlying incidence rate, λ_{it}. The log incidence rate was modelled as

$$\log \lambda_{it} = \beta_0 + f(s_i, t)$$

where β_0 was an intercept and for each month $f(\cdot, t)$ was a realisation of a Gaussian process over space with zero mean and Mátern covariance structure. Unlike in the other applications of Mátern random fields in this thesis, the parameters of the Mátern covariance function (the range, ρ, and marginal variance, σ^2) were not learned during the fitting process under a penalised complexity prior. This is because this prior provided insufficient regularisation (assessed by visual inspection) even under strict values of tuning parameters (see Section 2.2.2). Instead, the values of the range and marginal variance were chosen by a search over parameter space to maximise accuracy

when predicting incidence in held out locations. These parameters were optimised jointly over all months. The value of smoothness parameter, ν, was fixed at 1.

4.3.4 Prevalence model

Prevalence data from MIS surveys were also modelled using a Bayesian geospatial model, informed by environmental and socioeconomic covariates and the modelled incidence surfaces. Let Y_i be the number of infected individuals and N_i be the number of individuals tested in survey i ($i = 1, ..., M$), taking place in location s_i at time t_i. The results of the survey were modelled as a realisation of a binomial process

$$Y_i \sim \text{Binomial}(p_i, N_i)$$

with underlying prevalence p_i at this location and point in time. Prevalence was modelled as

$$\text{logit}(p_i) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0^{\text{inc}} g(\lambda(s_i, t_i)) + \beta_1^{\text{inc}} g(\lambda(s_i, t_i + 1)) + f(s_i)$$

where $\beta_0, \boldsymbol{\beta}, \beta_0^{\text{inc}}, \beta_1^{\text{inc}}$ were parameters (with β_0^{inc} and β_1^{inc} non-negative), $\boldsymbol{X}_i$ were covariate values and $\lambda(s, t)$ was the log incidence value from the modelled incidence surfaces at location s and time t. f was modelled as a realisation of a Gaussian process over space with Mátern covariance, parameterised by the range ρ and marginal variance σ^2, while g was a realisation of a Gaussian process over incidence with a squared exponential kernel, parameterised by the scale κ (with fixed variance 1). This allowed a non-linear effect of incidence on logit-prevalence while limiting model-complexity by assuming a priori that this relationship is smooth (by the choice of squared exponential kernel and placing an appropriate prior on the scale parameter). This flexibility was important as incidence is likely to be the main driver of prevalence in the model (while the additional covariates are likely to be less informative and therefore were

modelled as having linear effects) and assuming a linear effect produces prevalence-incidence relationships that do not match empirical observations [50]. Incidence in the subsequent month was included in addition to incidence in the current month, as the presence of parasites in the blood could result in a clinical case that is recorded in a health centre several weeks later. The parameters β_0^{inc} and β_1^{inc} allowed the model to learn the relative predictive power of incidence at these two time points.

The Bayesian model was completed by placing appropriate priors on the parameters and hyperparameters. Normal priors were placed on $\beta_0, \boldsymbol{\beta}, \beta_0^{\text{inc}}, \beta_1^{\text{inc}}$ centred around 0 with standard deviation 1 for β_0 and 0.25 for $\boldsymbol{\beta}, \beta_0^{\text{inc}}, \beta_1^{\text{inc}}$. The Mátern covariance parameters were given log-normal priors with $\log \rho \sim \text{N}(3, 0.1^2)$ and $\log \sigma \sim \text{N}(0, 0.1^2)$, shrinking the spatial term towards a fairly flat and smooth field. A log-normal prior was placed on κ with mean $\log \kappa \sim \text{N}(3, 0.1^2)$. Both the incidence and prevalence models were fitted using the Gaussian Markov Random Field approximation [67] with the Template Model Builder package [197] in R [198] and samples were drawn from a Laplace approximation to the posterior to produce associated uncertainty estimates, as discussed in Chapter 2. Uncertainty was quantified using the interquartile range of the posterior distribution and exceedance and non-exceedance probabilities. For a given prevalence threshold, the exceedance probability at a location is defined as the probability that a prevalence sample from the posterior at this location will exceed the threshold. Similarly, a non-exceedance probability is the probability a prevalence sample will not exceed a given value. The model was validated by fitting the model with data from the 2013 survey and making predictions for 2016 (to evaluate model performance when predicting prevalence in previously unobserved time points) and through k-fold cross validation.

4.3.5 Causal feature selection

In total there were 26 potential features, 18 dynamic (6 covariates each at 0, 1 and 2 month time lags) and 8 static. Feature selection was performed using the causal discovery procedure described in Section 3.5, which I refer to as *causal feature selection*.

causal feature selection

When applying the causal discovery procedure the prevalence observations were treated as point-level observations so the modified independence tests described in Section 3.3.6 for use with aggregated data were not needed. Each prevalence observation was associated with a single value for each covariate (the value in the same location and, for temporal covariates, the same month as the observation).

During the prewhitening step, the 2013 and 2016 observations were fit separately. Within each survey year, each variable was fit with a spatial Gaussian process with Mátern covariance. The range and scale parameters were learned within the fitting process and were given log-normal priors with $\log \rho \sim \mathrm{N}(2, 0.1^2)$ and $\log \sigma \sim \mathrm{N}(-2, 0.1^2)$. See Section 5.1.2.1 for more discussion on the choice of these priors. A spatial, rather than spatiotemporal, Gaussian process was used because the data was not sufficient to accurately identify temporal trends. As each survey took place over a short time period during the high transmission season, any the effect of temporal trends over this period was expected to be small.

After running the causal discovery procedure, potential feature sets were generated as follows. For any value of $\lambda \in (0, 1]$, a potential feature set could be produced by retaining any features X such that the certainty of the edge $X \to P$ exceeded λ, where P was the variable representing prevalence. In other words, any value of $\lambda \in (0, 1]$ defines a potential feature set,

$$F_\lambda = \{X \mid c(X \to P) > \lambda\}.$$

Varying λ between 0 and 1 produces a list of potential feature sets $\{F_\lambda | \lambda \in (0, 1]\}$.

The number of sets in this list is at most the number of possible features, since as λ decreases more features are included in order of certainty of the features being direct causes of prevalence.

In this case, the causal discovery algorithm was run on the dynamic and static features separately. By this I mean that the causal discovery algorithm was run on joint observations of all static features and prevalence and then again on joint observations of all dynamic features and prevalence. This produced sets of potential static feature sets S_i and dynamic feature sets T_j. This was done because if all features were included together, with the relatively small sample size it would have been difficult to choose a significance level which allowed edges between static and dynamic variables (which appear less related due to the fixed nature of the static variables) to be retained while not retaining the majority of edges between pairs of static or pairs of dynamic variables. With a larger sample size (as in Chapter 5) this is of less concern.

The static and dynamic feature sets identified (by varying λ) were then combined to produce a number of potential feature sets. For example, if the causal discovery algorithm identified three static sets, S_1, S_2, S_3, and two dynamic sets, T_1, T_2, (in each case by varying the threshold for inclusion of a feature) then there would be six potential feature sets — $S_i \cup T_j$ for each $i = 1, 2, 3$ and $j = 1, 2$. The final feature set used was the set that maximised cross-validation performance (within the training set when evaluating model performance), where performance was quantified as correlation between observed and predicted prevalence rate.

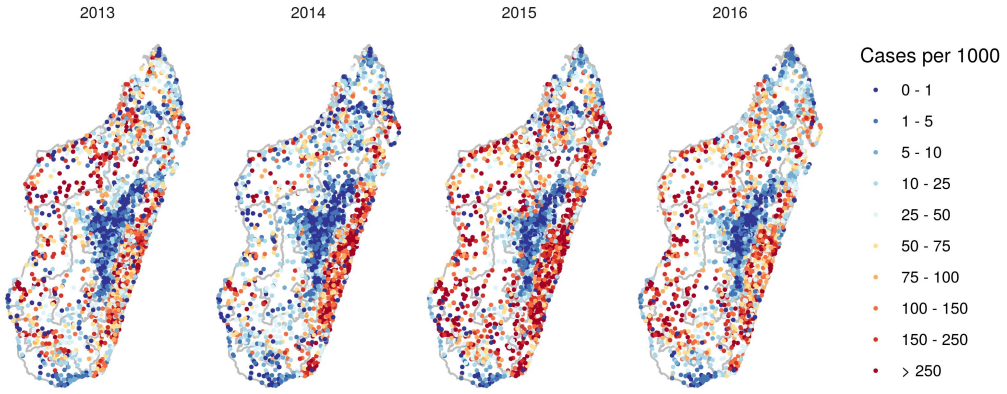


Figure 4.3: Annual incidence rates at each health facility based on routine case data and modelled catchment populations.

4.4 Results

4.4.1 Catchment populations

The median size of the modelled catchment populations was 2890 (LQ: 1710, UQ: 4740). The total population served according to the modelled populations increased from 10.15 million in 2013 to 11.02 million in 2016, corresponding to approximately 43% of the Malagasy population each year, which is largely in agreement with estimated treatment-seeking rates [77].

4.4.2 Incidence

Annual incidence rates at each health facility (calculated using reported cases and modelled catchment populations) are shown in Figure 4.3. Spatial patterns of incidence in 2013 and 2016 were similar to observed prevalence (Figure 4.1(b)) with lower rates in the central highlands and the south, and higher rates in the east and west coastal regions. Compared to 2013, incidence in 2016 was generally higher in health facilities in the southeast and southwest but lower in the north. Across the four years, rates were generally lowest in 2014 (with the exception of the east coast)

and highest in 2015. Figure 4.4 shows monthly incidence rates aggregated by ecozone. A seasonal pattern of a single peak in incidence between January and April can be seen to some extent in most regions, most clearly in the Southeast, Northeast and (despite low overall incidence) Highland fringe east ecozones. An increase in cases in 2015 was observed in almost all ecozones.

The optimal hyperparameters for the incidence model were $\rho = e^{-0.1}$ and $\sigma = e^{-2}$. Figure 4.5 shows the smooth incidence surfaces produced by this model aggregated annually, which as expected reflect the overall spatiotemporal trend of the observed data. It should be noted that these surfaces were only intended to be a spatial smoothing of the treatment seeking-adjusted case data reflecting relative spatial trends, rather than an enumeration of true incidence rates, and are an intermediate step in the modelling process rather than an output.

4.4.3 Prevalence

The causal discovery algorithm identified 4 dynamic and 4 static feature sets which combined to give 16 potential feature sets for the prevalence model. The features that were present in at least one of these sets are listed in Table 4.1 and these sets are listed in full in the appendix. The features used in the final model were rainfall [190] with no time lag, accessibility to cities [57], an aridity index [80] and distance to water [58]. The posterior mean and 95% credible interval for all the model parameters, including coefficients of these features, are shown in Table 4.2. As expected, incidence was the most important predictor, with incidence in the current month having a greater effect than incidence in the following month. The other covariates in the model had smaller and less consistent effects on prevalence. When fitting the model on 2013 data and making predictions for 2016, there was a correlation of 0.63 between predicted and observed values. For 3, 4 and 5-fold cross-validation, there were average correlations of 0.58, 0.59 and 0.6, respectively. These values are reasonably high given

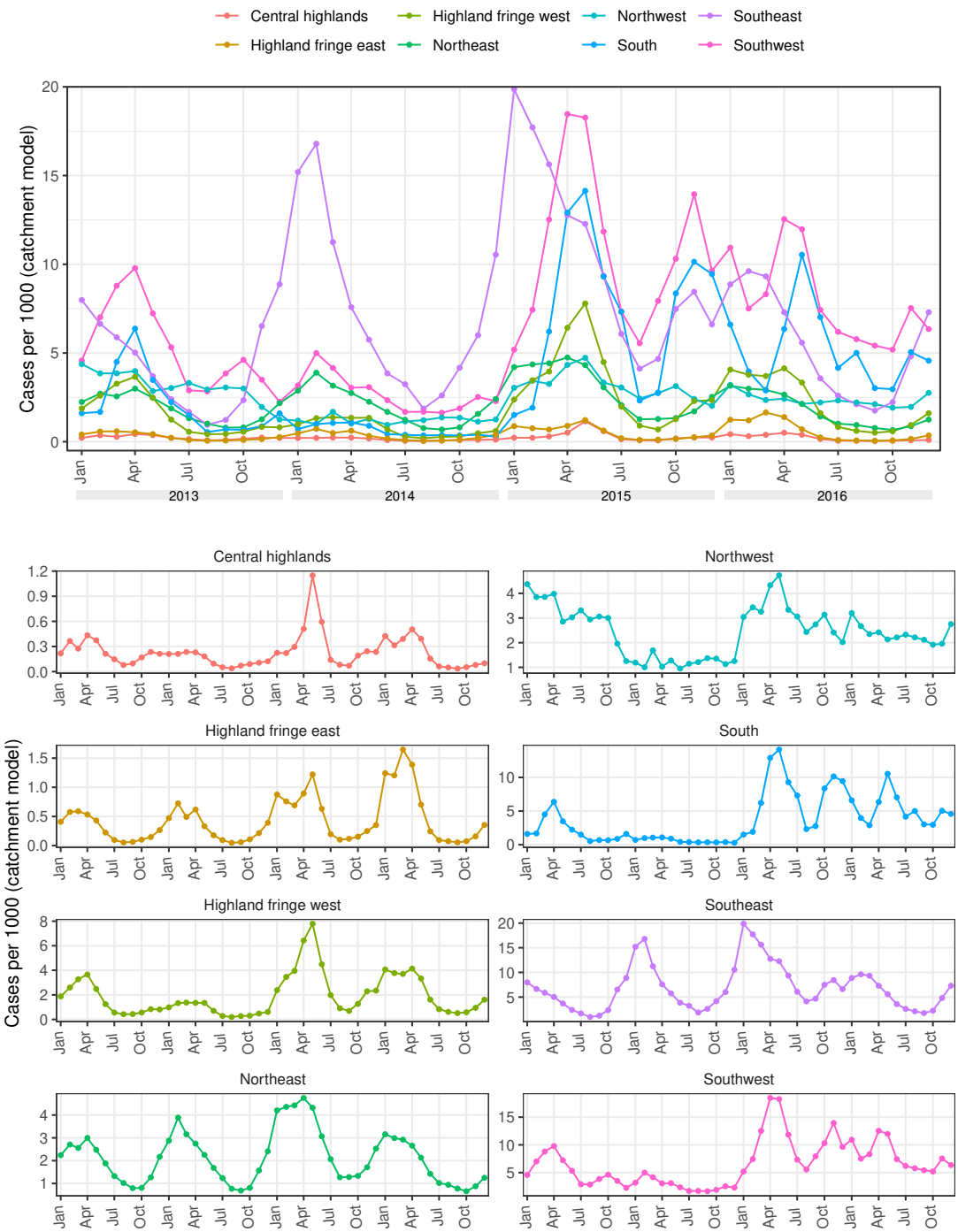


Figure 4.4: Monthly incidence rates at each health facility based on routine case data and modelled catchment populations stratified by ecozone (2013-2016).

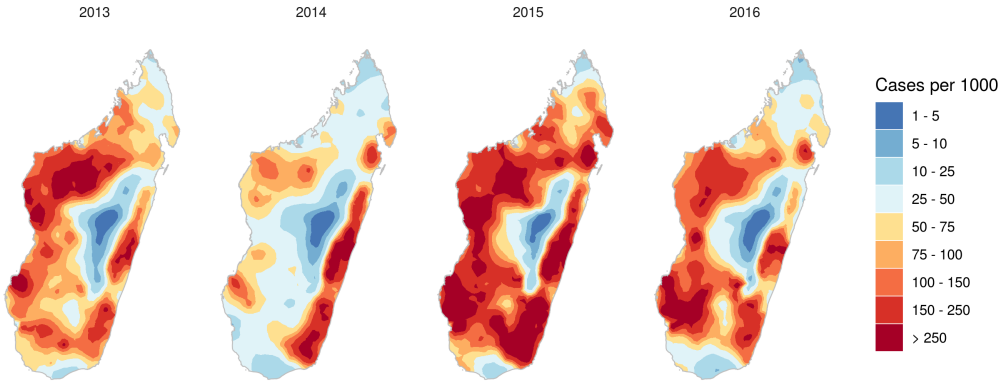


Figure 4.5: Incidence surfaces for all ages aggregated annually.

that the observed prevalence rates are themselves noisy point estimates of underlying prevalence; for context, the standard errors of the mean of binomial samples of 20 individuals (the median survey size) with true prevalence values of 0.01, 0.05 and 0.15 are 0.022, 0.049 and 0.080, respectively.

Table 4.2: Mean and 95% credible intervals for the prevalence model parameters.

Parameter	Mean	CI
Intercept	-3.956	-6.082, -1.926
Accessibility	0.060	-0.021, 0.150
AI	-0.066	-0.172, 0.042
Distance to water	-0.031	-0.108, 0.042
Rainfall (no lag)	0.019	-0.097, 0.128
$\log \rho$	1.480	1.263, 1.693
$\log \sigma$	-0.762	-0.968, -0.561
$\log \lambda_{\text{inc}}$	1.883	0.881, 3.043
β_0^{inc}	0.802	0.407, 1.22
β_1^{inc}	0.497	0.184, 0.879

Prevalence estimates for individuals between 6 and 59 months of age are shown aggregated annually in Figure 4.6(a). Spatial patterns were similar across all four years, with highest estimated prevalence near the southeast coast and other areas of high prevalence along the west coast. Prevalence estimates were consistently low in the far south of the country and in the central highlands. Prevalence in the north varied more

between years, although was generally low or moderately low. Population-weighted mean prevalence was lowest in 2014 (6.0%, 95% credible interval (CI): 3.3-8.8), followed by 2013 (6.4%, CI: 3.5-9.5) and 2016 (6.6%, CI: 3.6-9.6), and was highest in 2015 (8.9%, CI: 4.6-12.9). Population-weighted mean prevalence over time (Figure 4.6(b)) showed a clear seasonal pattern, peaking between February and April each year with the lower prevalence occurring between July and September. Prevalence was highest in 2015, with the peak prevalence across all four years occurring in April 2015 (11.4%, CI: 5.8-16.0) and notably high prevalence sustained into the lower transmission season. By the second half of 2016, prevalence appeared to have returned to similar levels to 2013 and 2014.

Figure 4.7 shows the mean prevalence over time for each ecozone. Prevalence was consistently highest in the Southeast ecozone, with peak monthly prevalence of 0.3 or more in all years, while consistently lowest in the Central highlands and Highland fringe east ecozones. Seasonal patterns of prevalence are present in most of these ecozones, including both higher transmission regions (Northeast, Southeast and Southwest) and the lower transmission highland ecozones. These regional seasonal trends are similar to the pattern observed for mean prevalence across the country, with a single peak occurring between February and April in most areas (though often earlier in the Southeast ecozone) and lowest prevalence around August. The Northwest and South ecozones exhibit some similar seasonal trends but these are less consistent. In most areas, prevalence in 2014 was slightly lower than in 2013, with large decreases in the Northwest and South ecozones. Increased prevalence in 2015 was observed in all areas. In the highland ecozones, prevalence was higher than normal around April but later in the year returned to similar levels to previous years. In the rest of the country, a high peak in April was generally followed by higher than average prevalence throughout the rest of the year and into 2016. The ecozones generally correspond well to these latest prevalence estimates (Figure 4.8 shows the

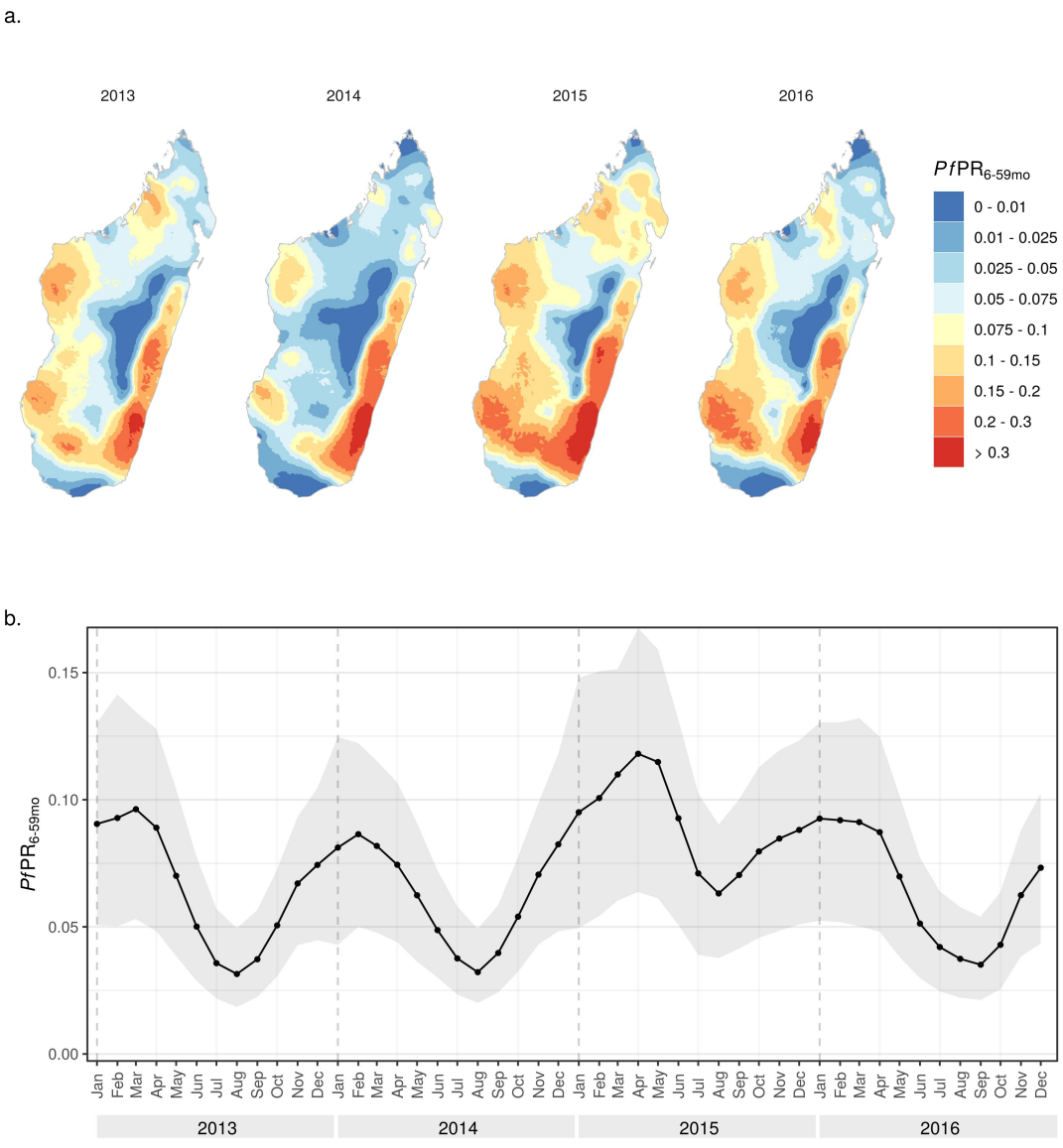


Figure 4.6: Prevalence estimates for individuals between 6 and 59 months of age, (a) aggregated annually and (b) population-weighted mean over time with 95% credible intervals.

2016 estimates with ecozone borders highlighted). The greatest heterogeneity within ecozones was in the Highland fringe west ecozone, which contained a region of higher prevalence to the south, and the Northwest ecozone.

Monthly prevalence estimates can also be used to identify areas where transmission is consistently low or high. Two examples are shown in Figure 4.9 for estimates of prevalence in 2016. Figure 4.9(a) shows areas where prevalence was always below 5% or below 5% for at least 9 months and Figure 4.9(b) areas where prevalence was always above 20% or above 20% for at least 6 or 9 months. Despite the seasonal trends in transmission (Figure 4.7), in much of the central highlands prevalence was below 5% year-round, while some areas in the western highlands only exceeded 5% for at most 3 months of the year. Transmission was also consistently low in parts of the desert south and far north of the country. The areas of consistently high prevalence were concentrated on the east and southwest coasts, following a similar pattern to overall prevalence.

Figure 4.12 shows two ways of visualising the uncertainty in prevalence estimates, using annual 2016 estimates as an example. The first (Figure 4.12(a)) maps the interquartile range of the posterior distribution. The spatial patterns of uncertainty are similar to the prevalence estimates, with higher (absolute) uncertainty areas of high prevalence. An alternative quantification of uncertainty is in terms of probability of estimates exceeding (or not exceeding) chosen values. Figure 4.12(b) shows the probability of prevalence (sampled from the posterior distribution) exceeding 0.15. This can be interpreted as the probability that the true prevalence in 2016 exceeded 0.15. In the majority of the country there is a very high certainty that prevalence was below 0.15, while on the east and west coasts there is fairly high confidence that prevalence exceeded this value. In the southwest, in particular, the probability of exceeding 0.15 is often above 90%. Similarly, we can visualise the probability of not exceeding a certain value, as shown in Figure 4.12(c) with the probability of not

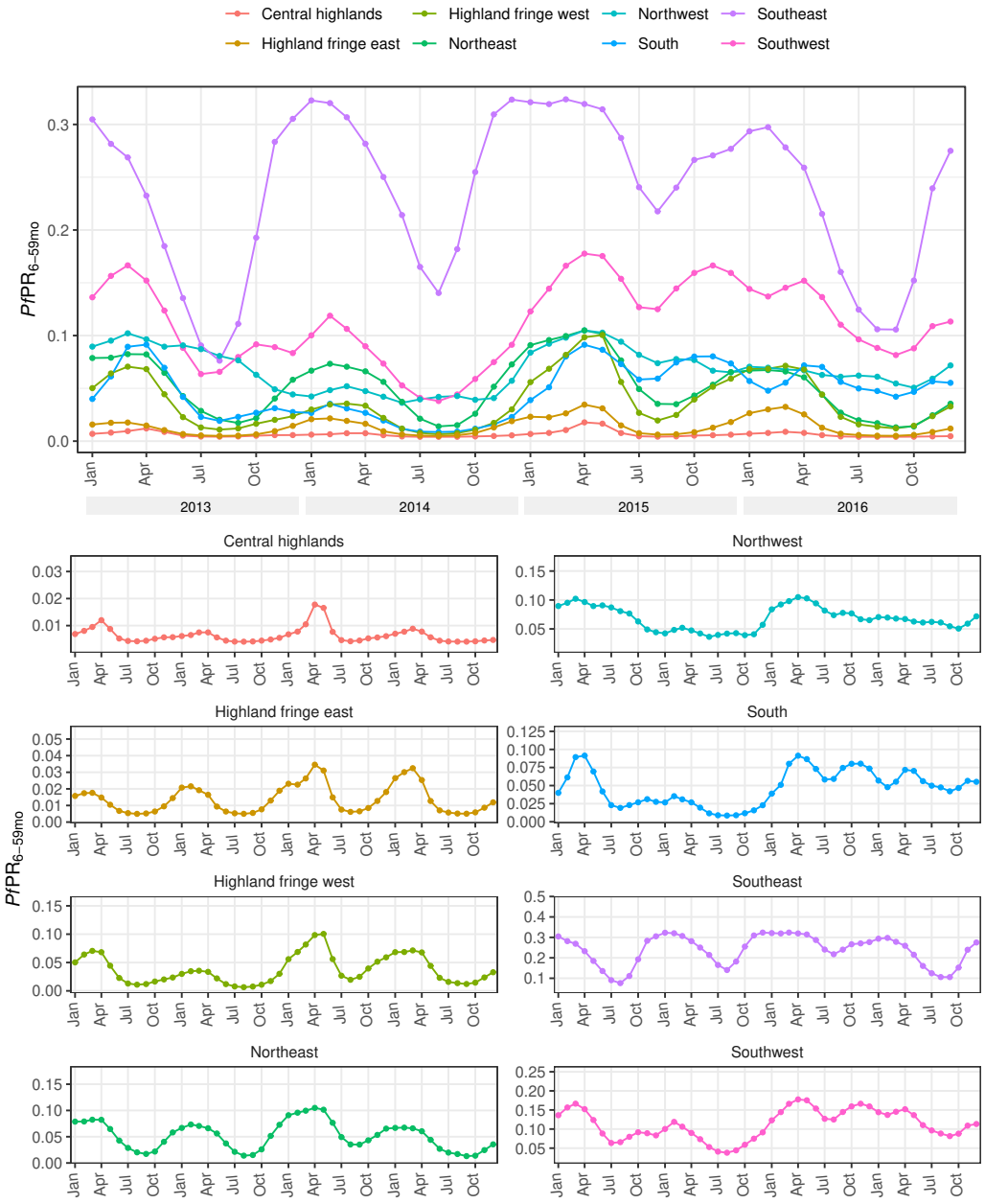


Figure 4.7: Population-weighted mean prevalence over time stratified by ecozone with 95% credible interval.

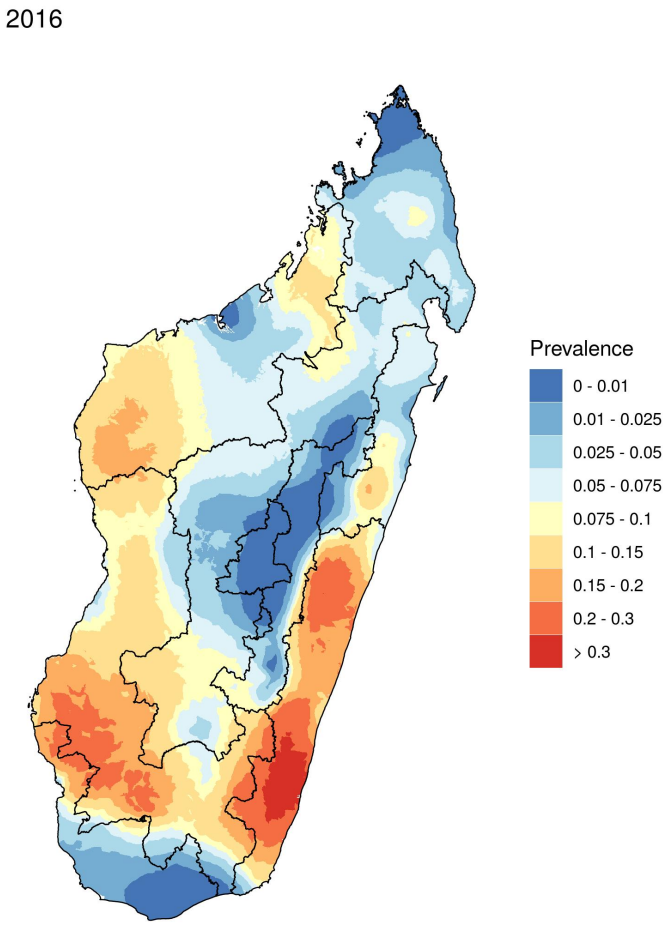


Figure 4.8: Prevalence estimates for 2016 with ecozones borders shown in black.

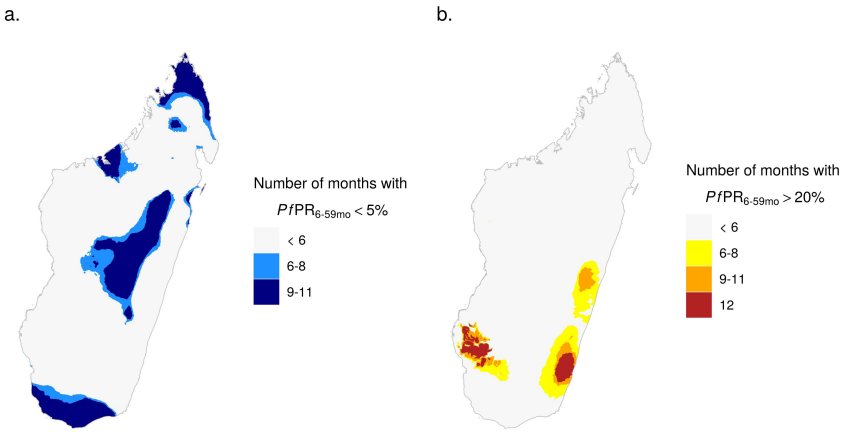


Figure 4.9: Number of months in 2016 with estimated prevalence (a) below 5% and (b) above 20%.

exceeding 0.05. Here we can identify areas in the highlands, south and far north where there is very high confidence of low prevalence. We also see more areas around the north where there is less certainty, with probabilities between 20-50% of prevalence not exceeding 0.05. Maps of uncertainty in 2013, 2014 and 2015 using the interquartile range and exceedance probabilities are included in the appendix (Figures A1-A3).

The sensitivity analysis performed using different catchment populations showed that the prevalence estimates were fairly robust to different methods of estimating these populations. Despite the systematic differences between these estimates, leading to significantly different incidence estimates (Figure 4.10), the resulting prevalence estimates were very similar (Figure 4.11).

4.5 Discussion

These results highlight the benefits of combining routine case data and cross-sectional survey data to provide a more complete understanding of prevalence over time. While the 2013 and 2016 MIS data show similar levels of prevalence in these two years, the routine case data suggests there was substantial variation in transmission between these time points, with decreasing case numbers in 2014 followed by a marked increase in 2015. Incorporating this case data into a formal modelling framework allowed us to quantify these trends in terms of changes in prevalence over time. The robust two-step modelling approach allowed us to learn from these noisy data sources effectively while retaining flexibility in important model components, in particular the incidence-to-prevalence relationship.

By producing monthly risk maps we are able to assess temporal trends of interest in detail. For example, in much of the country the increase in prevalence in 2015 was caused by increased prevalence throughout the year and into early 2016, rather than only higher than average prevalence during the high transmission season. The

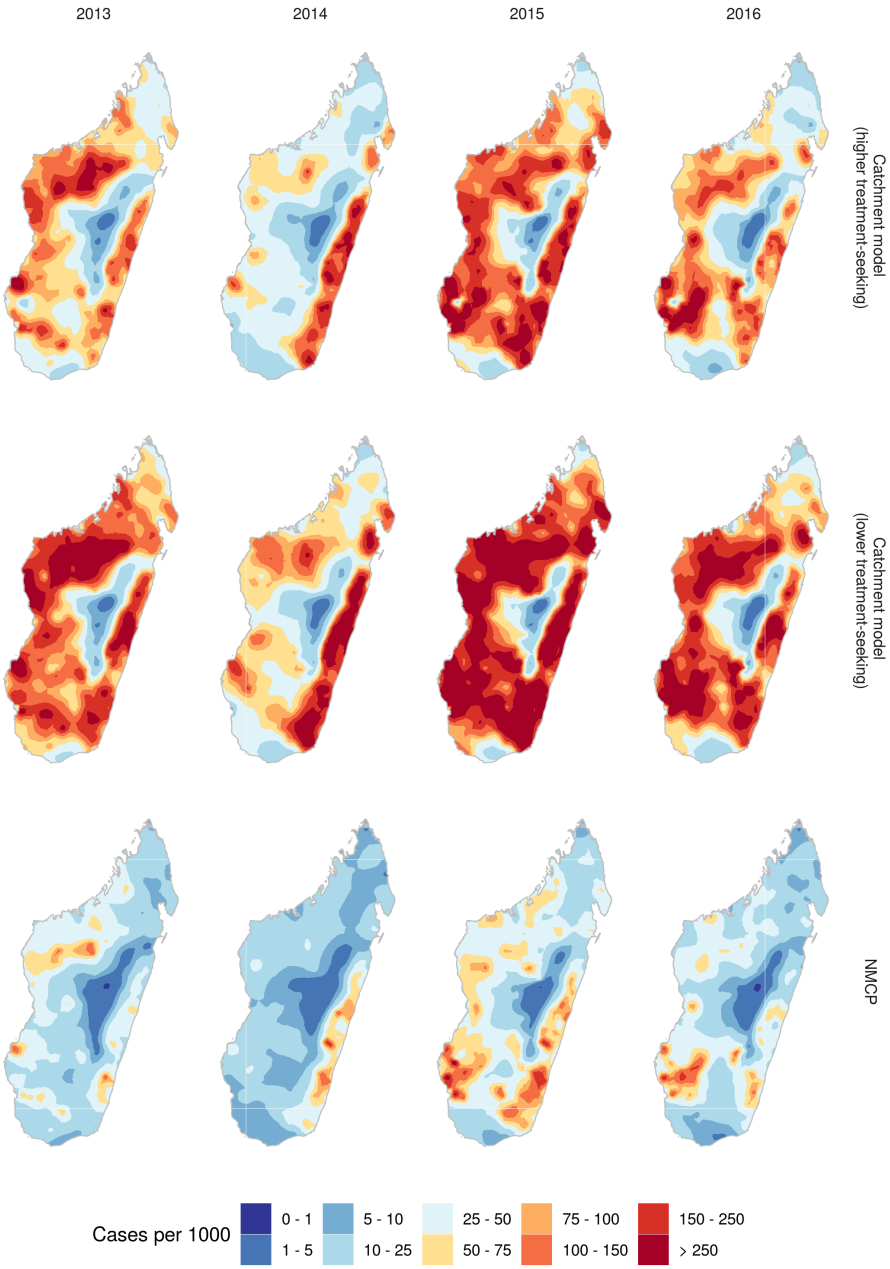


Figure 4.10: Annual incidence surfaces using three alternative catchment population estimates: estimates from the catchment model assuming higher treatment-seeking (top), estimates from the catchment model assuming lower treatment-seeking (middle) and estimates supplied by the NMCP (bottom).

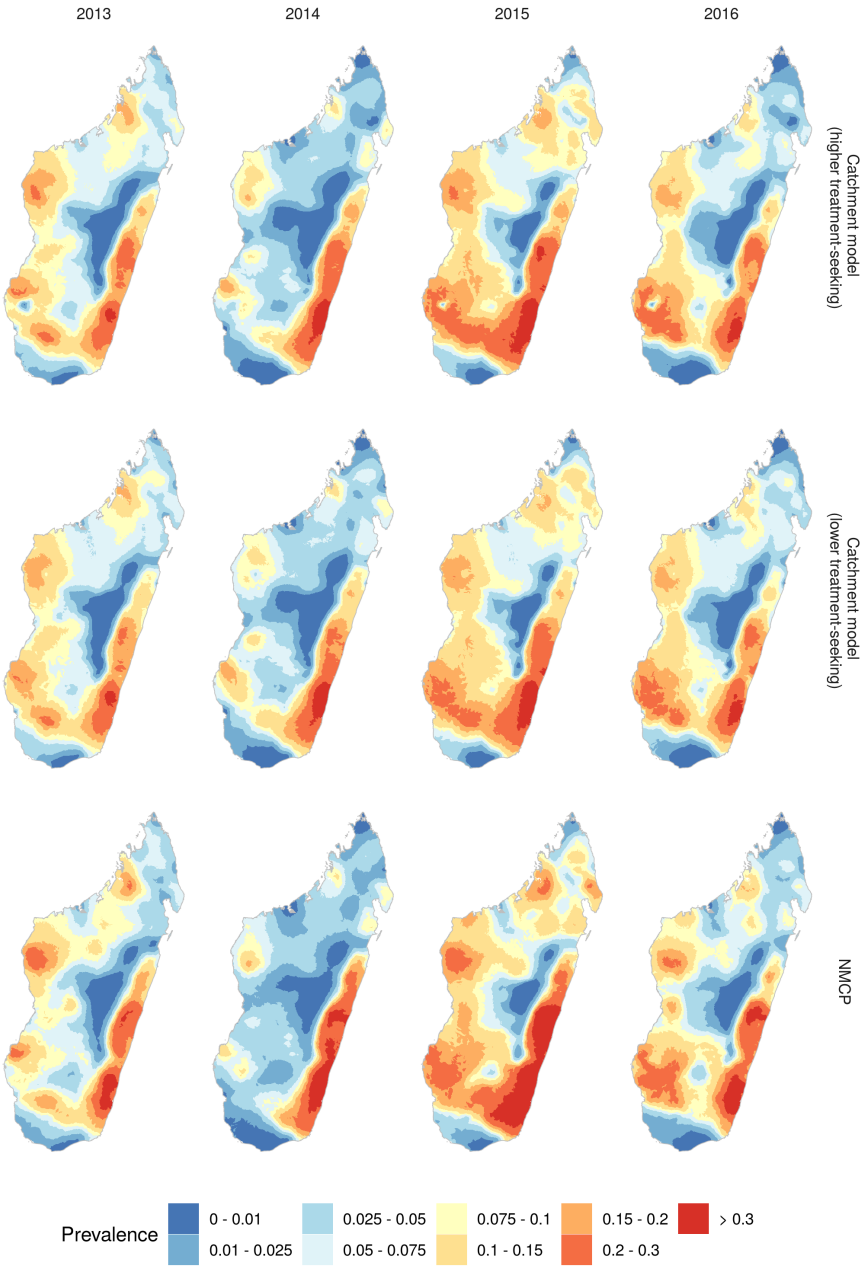


Figure 4.11: Annual prevalence estimates using three alternative catchment population estimates: estimates from the catchment model assuming higher treatment-seeking (top), estimates from the catchment model assuming lower treatment-seeking (middle) and estimates supplied by the NMCP (bottom).

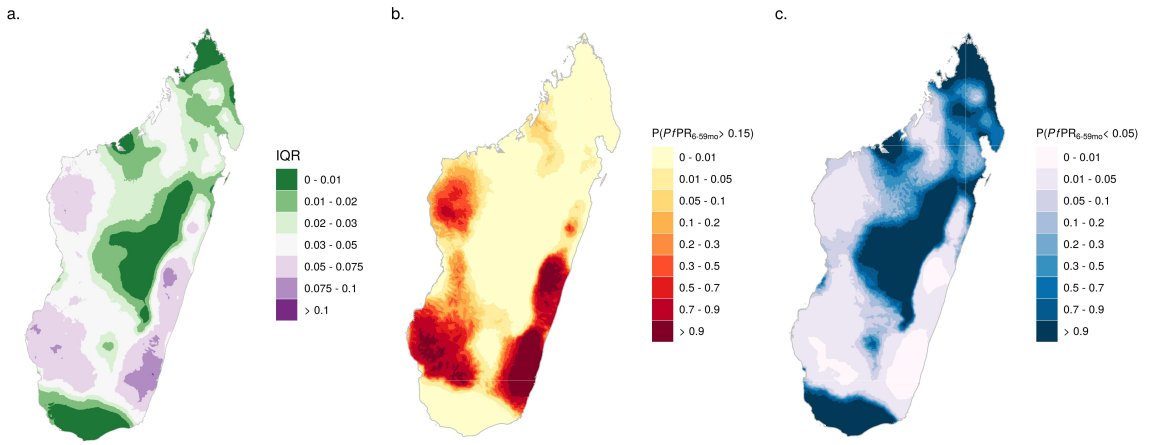


Figure 4.12: Uncertainty in 2016 annual prevalence estimates expressed with (a) interquartile range, (b) probability of prevalence exceeding 0.15 and (c) probability of prevalence not exceeding 0.05.

observed increase in cases in 2015 has been attributed to cyclones and the resulting flooding and lack of supplies in many parts of the country [199]. These results may help confirm this or identify other factors driving these increases in transmission (and similarly the decreases in 2014). We can also distinguish regions where prevalence appears to have returned to normal levels by the second half of 2016 from regions where increased transmission appears to have been sustained, such as the South and Highland fringe west ecozones. The latter may be potential targets for increased control efforts. Monthly prevalence estimates allowed us to evaluate seasonal trends in prevalence. These trends were largely similar across the country, with one seasonal peak occurring between February and April. This seasonality was clear in many of the high transmission areas but was also observed in highland ecozones, despite the low overall prevalence in these regions. A better understanding of baseline transmission patterns in low transmission settings may improve outbreak detection [176] or influence plans for moving towards pre-elimination and elimination in these areas.

Identifying areas of consistently low or high transmission may also influence operational planning. Currently ITN mass distribution campaigns (MDCs) are carried out everywhere except elimination and pre-elimination districts. The criteria for this

stratification are annual incidence and test positivity rates (elimination districts are defined by an incidence rate of less than 1 per 1000 people and pre-elimination by an incidence rate between 1-10 and less than 5% test positivity) [183]. Use of monthly prevalence estimates could help to distinguish areas that have consistently low transmission (as identified in Figure 4.9(a)) from low transmission areas with short but significant seasonal peaks which could benefit from more intervention. Similarly, highlighting areas with consistently high prevalence could inform the control methods used in these areas.

We can also consider the long term trends over the four years in the context of the MDCs which have taken place approximately every three years since 2009 [15]. The second MDC took place in November 2012 on the east coast and October 2013 in the rest of the country and the third from September to December 2015 [200, 201]. Although continuous ITN distribution was carried out to supplement these campaigns, a decline in net coverage and effectiveness over time could be a factor in the overall increase in estimated burden up to the end of 2015.

When comparing my results to the MIS data, the prevalence estimates from the MIS reports of 9.1% in 2013 [178] and 7.0% in 2016 [55] are higher than our annual point estimates (6.4% and 6.6% respectively) but are consistent with our credible intervals. However, our results show limited evidence of a decrease in prevalence in 2016 compared to 2013. It appears likely that the decrease observed in the raw data is a consequence of the timing of the 2016 survey which primarily took place between May and July, around a month later than the 2013 survey and slightly past the seasonal peak in prevalence for the year. The annual 2013 and 2016 risk maps made by Kang et al. [53] show a similar spatial pattern to our results, with low rates of prevalence in the highlands and south of the country and higher rates in the east and west coasts. However, Kang et al. [53] estimate high prevalence in the north of the country, whereas in our results areas of higher prevalence are largely in the northeast.

Similar to our results, their estimates show very similar prevalence in 2013 and 2016, although in general the overall mean estimated is slightly higher (9.3% in 2016) than our point estimates. We can also compare the monthly estimates of prevalence in each ecozone. While in many regions overall seasonal trends were largely similar, the monthly prevalence estimates made by Kang et al. [53] show less consistent seasonal patterns with greater uncertainty. We would expect the monthly estimates in our analysis to be more robust, particularly outside of the survey months (April - June in 2013 and May - July in 2016) and in lower transmission areas, as there is more data to inform each month.

Comparing prevalence estimates to the routine case data, we can see that the temporal patterns in prevalence are typically smoothed trends from the case data (due to the spatial smoothing when producing the incidence surfaces and the temporal smoothing from using incidence at two time points to the predict prevalence). However, we can see distinct spatial trends in the prevalence estimates, for example prevalence was consistently highest in the Southeast ecozone in the whole study period despite similar or higher observed incidence rates in the Southwest ecozone from April 2015 onwards.

As well as visualising uncertainty using the interquartile range, maps of exceedance (and non-exceedance) probabilities were used. In practice, the latter may be more interpretable and therefore more useful for communicating uncertainty. By choosing relevant thresholds, exceedance surfaces may be useful for identifying areas that are estimated to be high transmission with high confidence or areas that are most likely to be considered pre-elimination.

4.5.1 Limitations

A key modelling input in this study was monthly incidence rates at each health facility, which depend on the estimated catchment populations. The model for es-

timating these catchment populations is based only on travel time to each health facility and does not take into account other factors that may influence treatment-seeking or choice of facility, such as type of facility [192, 202]. However, the results of this analysis when using catchment population estimates from the NMCP and when varying treatment-seeking behaviour (Figures 4.10 and 4.11) demonstrate that this modelling approach is fairly robust to differences in catchment population estimates. This is likely due to the spatial smoothing of incidence data and learned incidence-to-prevalence relationship. Treatment-seeking behaviour may also vary throughout the year, which is not accounted for in this analysis and may bias monthly estimates of prevalence. Modelling these temporal dynamics is extremely challenging, with limited treatment-seeking data available and behaviour that is likely to vary on a small spatial scale based on local infrastructure and geography.

A strength of this approach is the ability to account for bias in reporting of case information due to the learned relationship between prevalence and incidence. However, differences in reporting across the country cannot be accounted for in this relationship (which is constant across time and space). Howes et al. [54] investigated the spatial variation in reporting in 2014 by looking at RDT stock-outs and proportion of distributed RDTs for which any result was reported to the HMIS by district. The proportion of RDT results reported was generally higher on the west coast, and therefore prevalence in this region may be somewhat overestimated, but across the rest of the country there were no strong spatial trends in reporting. Similarly, increases in reported cases over time may be partly due to increased access to healthcare or availability of RDTs, leading to overestimates of prevalence in more recent years, although the use of survey data in the first and last years of the study period should mitigate this.

Although the case data used in this study was for a different age range (all ages) to the estimates of prevalence (6-59 months), I believe this is unlikely to bias these

estimates. There was a strong relationship between the number of cases in individuals under 5 and all ages in the routine case data (correlation of 0.91 across all health facilities and months) and the relationship between incidence and prevalence was learned within the model. Therefore I would expect the predictive power of incidence in either age range to be similar. Incidence for all ages was chosen due to the larger sample sizes in the data, which should produce more reliable estimates of incidence.

The aim of this study was to estimate prevalence, however estimates of incidence may be of more use for public health purposes (for example, district-level risk stratification by the NMCP is based on estimated incidence [183]). Future work could therefore focus on mapping incidence. A common approach for generating incidence estimates is to use prevalence estimates and an established prevalence-incidence relationship [26, 27]. This approach is generally used where prevalence estimates are informed only by prevalence survey data (where routine case data is unavailable or too unreliable to be used) but a similar transformation could be applied to the prevalence estimates produced here. A more direct way of combining routine case data and survey data to estimate incidence would be to use a joint model [186]. To do so successfully would likely require a more complete understanding of the spatiotemporal trends in treatment-seeking and reporting completeness in order to account for bias in the routine case data due to these factors.

4.6 Conclusion

In this work, routine case data and survey data was used to produce monthly estimates of malaria prevalence between 2013 and 2016. The results suggest that while malaria endemicity was similar in 2013 and 2016, there was considerable variation in the intervening period, with a small decrease in 2014 followed by a substantial increase in 2015. In many areas, this increase in prevalence was sustained throughout

2015 and early 2016, and in the Northwest and South ecozones prevalence had yet to return to 2014 levels by the end of 2016. Considering these temporal trends at specific locations in relation to the timing of control efforts (such as ITN distribution) and climatic events may help to identify drivers of transmission and assess intervention effectiveness. Areas where prevalence remains higher than pre-2015 levels could be targets for increased control measures. Seasonality of transmission (with a single peak around March) was widely observed in high transmission areas and in some low transmission areas. The relative spatial patterns of prevalence were mostly consistent over time, with highest prevalence in the southwest and high prevalence in the southeast. Prevalence was consistently low in the highlands and in parts of the south of the country.

The Bayesian modelling approach applied here allowed us to make prevalence estimates with associated measures of uncertainty by learning temporal trends in transmission from the routine case data and calibrating these trends to prevalence observations from the survey data. These risk estimates could be an important tool for assessing the impact of control measures and the progress made towards the goals set out by the NMCP [15, 16] and for better understanding the drivers of changes in transmission. The results demonstrate the considerable additional information that can be gained by combining data sources in this way. While previous risk mapping efforts produced monthly prevalence maps for 2011, 2013 and 2016 [53], these were informed by a relatively small number of prevalence observations from surveys which only covered 3 months of each year. We were able to make monthly maps of prevalence over a four year period (including years in which no surveys took place), informed by a large amount of monthly case data in addition to the survey data. By using a two-step modelling approach, rather than a joint likelihood model, we were able to learn the relationship between incidence and prevalence within the model (thereby accounting for systematic under reporting in the case data) and ensure that the model

was informed by the more reliable prevalence data, despite the relatively small number of observations compared to the case data. The sensitivity analysis demonstrated that these estimates are robust to varying assumptions about treatment-seeking behaviour and reporting incompleteness. A similar modelling strategy may be effective in other multi-metric disease modelling settings, especially where there is an imbalance between the number of observations or reliability of different metrics.

Incomplete reporting, treatment-seeking behaviour, and inconsistent standardisation of clinical diagnoses affect the quality of routine case data in many countries of sub-Saharan Africa [17]. Consequently, the World Health Organization estimates for malaria burden are often based on community prevalence surveys alone [57, 96]. The methods presented here may be a useful starting point for incorporating routine case data into estimates of malaria burden more widely.

Chapter 5

Causal feature selection for forecasting incidence in Madagascar

In the previous chapter, I applied my proposed causal discovery procedure as a feature selection method when mapping malaria prevalence with point-level response data. In the work described in this chapter, I again use this causal feature selection method, this time in the context of predicting malaria incidence with area-aggregated response data. Importantly, I also compare the performance of the model forecasts when using causal feature selection to two alternatives—using a classical feature selection method (spike-and-slab regression) and not using any feature selection (i.e., instead including all available features in the model). The response data used was the routine surveillance data from Madagascar introduced in Chapter 4 (where it was instead used as an explanatory factor when predicting prevalence) and the forecasting task was to learn on 12 months of incidence data from a subset of health facilities and then predict the incidence rate at all facilities over the next 24 months.

Compared to the prevalence mapping task in the previous chapter, this was a more suitable problem for benchmarking the performance of these selection strategies. The prevalence estimates were mainly informed by the smoothed incidence surfaces, with the environmental and socioeconomic covariates included in the final model having a relatively small effect; hence, the manner in which the environmental and socioeco-

nomic covariates were selected is unlikely to have had a large effect on the predictive accuracy. In contrast, when forecasting incidence here, prevalence was not included as an explanatory variable due to insufficient temporal coverage. Moreover, since the task is to predict forwards in time, estimates could not be based on spatial interpolation of nearby observations at the same point in time. Indeed, in the model used in this chapter the dynamic component of the predictions was designed solely as a function of the explanatory variables. Therefore, the environmental and socioeconomic covariates included in the model, and hence the selection method used, had a large effect on the final estimates.

The incidence model examined in this chapter is implemented with two variations of its regression structure, one assuming a linear effect of each covariate and the other allowing for non-linear effects. I begin this chapter by detailing the incidence model and then the two feature selection methods used. I then present and discuss the results of the analysis. Parts of this work were published as a preprint, co-authored with my supervisors Ewan Cameron and Peter Gething [1].

5.1 Methods

5.1.1 Modelling incidence

In this section, I describe the geostatistical model used to forecast incidence. The feature selection procedures (causal selection and thresholded spike-and-slab regression) used to choose the features included in this model are detailed in the subsequent sections. The observed incidence data was the number of confirmed cases of malaria at each health facility each month. This can be thought of as 'aggregated' data, in the sense that the cases observed at a given health facility are drawn from the surrounding area, rather than originating in the same location as the health facility. Incidence was therefore modelled using disaggregation regression [94, 203], a method designed for

use with aggregated response data, in which the case generating process is modelled on a fine grid of pixels (introduced in Section 2.4). The likelihood of the observed data is then given by a weighted sum of these processes over the modelled catchment area of each facility. As in Chapter 4, another important factor to consider was that a large proportion of the population may not seek treatment for fever within the formal healthcare system and therefore the cases observed at health facilities represent only a certain fraction of all malaria cases. Accounting for this treatment-seeking behaviour allows for relationships between covariates and incidence to be learned more reliably and for predictions of true incidence to be made at the pixel level.

5.1.1.1 Catchment model

Adjustments for treatment-seeking behaviour were made based on distance to the nearest health facility, as described in Section 4.3.2, to produce treatment-seeking adjusted population estimates for each pixel. The catchment model used was also very similar to the one described in Section 4.3.2, but in this case health facility attractiveness weights were included. For completeness, I describe this catchment model in full here. The model was based on the friction surface-derived travel time estimates calculated as described in Section 4.3.2. Of individuals in pixel i who seek treatment, the proportion seeking treatment at health facility j, denoted $p(\text{pixel}_i \rightarrow \text{HF}_j)$, was assumed to be zero if the travel time to heath facility j was more than 200 minutes, and otherwise modelled as proportional to the health facility attractiveness (a per-facility parameter) divided by the square of the travel time to that health facility. That is, defining

$$\tilde{p}(\text{pixel}_i \rightarrow \text{HF}_j) := \begin{cases} t(\text{pixel}_i \rightarrow \text{HF}_j)^{-2} w_j & \text{if } t(\text{pixel}_i \rightarrow \text{HF}_j) \leq 200 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where $t(\text{pixel}_i \rightarrow \text{HF}_j)$ was the travel time from pixel i to health facility j in minutes and w_j was the attractiveness of health facility j, the proportion seeking treatment at facility j was

$$p(\text{pixel}_i \rightarrow \text{HF}_j) = \frac{\tilde{p}(\text{pixel}_i \rightarrow \text{HF}_j)}{\sum_{k=1}^{N_{\text{HF}}} \tilde{p}(\text{pixel}_i \rightarrow \text{HF}_k)} \tag{5.2}$$

where N_{HF} was the total number of health facilities. The attractiveness parameter for each health facility allows for the impact of facility-specific latent factors that may affect treatment-seeking preferences. These weightings were learned jointly with the incidence surface within the hierarchical Bayesian structure of the disaggregation regression model.

5.1.1.2 Disaggregation regression model

The number of cases occurring in pixel i in month t that would be seen in the health-care system, Y_{it}, was modelled as a Poisson random variable,

$$Y_{it} \sim \text{Poisson}(\lambda_{it} \times \text{pop}_i)$$

where λ_{it} was the latent incidence rate per person in that pixel and month, and pop_i was the population in pixel i adjusted for treatment-seeking behaviour. The log incidence rate was modelled as the sum of the effect of environmental and socioeconomic covariates and a spatial Gaussian process,

$$\log \lambda_{it} = f(X_{it}) + \text{GP}(s_i)$$

where X_{it} were covariate values (in pixel i at time t), GP was the Gaussian process and s_i was the location of pixel i. The two functional forms of f that were considered are described below. A Mátern kernel was used for the Gaussian process, parameterised by the range, ρ, and marginal standard deviation, σ, which were learned. The number

of cases observed at health facility j in month t, Y_{jt}^{HF}, was then a weighted sum of cases in each pixel, with the weights from the catchment model,

$$Y_{jt}^{\text{HF}} = \sum_{i=1}^{N_{\text{pixel}}} p(\text{pixel}_i \to \text{HF}_j)\, Y_{it}$$

where N_{pixel} was the total number of pixels. In other words, the number of cases observed at health facility j was the sum over all pixels of the proportion of cases occurring in each pixel that would seek treatment at health facility j. Note that for any given facility, the weights from many pixels were zero. Assuming that, conditional on the underlying rate λ_{it}, the number of cases in each pixel was independent, this sum also follows a Poisson distribution,

$$Y_{jt}^{\text{HF}} \sim \text{Poisson}\left(\sum_{i=1}^{N_{\text{pixel}}} p(\text{pixel}_i \to \text{HF}_j) \times \lambda_{it} \times \text{pop}_i\right)$$

which gives the likelihood of the response data. Two alternative functional forms were explored for $f(\cdot)$, the additive impact of covariates on the log incidence rate, as follows.

1. *Linear model.* The covariates were assumed to have a linear effect on log incidence,

   $$f(X_{it}) = \beta_0 + \beta^T X_{it}$$

   where β_0, β were parameters to be learned.

2. *Gaussian process model.* Non-linear effects of covariates were considered by modelling the effect of each with a univariate Gaussian process,

   $$f(X_{it}) = \beta_0 + \sum_{l=1}^{N_{\text{cov}}} g_l((X_{it})_l)$$

   where N_{cov} was the number of covariates and g_l was a Gaussian process over $\mathbb{R}$

for each $l = 1, ..., N_{\text{cov}}$. These covariate space Gaussian processes were assigned a radial basis function kernel (with unit marginal variance) to structurally promote smoothness,

$$k_l(x, y) = \exp\left(-\frac{(x-y)^2}{2\lambda^2}\right)$$

where the parameter λ was common across all processes and was learned during model fitting.

In each case, the model was completed by assigning prior distributions to the various parameters. A penalised complexity prior was used for the range and standard deviation of the Mátern kernel of the spatial Gaussian process [64], chosen such that

$$p(\rho < 1) = 0.01, \text{ and}$$
$$p(\sigma > 0.5) = 0.01.$$

The attractiveness weights of the health facilities were given independent log-normal priors,

$$\log w_j \sim \text{N}(0, 0.25^2)$$

which placed most of the density on values of w_j between $1/\sqrt{3}$ and $\sqrt{3}$, therefore assuming a-priori that it was unlikely for any health facility to be more than 3 times as attractive as any other. The intercept term, β_0, was given a normal prior with a mean 0 and standard deviation 2. In the linear model the entries of the vector β were given independent normal priors with mean 0 and standard deviation 1 and in the Gaussian process model a log-normal prior was applied to λ with $\log \lambda \sim \text{N}(1.5, 0.1^2)$. These priors were chosen to produce prior predictive incidence rates that were on a similar scale to those observed.

5.1.1.3 Forward temporal estimation task

The power of the incidence model (for both linear and Gaussian process variants) to forecast malaria incidence was assessed under three scenarios: (i) using all available features; (ii) using only those selected by the causal inference procedure, and (iii) using only those selected using thresholded spike-and-slab regression. In each case, the incidence model (as described above) was fit using one year of data from training locations (a subsample of all health facilities) and then forward estimates were made for the monthly incidence rate over the following two years at all health facilities. Since the model-based estimates for each month depend on covariates from the same or recent months I do not consider these outputs to be, in a strict sense, forecasts (*cf.* Friedman et al. [204]). However, the ability to provide forward estimates of incidence to months beyond those for which case data are already available is nevertheless of value to malaria control programs. Covariate data are often ready for use many months earlier than case data, for which a lead time is incurred through the stages of collection, processing, validation, and aggregation in many malaria endemic settings.

The number of training locations was varied, as was the starting month of the three year time period studied (which are referred to as 'iterations'). For example, in the first iteration the model was fit to data from January 2013 to December 2013 and forward estimates were made from January 2014 to December 2015, and in the second iteration the model was fit using data from February 2013 to January 2014 and forward estimates were made from February 2014 to January 2016. With four years of data in total, there were therefore 13 iterations.

The main metric of performance used was correlation between observed and predicted incidence rates. Correlation is often the preferred metric when evaluating malaria risk mapping, as relative spatial and temporal trends are of most interest to program managers. For example, the allocation priority of limited resources or targeting of interventions may be based on the relative estimated risk of different regions.

A model which accurately distinguishes lower and higher risk areas can therefore be of more use than a model which has smaller absolute error but is less accurate at establishing relative rankings. I considered overall correlation between predicted and observed rates (over all locations and months) and 'temporal correlation', which was defined as the correlation between the predicted and observed time series (over the 24 months forecasted) at each location, averaged over all locations. Accurately estimating temporal patterns in transmission is useful for interventions that are based around high transmission periods (such as seasonal malaria chemoprevention and indoor residual spraying). I also considered this temporal correlation limited to only the second year of predictions (months 13-24) in order to investigate model performance further into the future. Predictions were also assessed in terms of root mean square error.

5.1.2 Causal selection

Causal feature selection was performed by applying the causal discovery procedure, as described in Section 3.5, to all features simultaneously (in contrast to the analysis in Chapter 4, where static and dynamic variables were considered separately). Again here the term 'feature' refers to either a static variable or a dynamic variable at a specific time lag. Unlike in Chapter 4, cross-validation was not used to select the final feature set. Instead, features were ranked by the certainty of the edge $X \to I$, where X was the feature of interest and I was the variable representing incidence. The final feature set was then the top four static and top four dynamic variables. To achieve a fair comparison with thresholded spike-and-slab regression the decision was made to select the top four static and top four dynamic features in each case.

When applying the causal discovery procedure, the aggregated nature of the incidence data required a different prewhitening approach than was used in Chapter 4. It also required the use of the modified independence tests introduced in Section 3.3.6.

These steps are described in more detail here.

5.1.2.1 Prewhitening

The first step of the causal feature selection procedure is a prewhitening step to remove large-scale spatial and temporal trends from all variables. The model used for the incidence data was a spatiotemporal version of the model described in Section 5.1.1 which used no covariates. In this version, the incidence rate in pixel i at time t was modelled as,

$$\log \lambda_{it} = \beta_0 + \mathrm{GP}(s_i, t)$$

where β_0 was a parameter to be learned, s_i was the location of pixel i, and GP was a separable Gaussian process with Mátern structure in the spatial component and an AR(1) structure over time. The rest of the model was the same as described in Section 5.1.1. Health facility attractiveness weights were learned as part of this model and note that the resulting proportions of the population in each pixel seeking treatment at each facility were used in the aggregated independence tests (see below). The residuals for each health facility were then calculated by subtracting the rate in the fitted model from the observed rate. These residuals were the values representing incidence that were used in the causal discovery algorithm. In order to enforce that this prewhitening model learned broad spatial patterns, the range and scale parameters of the Mátern covariance were given log-normal priors with $\log \rho \sim \mathrm{N}(2, 0.1^2)$ and $\log \sigma \sim \mathrm{N}(-2, 0.1^2)$.

To prewhiten the environmental and socioeconomic covariates, these variables were fit with a spatiotemporal Gaussian process of the same form directly. For a temporal covariate, the expected value in pixel i at time t, Y_{it}, was modelled as,

$$Y_{it} \sim \mathrm{N}(\beta_0 + \mathrm{GP}(s_i, t),\, \sigma^2)$$

where again the Gaussian process was the product of a Mátern process over space and AR(1) process over time. For a spatial covariate (such as elevation) the value in pixel i, Y_i, was modelled as,

$$Y_i \sim \mathrm{N}(\beta_0 + \mathrm{GP}(s_i),\, \sigma^2)$$

with a Mátern covariance. In both cases, the same log-normal priors were used as in the prewhitening model for incidence. For both the prewhitening of the incidence and covariates, I erred on the side of caution with these priors, opting to ensure that I did not remove too much of the signal from the original data at this step. Therefore it is possible that less spatiotemporal structure was removed than in the ideal prewhitening step. I would note here again that residual spatiotemporal structure is expected to (on average) result in the causal discovery algorithm incorrectly retaining edges, which is less problematic for feature selection than incorrectly removing edges.

5.1.2.2 Aggregated independence tests

As in the disaggregation regression model, when performing conditional independence tests involving incidence, each incidence observation at a health facility was modelled as an aggregation of cases occurring in nearby pixels. To perform independence tests involving this aggregated incidence variable and pixel-level covariates, I implemented an aggregated version of the RCIT as described in Section 3.3.6. The prewhitened incidence rate in health facility j and time t, r_{jt}, was associated with prewhitened covariate values at time t from all pixels, weighted based on the probability of seeking treatment at that facility calculated using the catchment model. More precisely, let x_{it} be the value of a covariate in pixel i at time t. Then the rate r_{jt} was associated with the set of values of this covariate in each pixel, $\{x_{it}\}_{i=1}^{N_{\text{pixel}}}$, where the value in

pixel i was weighted by

$$p(\text{pixel}_i \rightarrow \text{HF}_j) \times \text{pop}_i$$

and N_{pixel} was the total number of pixels. The weights w_j that were needed to calculate $p(\text{pixel}_i \rightarrow \text{HF}_j)$, as described in (5.1) and (5.2), were learned in the prewhitening step. Note that for many pixels the value of $p(\text{pixel}_i \rightarrow \text{HF}_j)$ will be zero so in practice not every covariate value is associated with every health facility. In other words, each health facility is associated with covariate values in the pixels where some proportion of population would seek treatment at that facility, with these values weighted by the proportion of the total catchment population expected to come from each pixel.

5.1.3 Spike-and-slab regression

Spike-and-slab regression is a sparsity-promoting Bayesian modelling technique originally designed for linear regression [205, 206, 207, 208]. The name refers to the characteristic mixture model prior applied to the regression coefficients, made up of a point mass at zero (the 'spike') and a wide, flat distribution (the 'slab'). Spike-and-slab models are advantageous for variable selection as they result in selective shrinkage, shrinking small coefficients towards zero while leaving large coefficients relatively unchanged, in a similar way to the LASSO in the frequentist context [209]. In this implementation I used a continuous bimodal prior for the variance of the prior on the coefficients, rather than a mixture prior on the coefficients, to achieve the same effect, as suggested by Ishwaran and Rao [208]. In the case of the linear model (as defined in Section 5.1.1.2), the only difference between the final regression model and the spike-and-slab model was the prior placed on the regression coefficients. The

prior hierarchy used was as formulated by Ishwaran and Rao [208],

$$\begin{aligned}
\beta_l|\phi_l, \tau_l^2 &\overset{\text{ind}}{\sim} \mathrm{N}(0, \phi_l\tau_l^2) \qquad\qquad l = 1, ..., N_{\text{cov}} \\
\phi_l|\nu_0, w &\overset{\text{iid}}{\sim} (1-w)\delta_{\nu_0}(\cdot) + w\delta_1(\cdot) \\
\tau_l^{-2}|a_1, a_2 &\overset{\text{iid}}{\sim} \mathrm{Gamma}(a_1, a_2) \\
w &\sim \mathrm{Uniform}[0, 1]
\end{aligned}$$

with $a_1 = 40, a_2 = 0.025$ and $\nu_0 = 0.01$ to produce a bimodal prior for $\phi_l\tau_l^2$. For the Gaussian process model, a modified form of f was used,

$$f(X_{it}) = \beta_0 + \sum_{l=0}^{N_{\text{cov}}} \beta_l \mathrm{GP}_l((X_{it})_l)$$

with the same prior hierarchy described above. The posterior was sampled using a Metropolis–Hastings algorithm (a form of Markov chain Monte Carlo), with four chains for each repeat with a burn-in of 500 and 3000 steps, to obtain inclusion probabilities for each variable. In a strict sense the spike-and-slab model does not produce a hard selection of variables unless an *a posteriori* processing is applied to summarise the posterior in this manner. In settings where there are many features to choose from it has been shown that selection by marginal inclusion probability threshold can outperform selection by highest posterior mass [210]. For optimal comparability with the causal feature selection procedure, the thresholded spike-and-slab selection was defined as the restriction to the four static and four dynamic covariates with highest marginal inclusion probabilities respectively. I term this procedure 'spike-and-slab selection'.

5.1.4 Data

Monthly malaria case data between January 2013 and December 2016 were available from the NMCP, as described in Section 4.3.1. These data represent individuals who seek care at a health facility where a diagnosis of malaria is made and where that case is then recorded and entered into the national surveillance system. Such data are termed 'passive' case detection and rarely capture all malaria incidence in the community, as some malarious individuals might not seek care or may do so only from informal or private providers. These data do not differentiate by the species of malaria parasite causing the infection. Data were available from 3342 health facilities in Madagascar, of which 2801 were geolocated (NMCP, pers comm) and used for the present study. The location of these facilities is shown in Figure 5.1. The facilities that could not be geolocated are believed to be smaller basic health centres and these 541 facilities accounted for only 5.4% of total cases observed at all facilities over the four year period. Therefore the vast majority of the reported cases were included in our analysis. While the precise location of these facilities was not known, the district in which they were located was. The distribution of these facilities that could not be located precisely is shown in Figure 5.2. While there was heterogeneity in the proportion of facilities that were not geolocated in each district, there was no clear spatial trend.

The spatial data products gathered to assist with probabilistic interpolation and forecasting of the incidence rates reflected in the routine case data are summarised in Table 4.1. One additional dynamic covariate was included, a temperature suitability index for *P. vivax* [211]. When modelling prevalence in the previous chapter, only the temperature suitability index for *P. falciparum* was used, as this is the dominant parasite in the country and the two indices are largely similar. Both indices were included here to investigate whether the feature selection methods would preferentially select one or the other. Dynamic variables were included at 0, 1, 2, and 3 month time

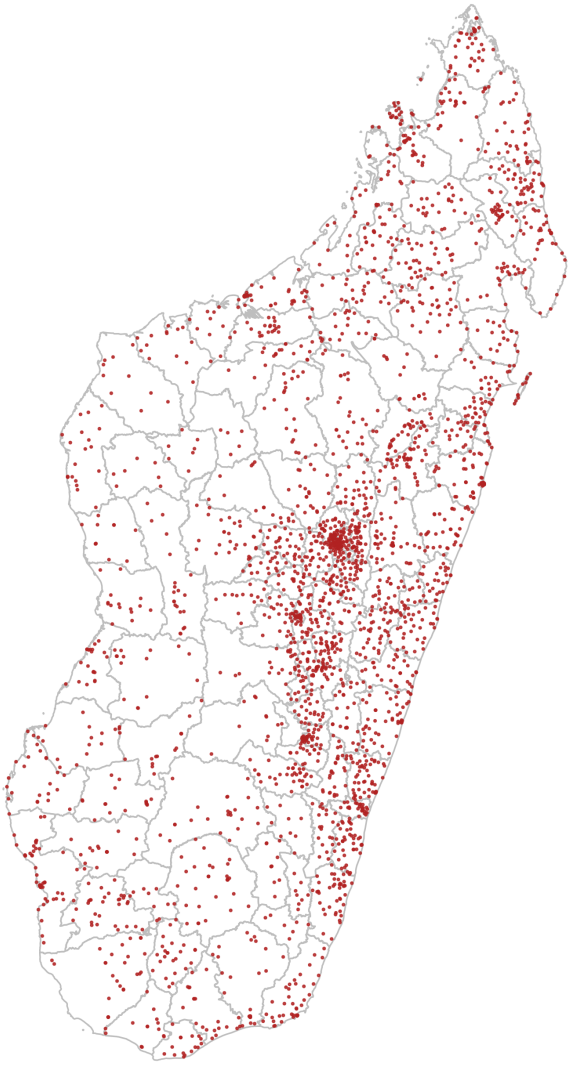


Figure 5.1: Geolocated health facilities.

lags to allow covariate selection to identify the optimal lag (or lags) to include in the final model. As we have discussed, all of these variables are commonly used in malaria risk mapping and have putative causal connections to malaria incidence, however the relative importance of these factors in any given location and (for dynamic variables) the exact time scales on which they act is not necessarily known (see e.g. Weiss et al. [21]). In total there were 36 features available—8 static and 28 dynamic (each of the 7 dynamic variables at 4 different time lags). In the rest of this chapter, I continue to

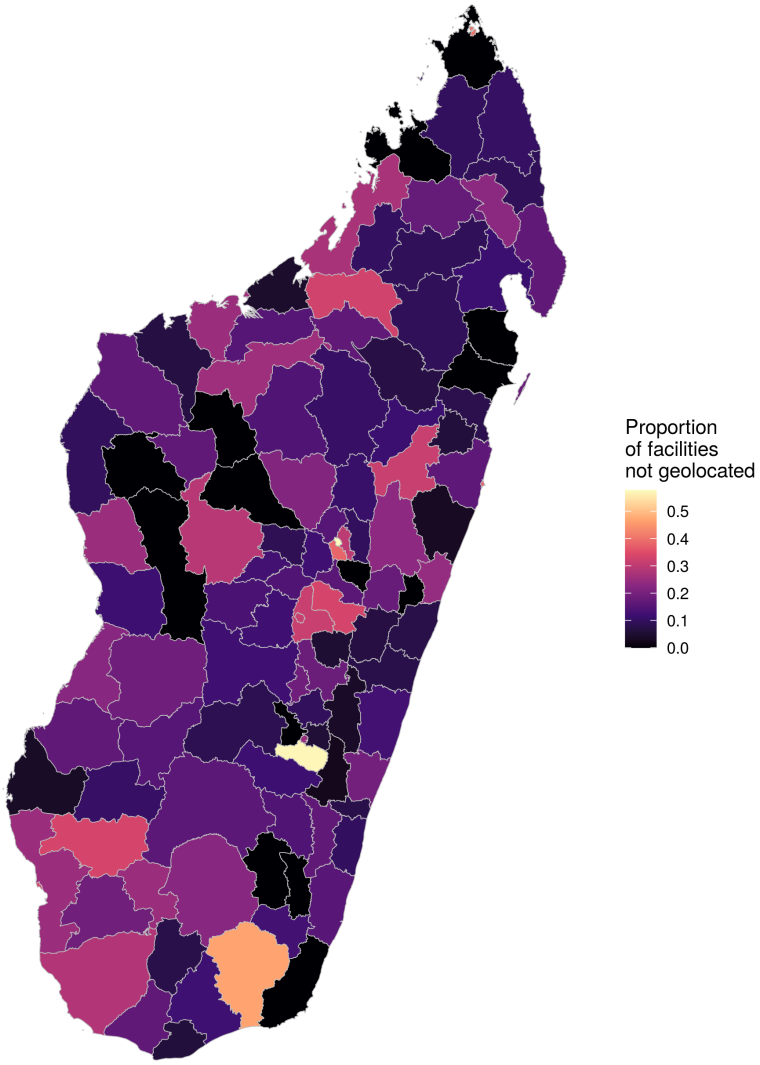


Figure 5.2: Proportion of health facilities in each district that could not be precisely geolocated.

use 'feature' to mean either one of the static variables or one of the dynamic variables at a specific time lag (for example, I refer to rainfall as a variable and rainfall at a 1 month time lag as a feature).

5.2 Results

5.2.1 Feature importance

The importance rankings of features returned by the causal inference approach, and by the linear and Gaussian process spike-and-slab models, are presented in Figures 5.3, 5.4 and 5.5, respectively. A total of thirteen are shown for each, corresponding to the thirteen overlapping windows of data created for testing purposes by changing the starting month (referred to here as 'iterations'). In the case of the causal inference approach, the importance of a feature was quantified by the proportion of outputs in which this feature was a direct cause of incidence from repeated applications of the PC algorithm, as described in Section 5.1.2. For each of the spike-and-slab models, on the other hand, the importance of a feature was quantified by its marginal posterior inclusion probability. Since the causal feature selection procedure is model-free, these rankings were used to select variables for both the linear and Gaussian process models applied subsequently; whereas feature selection under the spike-and-slab approach is intrinsically tied to the corresponding regression model.

A comparison of Figure 5.3 against Figures 5.4 and 5.5 reveals a striking contrast between these approaches: feature selection under the causal method is highly stable across the iterations, while spike-and-slab selection is highly unstable. In the former, three static features (elevation, aridity and PET) were consistently high ranked, as were all lags of the LST night and TSI (both *P. falciparum* and *P. vivax* versions) dynamic features. In the linear spike-and-slab model, four static features (accessibility, aridity, PET and slope) were somewhat consistently high ranked, while the ranking of all dynamic features varied greatly between iterations. In the Gaussian process spike-and-slab model, all static covariates were somewhat consistently high ranked, and again the dynamic features showed great variation. The between-iteration correlation coefficients for the rankings returned by each approach were 0.87 (causal), 0.13

(linear spike-and-slab) and 0.21 (Gaussian process spike-and-slab).

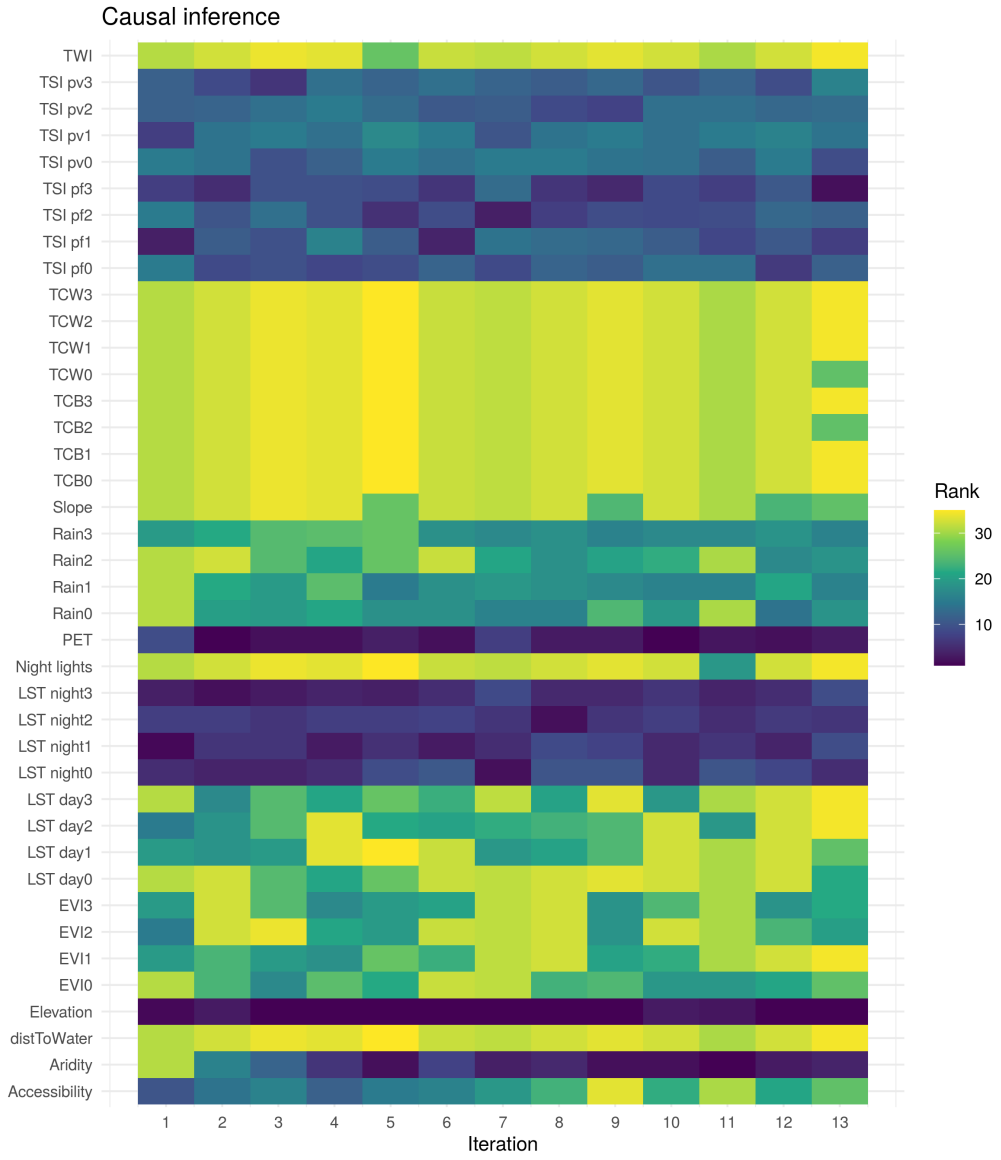


Figure 5.3: Feature rankings by importance (quantified by the proportion of repeats in which that variable was a direct cause of incidence) according to the causal inference approach, across each of the thirteen data iterations.

5.2.2 Model performance

The correlation between true and predicted incidence rates (across all health facilities and months) when using causal and spike-and-slab selection is shown in Figure 5.6.

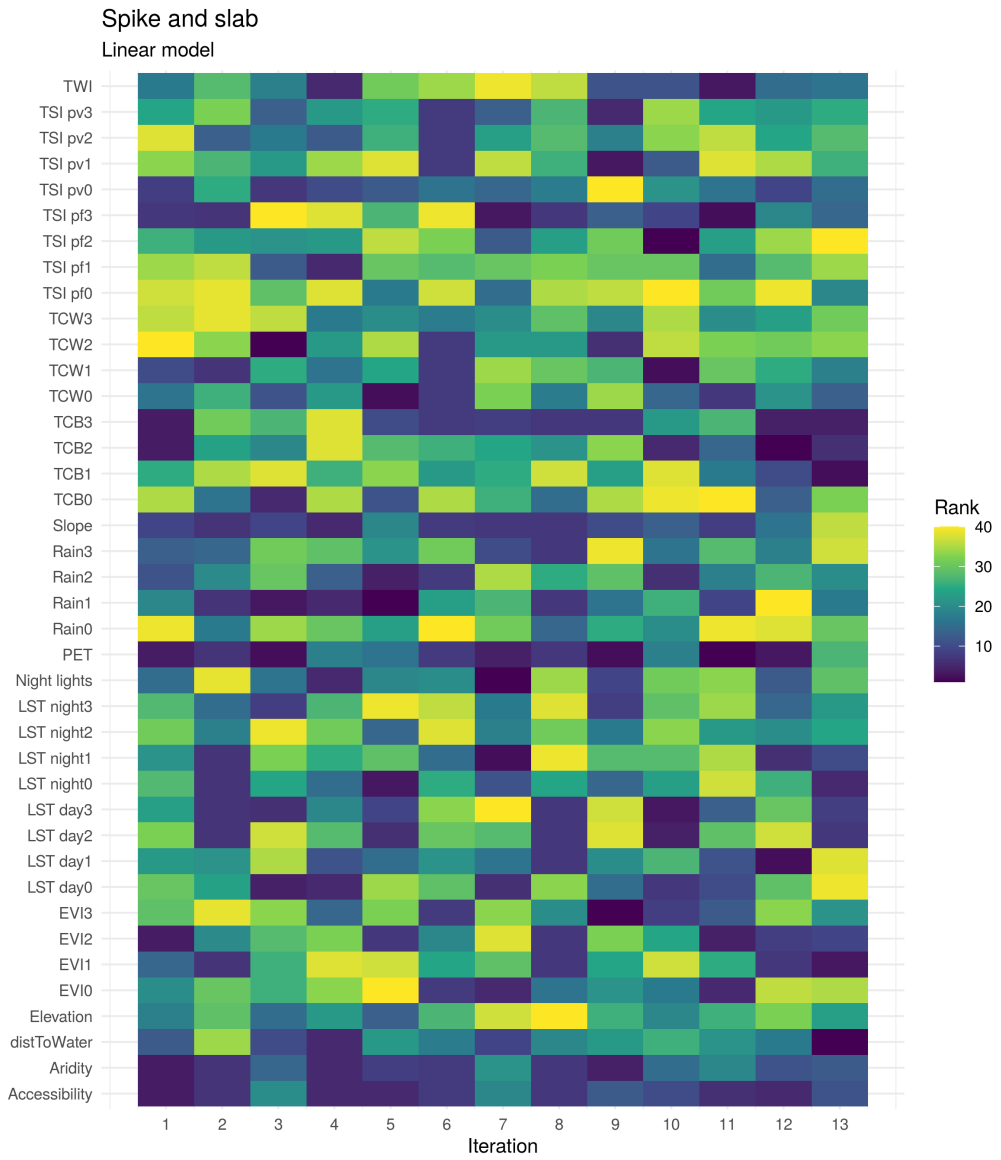


Figure 5.4: Feature rankings by importance (quantified by the marginal posterior inclusion probability) according to the linear spike-and-slab model, across each of the thirteen data iterations.

For the linear model, overall there was no clear advantage for either selection method but there is a trend of better performance with causal selection for smaller training sets and with spike-and-slab selection for larger training sets. For the Gaussian process model, however, performance was superior with causal selection in almost all cases except when using the smallest training sets. The model performance according to this metric under the alternative of no feature selection is compared with that un-

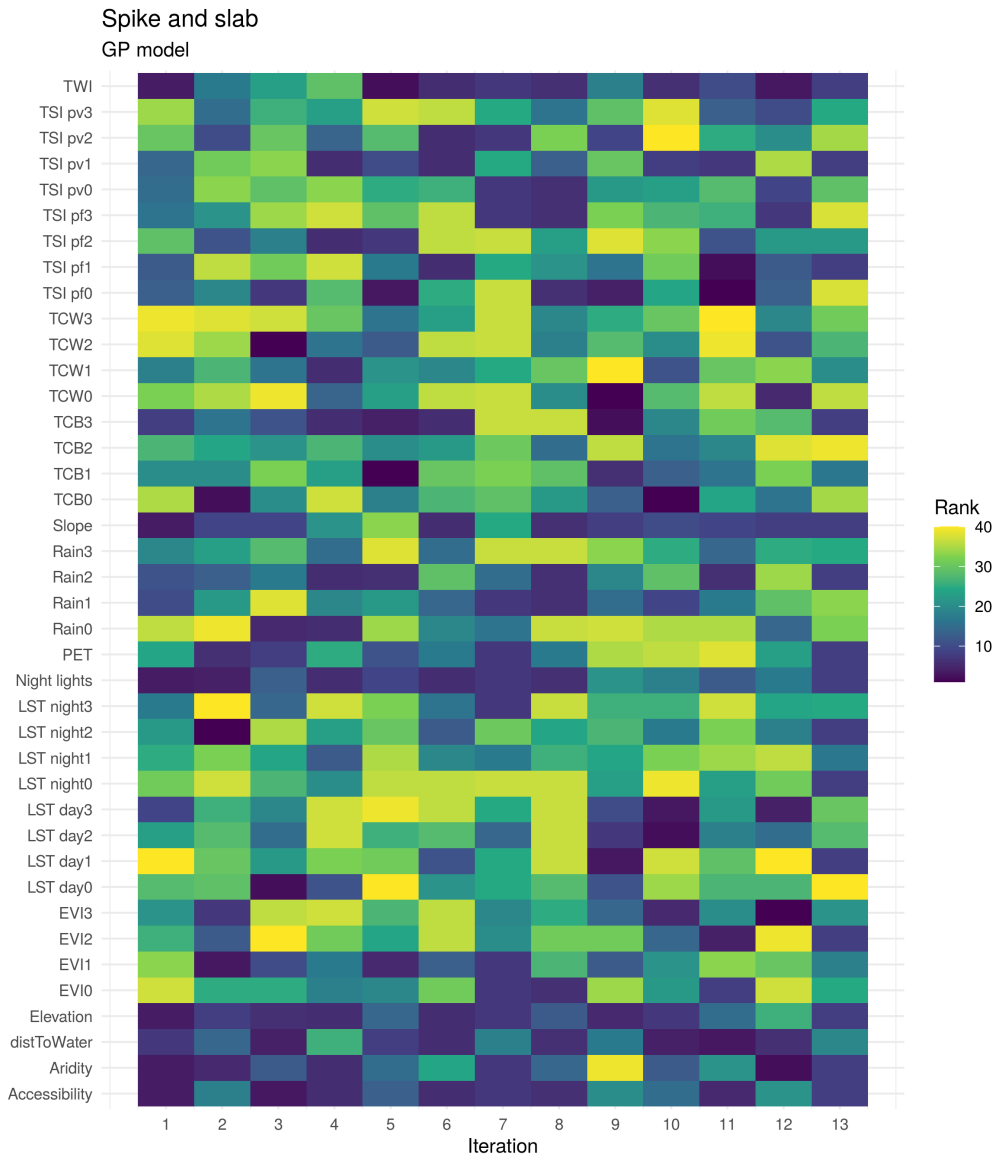


Figure 5.5: Feature rankings by importance (quantified by the marginal posterior inclusion probability) according to the Gaussian process spike-and-slab model, across each of the thirteen data iterations.

der causal feature selection in Figure 5.7. The linear model usually performed better with no selection, except for the smallest size of training set where correlation was consistently higher with causal selection. For the Gaussian process model, performance was improved when using causal selection in the large majority of cases, with this effect being clearest for smaller training sets.

Figure 5.8 shows the comparison between model performance in terms of temporal

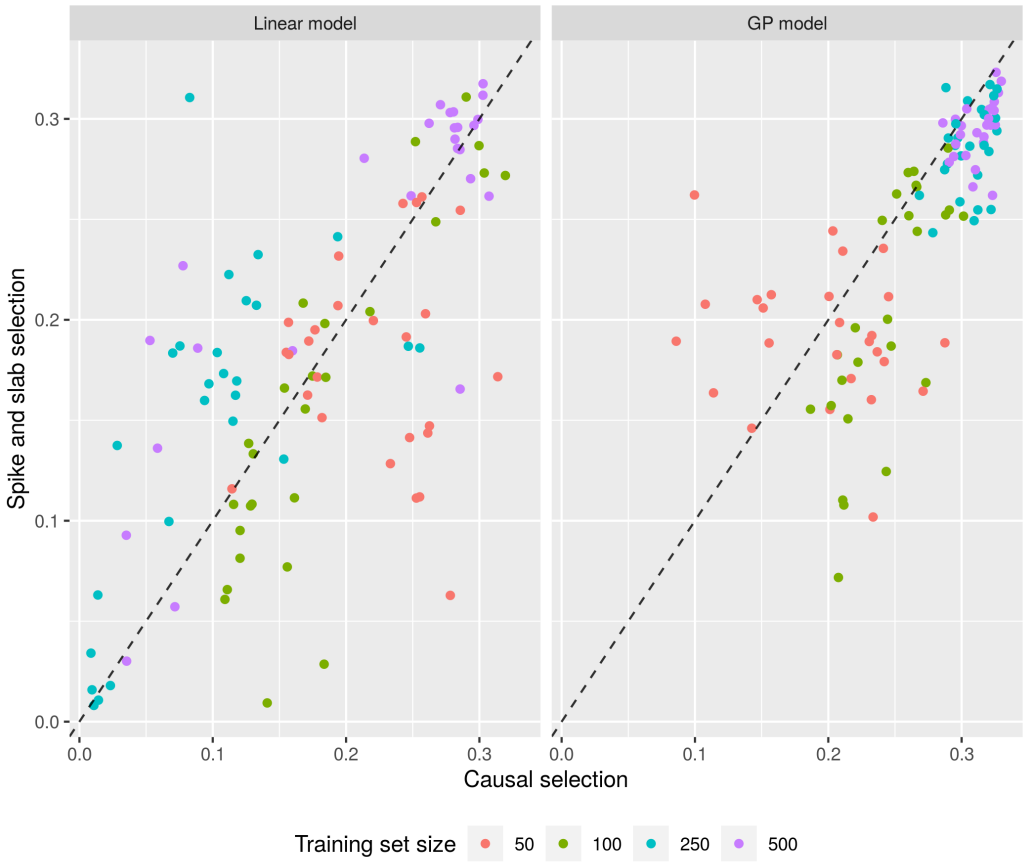


Figure 5.6: Comparison of correlation between predicted and true incidence rates for the linear (left) and Gaussian process (right) models when using spike-and-slab and causal selection.

correlation when using causal and spike-and-slab selection. Recall that temporal correlation was defined as the correlation between forecasted and observed time series at each health facility averaged over all facilities. This measures the ability of the model to predict temporal trends at each location. As well as calculating the temporal correlation over the entire 24 months forecasted, this metric was also calculated under restriction to months 13-24 in order to emphasise long-term predictive performance. Overall, both models appear to perform better in terms of temporal correlation when using causal selection compared to spike-and-slab selection. For the linear model, causal selection appeared to improve temporal correlation in the majority of the cases. When this was restricted to just the second year of predictions, the performance using

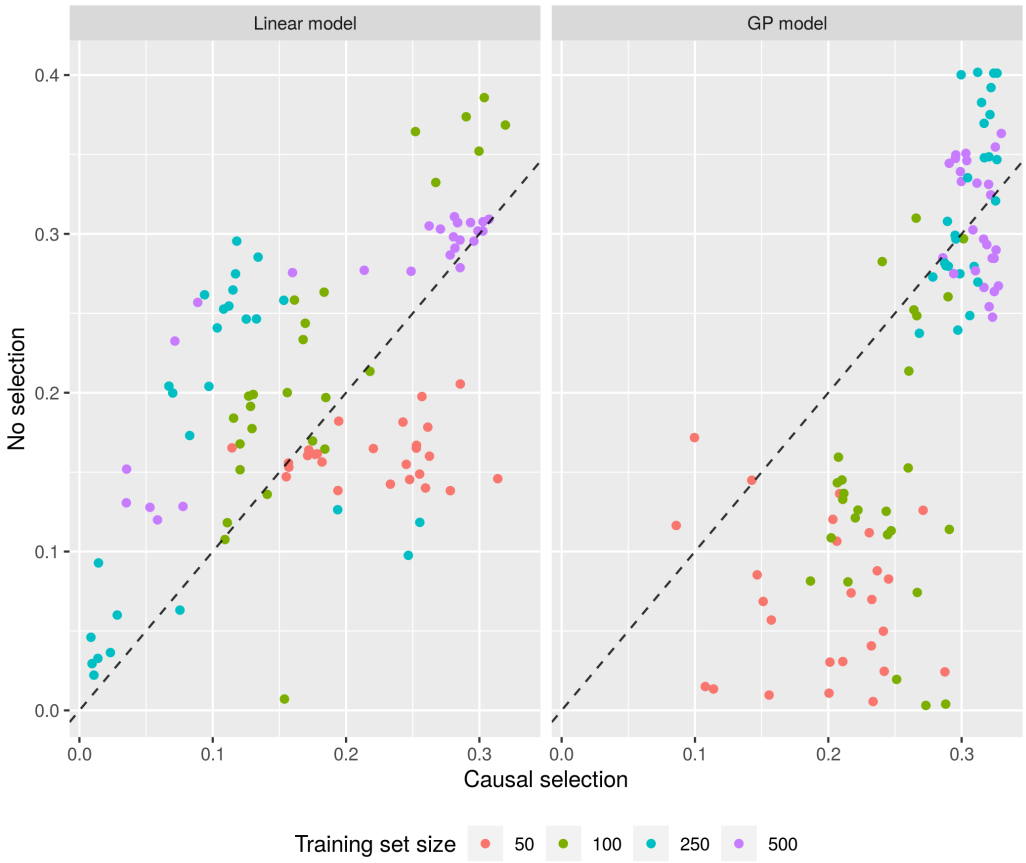


Figure 5.7: Comparison of correlation between predicted and true incidence rates for linear (left) and Gaussian process (right) models when using all features (no selection) and causal selection.

causal or spike-and-slab selection was more similar. The performance of the Gaussian process model in terms of temporal correlation was superior using causal selection in almost all cases, both over the whole time frame and in the second year.

Causal selection is compared to performing no feature selection in terms of temporal correlation in Figure 5.9. For the linear model, results were more favourable (over both time frames) to including all features in the model. The exception to this was for the smallest training sets, in which case temporal correlation was usually higher when using causal selection. For the Gaussian process model, there was clear advantage to using causal selection over no selection, over both years and in the second year alone. This improved performance was most notable for smaller training sets,

while for the largest training sets temporal correlation was similar with no selection or causal selection.

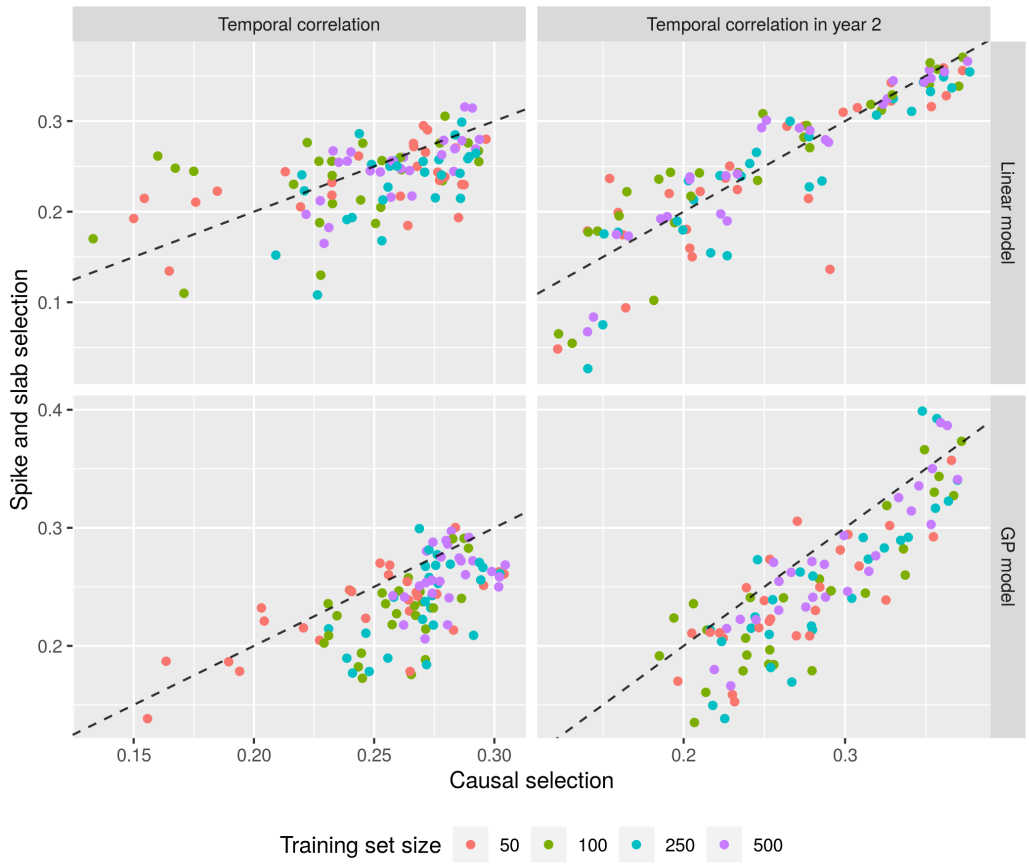


Figure 5.8: Comparison of temporal correlation for the linear (top) and Gaussian process (bottom) models when using spike-and-slab and causal selection. The time-series considered are all 24 months forecasted (left) and months 13-24 (right).

Using the metric of root mean square error, similar patterns in performance were seen. These comparisons are included in the appendix (Figures A4 and A5). For the linear model, using causal selection resulted in greater predictive errors with larger training sets and smaller errors with smaller training sets when compared to both spike-and-slab selection and no feature selection. For the Gaussian process model, predictive error was usually smaller when using causal selection compared to no selection. There was less of a clear difference for this model when comparing causal selection and spike-and-slab selection.

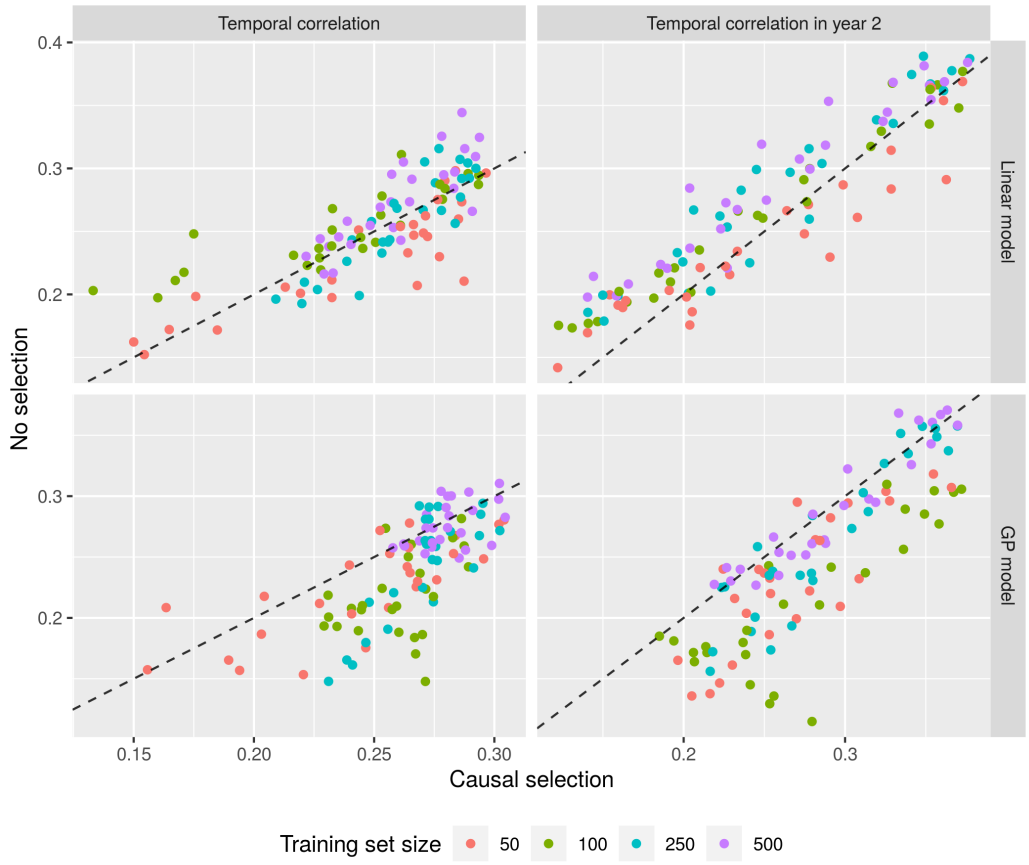


Figure 5.9: Comparison of temporal correlation for the linear (top) and Gaussian process (bottom) models when using all features (no selection) and causal selection. The time-series considered are all 24 months forecasted (left) and months 13-24 (right)

5.3 Discussion

A notable feature in these results is the stability of the causal feature selection procedure across the different iterations when compared to spike-and-slab selection. Instability in a selection method, defined loosely as large changes in the selected feature sets arising from small perturbations of the data, undermines the idea that the selected feature set represents a meaningful set of driving factors. A consistently selected feature set, on the other hand, is more likely to reflect true associations between explanatory and response variables and therefore result in better out-of-sample predictions. Many common feature selection methods are known to be unstable when

applied to linear or logistic regression [212, 213, 214]. One reason for this is that given a large number of potential features (and therefore a large number of candidate models), when using goodness-of-fit or cross-validation measures it is difficult to distinguish combinations of features that happen to fit the observed data well (either within-sample or in predictions for held-out observations) from truly informative feature sets. Bootstrapping samples has been proposed as a solution to this problem, for example the 'little bootstrap' method proposed by Breiman [215] to estimate predictive error which is then used to compare models. Austin and Tu [216] proposed a bootstrapping selection method where backwards elimination selection was applied to bootstrapped samples of the data and, as in our method, the features were ranked by the number of times they were selected, with features that were selected in at least a fixed proportion included in the final feature set. However, these methods do not address the possible misspecification of the models (resulting from assuming a linear effect of each feature). If the model is well-specified, informative features should perform well (in terms of goodness-of-fit or predictive error) in many of the bootstrapped samples but in the presence of misspecification this is not necessarily the case. The stability of the causal feature selection is likely to derive from both the use of bootstrapped samples and the non-parametric nature of the causal discovery algorithm. Spike-and-slab selection, on the other hand, showed relatively little consistency between iterations, suggesting that features were often selected due to chance associations in the training data that agreed with the model structure. While the more flexible Gaussian process model should suffer less from misspecification than the linear model, the assumed smoothness of these effects and lack of interactions between variables may still have prevented important variables being identified using spike-and-slab selection.

The ranking of variables by causal selection is also plausible—elevation is well-known to have a strong connection to transmission in Madagascar and aridity and

PET are likely to influence the presence of mosquito breeding sites. Both *P. falciparum* and *P. vivax* are endemic in the country and therefore it is reasonable that the temperature suitability of both these species is an important factor. It is less clear why nighttime temperature would be a more causal factor than temperature during the day, although this could relate to night-biting behaviour of the mosquito vector. I would caution against interpreting the reasonableness of these results as validation of the causal selection procedure, however, the results could direct further investigation.

While the causal selection algorithm produced consistent rankings of dynamic variables, the relative importance of the different time lags for each variable was less clear (for example whether nighttime temperature at a 0, 1, 2, or 3 month time lag was mostly likely to be a direct cause). This is perhaps unsurprising, as there is likely to be variation in the time taken for a change in dynamic variable to affect the observed case count (due to variation in, for example, the duration of the infection process, time taken to seek treatment, and reporting delays). However, it is also possible that the causal selection process was unable to infer meaningful differences between the same variable at different time lags and that different time lags were selected across repeats based on noise. This may be a consequence of the stability of the algorithm, as we would expect a trade off between the stability and specificity of any selection procedure.

When comparing the predictive performance of the models using causal selection and spike-and-slab selection, there was no clear superior strategy for the linear model in either metric, although on average causal selection produced slightly higher temporal correlations. In contrast, the Gaussian process model consistently performed better across both metrics when using causal selection. That is, causal selection was able to identify a single feature set that on average performed as well or better in both models compared to the model-specific features selected with spike-and-slab regression. Causal selection had the greatest benefit in the Gaussian process model,

possibly because the non-parametric discovery algorithm selected features that had non-linear relationships with log incidence which were better captured in the more flexible model.

Somewhat similar patterns were observed when comparing model performance when using causal selection or when performing no feature selection (instead including all available features), but in this case there was a clearer benefit to using all features for the linear model and causal selection for the non-linear model. This difference in performance of the linear model suggests that there were features that were informative within this model framework that were not selected by the causal selection procedure. It is worth noting that the linear model using all features also performed better than when using spike-and-slab selection, suggesting that these useful features were missed by both selection schemes. A notable exception was when the training sets were small, in which case causal selection often improved predictions. The superior performance of the Gaussian process model under causal selection suggests that when using all features this more flexible model overfit the data.

These results provide the first real evidence base to support the use of causal discovery as a feature selection tool for mapping malaria incidence and likely for disease mapping more generally. When using simple models with sufficiently large datasets, it may be that no feature selection is required and that important features can be identified while the model is fit. However, with more complex models or in lower data settings, using causal selection is likely to improve performance. More notably, these results suggest that causal feature selection is also likely to improve model performance when compared to classical feature selection methods. This is supported both by the improved performance under causal selection of the specific models considered in this work and the remarkable stability of the causal selection process (which is independent of any model).

5.3.1 Limitations

Having seen that the proposed causal discovery procedure shows promise as a feature selection tool, it is important to also consider the limitations of this approach. While the causal discovery procedure scales well with the number of data points (due to the use of linear-complexity conditional independence tests), the PC algorithm itself is likely to scale poorly with large numbers of variables. Even though this algorithm runs in polynomial time in the number of variables when the true underlying DAG is sparse, the order of this polynomial (which depends on the number of adjacencies of each vertex in the true DAG) may still cause computational issues. In terms of feature selection, this means this method may not be easily applied if the number of potential features is very large. More efficient causal discovery algorithms exist (see e.g. Maathuis et al. [217], Nandy et al. [218]) but require stronger assumptions about the data generating process. The PC algorithm may also be inappropriate due to the assumption that there are no unobserved confounders. In this case, however, our approach can be modified to use a discovery algorithm that is designed to be robust in the presence of some unobserved confounders, such as the FCI algorithm [40]. I also note that unobserved confounders are more likely to lead to edges being incorrectly retained, rather than incorrectly removed, in the PC algorithm and therefore the resulting parent set of the variable of interest is more likely to be a superset of the true set of parents than it is to be missing true parent variables. In many applications, including these extra variables in the feature set may not be a significant problem. This is in contrast to the problem of estimating causal effects, where the presence of unobserved confounders is more problematic. Moreover, these methods may be useful as a non-parametric feature selection method based on conditional independencies, even when true causal relationships are difficult to infer. Finally, it is possible that the causal structure that the PC algorithm aims to infer is different in different parts of Madagascar due to the diverse ecological landscape. In this case, it may be more

appropriate to run the selection procedure separately in different areas and aggregate the resulting feature sets somehow or simply to model different areas independently.

5.4 Conclusion

In this chapter I investigated the performance of the causal discovery procedure proposed in Chapter 3 as a feature selection tool for predicting malaria incidence. This was compared to feature selection using spike-and-slab regression and the alternative of not performing any feature selection. Two versions of the regression structure were used (one assuming a linear effect of each covariate on log incidence and the other allowing for non-linear effects) and the size of the training datasets was varied. Unlike the work in Chapter 4, this application of the causal discovery procedure made use of the methods for testing independence between variables observed at different spatial scales introduced in Section 3.3.6.

The results of this analysis suggest that causal feature selection is a promising selection tool. The stability of the causal feature selection method, which produced highly consistent rankings of features by importance in the different training sets, suggests that meaningful features were identified and was in contrast to spike-and-slab selection for either model (echoing previous observations of instability in various selection methods for linear and logistic regression [219, 214]). This stability is likely a result of both the non-parametric independence tests and bootstrapping of the data in the causal discovery procedure.

When comparing predictive performance, causal selection in general led to better predictions than using spike-and-slab regression in both models. This was despite the fact that causal selection is model-independent while the spike-and-slab regression implementations were specific to each model. Compared to using all features (i.e. no feature selection), causal selection resulted in superior predictive performance with

the more complex Gaussian process model and when using the linear model with smaller training sets. When using the linear model with large training sets, using all available features generally led to better predictions. Considering the benefits of using smaller feature sets (including greater interpretability and lower computational demands), the relatively small datasets common in disease mapping and the utility of more flexible non-linear models (see Chapters 4 and 6) these results provide strong support for using causal feature selection in similar contexts.

Chapter 6

Evaluating the effectiveness of malaria control methods: challenges and opportunities

In this chapter, I discuss the problem of estimating or predicting the effect of malaria control methods on transmission. This causal inference task is distinct from the causal discovery problems discussed in previous chapters. In the latter we focused on inferring the DAG describing the causal structure between variables; the challenge of estimating treatment effects begins by supposing this structure to be known, from either background knowledge or a previous application of a causal discovery method. The aim here is to understand how intervening on one variable would affect the other variables in the system. In this chapter I describe some of the challenges faced when applying these causal inference ideas in a malaria risk mapping setting and discuss potential approaches. I focus in particular on the task of identifying variation in intervention effectiveness at a sub-national level, with the ultimate aim of making maps of impact, similar to the maps of prevalence or incidence that we have seen in previous chapters. Using simulated data, I illustrate why using Gaussian processes (or other non-parametric methods) may be necessary when inferring intervention effectiveness on this sub-national scale. I also propose a two-step method for mapping individual-level treatment effectiveness.

There is a large body of work on estimating causal effects from observational data

in the causal inference literature, including the potential outcomes framework [36, 220] and the related g-methods and target trial approaches [30, 34, 221], and the theory of do-calculus [222]. Much of this theory aims to describe under which conditions causal effects are identifiable in complex scenarios involving, for example, multiple treatments or outcomes, time-varying treatments, or mediating variables. Where these effects are identifiable, these methods can then be used to derive expressions for these effects in terms of conditional distributions (which can be estimated from observed data). In the examples in this chapter I consider a simple causal structure, in which there is one treatment variable, one response variable and a number of observed confounding variables. As we will see, in this case the conditions for identifiability are straightforward, as is the relationship between the causal effect and conditional distributions. The challenge is instead primarily in determining whether the criteria for identifiability are met by the available data and in estimating conditional distributions with relatively small datasets.

The ability to evaluate, predict and understand the spatial variation of effectiveness of malaria control measures would clearly be a useful tool for operational planning and targeting of interventions but to date there has been relatively little work on this topic. Two notable examples that I discuss are the work of Bhatt et al. [26] and Weiss et al. [32] on predicting malaria prevalence in counterfactual intervention coverage scenarios in Africa. Both of these are focused on continent-wide predictions, however, and it is unclear how well variation is captured at a fine scale. Kraemer et al. [223] identify causal inference as a promising tool in spatial mapping but fail to examine in any real detail the opportunities or challenges of applying causal inference methods in this field. This chapter therefore serves to inform future work on this topic by highlighting potential approaches and challenges and emphasising the need for appropriate model validation.

6.1 Average and individual treatment effects

The potential outcomes framework, based on the work of Rubin and Neyman [36, 220], can be used to define and estimate the effect of a treatment on an outcome of interest. Suppose we are interested in the effect on individuals of a binary treatment variable $T \in \{0, 1\}$ with respect to an outcome variable of interest Y. In this framework, for each individual there is a fixed value that the variable Y would take if the treatment was applied, denoted $Y_{T=1}$, and another fixed value that it would take without the treatment, $Y_{T=0}$. The values $Y_{T=1}$ and $Y_{T=0}$ are the potential outcomes with and without treatment. For each individual, only one of these outcomes is observed, while the other is a counterfactual outcome. For example, if the treatment is given to an individual then for that individual the observed outcome is $Y_{T=1}$ and the counterfactual outcome is $Y_{T=0}$. The effect of the treatment on that individual is then defined as the difference between these outcomes,

$$Y_{T=1} - Y_{T=0}.$$

Note that the 'individual' here refers to whatever unit the treatment is applied to. In the context of disease control, this could be an individual, a household or a community. Defined in this way, the treatment effect for an individual can never be calculated as only one outcome is ever observed, a fact that is sometimes known as the 'fundamental problem of causal inference'. However, a number of related quantities can be defined that can be estimated from observations of multiple individuals. One example is the *average treatment effect (ATE)*, defined as the expected individual treatment effect over the whole population,

ATE

$$\text{ATE} = \text{E}[Y_{T=1} - Y_{T=0}].$$

Suppose that N individuals are sampled from the population to participate in a trial where the treatment is randomly assigned. Let the assigned treatments be $t_1, ..., t_N$ and the observed responses be $y_1, ..., y_N$. Then we could propose estimating the average treatment effect by the difference between the average response in the treatment group and the control group:

$$\widehat{\text{ATE}} = \frac{1}{N_{T=1}} \sum_{t_i=1} y_i - \frac{1}{N_{T=0}} \sum_{t_j=0} y_j \tag{6.1}$$

where $N_{T=1}$ is the number in the treatment group and $N_{T=0}$ is the number in the control group, since we would not expect any systematic differences between these groups before treatment. This is indeed a consistent estimator for the average treatment effect under an additional condition, the *stable unit treatment value (SUTVA)* assumption [220]. This is the assumption that

1. there is only one version of the treatment, and

2. the outcome for an individual is unaffected by the treatment status of any other individual (also known as the 'no interaction between units' assumption [224]).

The first condition would be violated if, for example, the treatment was a drug and different doses had been used (in which case the estimand would not be well defined). The second condition might not hold if the effect of the treatment can 'spill over' to other individuals. For example, in a vaccination trial (for an effective vaccine) if two individuals live in the same house and one receives the true vaccine then this could reduce the risk to the other, as they are less likely to be infected at home. With randomised treatment assignment the individuals in the treatment and control groups are exchangeable because the random assignment ensures that (before the response is observed) the distribution of $Y_{T=1}$ is the same for any individual in the treatment group or the control group, and so is the distribution of $Y_{T=0}$. Therefore, the average outcome in the treatment group is an unbiased and consistent estimator

of the counterfactual outcome of an individual in the control group had they received the treatment, and vice versa.

When working with data where the treatment was not randomly assigned, exchangeability of the treated and untreated groups does not necessarily hold. In many cases, the treatment and response variables are confounded; that is, there is a variable that is a common cause of both. For example, suppose we are interested in the effect of a drug, T, on the risk of heart attacks, Y, but this drug was preferentially prescribed to patients who have high blood pressure (which is also assumed to be a cause of heart attacks). In this case, the rate of heart attacks in the treated group is not a good estimator for the rate we would have expected to see in the untreated group had they been treated. This is because, on average, the treated group was already at higher risk. In the language of potential outcomes, the distribution of the potential outcome under treatment $Y_{T=1}$ was not the same in the treated and untreated groups (and similarly for the distribution of $Y_{T=0}$). If (6.1) was computed in this case, we would expect it to underestimate the beneficial effect of the drug. In this example, blood pressure is a cause of both treatment and outcome.

If nothing is known about how the treatment is assigned then in general the problem of estimating the effect of this treatment is intractable, as there is no way of computing a principled estimate of the counterfactual outcome of the treated had they not been treated based on the outcomes in the untreated group, and vice versa. However, if the confounding variables are known and measured then the estimation of treatment effects may proceed as follows. Suppose that for some set of variables X representing individual traits, $Y_{T=i}$ and T are independent given X, for $i = 0, 1$. This assumption is sometimes called the *unconfoundedness* assumption and it is clear (by observation or from d-separation) that X must be a contain any confounding variables that exist between the treatment and outcome, other than those for which the effect on both the treatment and outcome is mediated by a single other variable. A sufficient

unconfoundedness

condition for such a set, X, to satisfy this assumption is that the probability of treatment is the same for any individuals with the same values of X. The probability of treatment given $X = x$, $P(T = 1|X = x)$, is known as the *propensity score*. If such a set of variables X exists, then within the subpopulation with the same value of X the treated and untreated individuals are exchangeable, since within this group treatment status is independent of $Y_{T=i}$. We also assume *overlap*, that for any value of X, $0 < P(T = 1|X = x) < 1$, which ensures that any individual in the population has the possibility of both being assigned and not being assigned the treatment. Under these assumptions, there are a number of ways to estimate the average treatment effect, including propensity score matching [225, 226] and inverse probability weighting [34]. We can also define the *average treatment effect on the treated (ATT)*,

overlap

$$\text{ATT} = \text{E}[Y_{T=1} - Y_{T=0}|T = 1]$$

and the *average treatment effect on the untreated (ATU)*,

$$\text{ATU} = \text{E}[Y_{T=1} - Y_{T=0}|T = 0].$$

These effects are not necessarily equal to the average treatment effect, or to each other, due to the differences between the treated and untreated groups. Returning to the example above of the drug for reducing the risk of heart attacks, it could be the case that the reason this drug is preferentially prescribed to individuals with higher blood pressure is that it is more effective for these individuals. In this case, we would expect the ATT to be greater than the ATU. We can also define treatment effects on other subsets of the population, including the *individualised treatment effect (ITE)*,

ITE

$$T(x) = \text{E}[Y_{T=1} - Y_{T=0}|X = x].$$

If the subpopulation with $X = x$ was large enough for each value x, then to estimate $T(x)$ we could restrict our attention to this subpopulation and this problem reduces to estimating the ATE under random treatment assignment. In practice such subpopulations are often small relative to the available data so this is impractical. Instead, the outcomes $Y_{T=1}$ and $Y_{T=0}$ are usually modelled as functions of the variables X, allowing the sharing of information between observations with different covariate values. For example, the outcomes could be treated as Gaussian random variables and modelled separately,

$$Y_{T=i}|(X = x) \sim \mathrm{N}(f_i(x), \sigma_i^2)$$

for $i = 0, 1$, or the outcomes could be modelled jointly,

$$Y_{T=i}|(X = x) \sim \mathrm{N}(f(x, i), \sigma^2) \tag{6.2}$$

where $\epsilon \sim N(0, \sigma_i)$. The latter formulation usually involves stronger assumptions about the form of the effect (for example, it could assume a constant effect of the intervention for all values of X). This may result in a less flexible model but can be useful in low data settings or when X is high dimensional. Once the functions f_i or f have been estimated, with functions $\hat{f}_i$ and $\hat{f}$, an estimate for the treatment effect when $X = x$ is given by

$$\hat{T}(x) = \hat{f}_1(x) - \hat{f}_0(x)$$

or

$$\hat{T}(x) = \hat{f}(x, 1) - \hat{f}(x, 0).$$

Other distributions for the potential outcomes can similarly be considered, for example if the outcomes are case counts we could model this as

$$Y_{T=i}|(X = x) \sim \mathrm{Poisson}(f(x, i))$$

where f_i is the function to be learned for $i = 0, 1$. In this chapter, the quantities of interest (spatially varying effects of interventions) can usually be interpreted as individualised treatment effects.

So far we have defined all these concepts in terms of a binary treatment variable but these ideas can be extended to accommodate treatment variables that can take several discrete values or are continuous. In this case, effects can be defined with respect to pairs of values of this variable. For example, the treatment effect on an individual of increasing a continuous treatment variable T from t_1 to t_2 can be defined as

$$Y_{T=t_2} - Y_{T=t_1}.$$

When it comes to modelling individual effects of a continuous treatment, modelling the potential outcomes as a single function of confounding variables and this treatment, as in (6.2), is most appropriate.

The idea of individualised treatment effects can also be arrived at using the theory of do-calculus developed by Pearl and collaborators [222, 227]. This framework defines the 'do' operator, which represents intervening on a variable T and setting it to a desired value t. This allows us to distinguish a conditional distribution $Y|T = t$ from a *manipulated distribution* $Y|\text{do}(T = t)$. The former is the distribution of Y in the subset of the population for which the value $T = t$, while the latter is the distribution of Y in the whole population if T was intervened on directly and (for each individual) set to t. For example, consider the (deterministic) relationship between the position of a light switch, T, which controls whether a light is on or off, Y. Suppose that the switch is flipped at random times such that any time it is observed there is a 50% chance that the switch and light are both on (and otherwise they are both off). The conditional distribution $Y|T = 0$ is a point mass at $Y = 0$, since if we observe that the switch is off then the light must be off. Similarly, the conditional distribution $T|Y = 0$ is a point mass at $T = 0$, since if we see that the light is off we know

the switch must be off. The manipulated distribution $Y|\text{do}(T = 0)$ is the same as the conditional distribution $Y|T = 0$, since if we intervene on the system to put the switch in the off position (regardless of its position before) then the light will turn off. On the other hand, if at some point we intervene on the light directly to turn it off, for example by removing it from the circuit, then this does not give us any information about the position of the switch. This means that the manipulated distribution $T|\text{do}(Y = 0)$ is not the same as the conditional distribution $T|Y = 0$. In fact, in this case the manipulated distribution is identical to the unconditional distribution of T. Manipulated distributions can also be defined for subsets of the population, for example $Y|\text{do}(T = t), X = x$ is the distribution of Y where the value of T is set to t in the subset of the population where $X = x$.

When we talk about the effect of interventions, we are interested in manipulated distributions, rather than conditional ones. If T is an intervention and Y is malaria transmission some time after the intervention is applied (where it is used), then if the intervention is targeted at high risk areas it is plausible that the conditional distribution $Y|T = 1$ will be more concentrated on higher values than $Y|T = 0$ even if the intervention reduces transmission. These conditional distributions therefore do not represent the causal effect. On the other hand, the difference between manipulated distributions $Y|\text{do}(T = 1)$ and $Y|\text{do}(T = 0)$ does capture this effect. The average treatment effect can be written in terms of these distributions as

$$ATE = \text{E}[Y|\text{do}(T = 1)] - \text{E}[Y|\text{do}(T = 0)]$$

as can the individualised treatment effect,

$$T(x) = \text{E}[Y|\text{do}(T = 1), X = x] - \text{E}[Y|\text{do}(T = 0), X = x].$$

While we can define the concepts of the 'do' operator and manipulated distributions,

the question remains of how to estimate these distributions. Pearl [222] provides three rules of do-calculus that describe conditions under which manipulated distributions are equal to other manipulated distributions or to conditional distributions. These rules can be applied recursively to determine whether a specific manipulated distribution is identifiable from the observed data and, if it is, obtain an expression for it in terms of conditional distributions. For our purposes, we simply need the result that if Y is independent of T given some variables X, then

$$P(Y|\text{do}(T=t), X=x) = P(Y|T=t, X=x)$$

for all t and x. This is a similar result to the requirement of unconfoundedness when estimating individualised treatment effects.

6.2 Effect of malaria interventions

Numerous RCTs have been carried out to investigate the effect of different malaria control methods, including insecticide-treated bednets (ITNs) [228] and indoor residual spraying (IRS) [229]. ITNs act as a physical barrier preventing mosquitoes from biting those sleeping under them and the insecticide repels and kills mosquitoes. At a high enough coverage in a community, ITNs can cause a sufficiently large reduction in vector abundance to decrease transmission. IRS involves spraying the indoor walls of buildings with long-lasting insecticide, which kills mosquitoes that rest on them. Mosquitoes typically rest on walls after feeding, so IRS does not prevent individuals from being bitten directly but reduces onward transmission and, like ITNs, has the potential to reduce vector abundance.

RCTs have been carried out at both the community level [230, 231, 232], with different villages or study areas randomly assigned the treatment or the control (usually simply no treatment), and the individual level [233]. Other studies have used observa-

tional data to estimate effect sizes at specific locations and points in time [234, 235]. There is clear evidence that both ITNs and IRS are effective at reducing transmission [229, 228]. However, from the different trials there is also evidence of spatiotemporal heterogeneity in effect sizes (even within the same study [236]). Effectiveness of control measures may be affected by a number of factors, including baseline transmission; relative abundance of different vector species and vector behaviour (both IRS and ITN effectiveness is likely to depend on propensity for indoor and nighttime biting); insecticide resistance; and socioeconomic factors such as housing quality. In many countries, interventions are targeted based on estimated transmission in order to make the best use of limited resources. Sub-national estimates of intervention effectiveness could be another useful tool for optimising the distribution of these resources.

One approach for estimating and comparing the effects of interventions is through mechanistic modelling [237, 238]. Unlike the statistical models described in this thesis, which learn associations between explanatory and response variables, mechanistic approaches aim to explicitly model some of the biological processes of disease transmission. For example, the infection rate of an individual could be some function of biting rate (the rate at which they are bitten by infected mosquitoes), with the functional form of this relationship informed by biological knowledge, experiments and field studies. Using highly detailed agent-based models, such as the Epidemiological MODeling software (EMOD) [239], the impact of an intervention can be predicted while varying the level of coverage of this intervention, baseline transmission, coverage of other interventions, population demographics and many other factors. However, when aiming to predict the spatial variation of intervention effectiveness, these models may be of limited use because the values of the factors that affect intervention effectiveness within the model (including biological parameters, such as biting rate, and the other modifiable parameters mentioned) would themselves need to be known

at each location of interest, which is rarely the case. For example, to predict effectiveness of a fine grid of pixels, these input variables would need to be known at each pixel.

In order to attempt to learn spatial heterogeneity of intervention effectiveness from observed data, it may therefore be more useful to take a statistical approach. The first notable attempt to evaluate the spatiotemporal effect of interventions was the work by Bhatt et al. [26] on assessing the impact of interventions on malaria transmission in sub-Saharan Africa between 2000 and 2015. In this work, a Bayesian spatiotemporal model, similar to those used in previous chapters, was fit to prevalence survey data and clinical incidence data. Malaria risk was modelled as a function of environmental and socioeconomic variables, estimated coverage of interventions (artemisinin-based combination therapy (ACT), ITNs and IRS), country-level random effects and spatiotemporally structured noise. This model was used to both estimate the actual prevalence of malaria across the region between 2000 and 2015 and to produce counterfactual predictions of prevalence under different intervention scenarios. Counterfactual scenarios in which different interventions were removed were created by making a posteriori predictions with the fitted model in which the coverage of these interventions was set to zero. More recently, Weiss et al. [32] used a similar updated model to predict the effect of reduced intervention coverage due to the COVID-19 pandemic on malaria mortality and morbidity in Africa. A similar geospatial model was used and a baseline level of transmission was predicted, this time using the fitted model with pre-pandemic coverage levels. Alternative scenarios in which intervention coverage was reduced due to the pandemic were also predicted using this same fitted model. In making counterfactual predictions across space and time, these models are essentially estimating individualised treatment effects via a regression structure, where an 'individual' is a location in space and time.

6.2.1 Simulated data example 1: the impact of unobserved confounders

In Section 6.1, we saw that we can infer causal effects from observational data if we have observed a set of variables, X, for each individual that satisfies the unconfoundedness assumption—that the treatment and outcome variables are independent conditional on X. While there exist methods for inferring causal effects when such a set X does not exist (such as instrumental variables approaches [240] and the front-door formula [222]), they require the causal relationships between variables to be structured in specific ways. The more general approaches to inferring causal effects, which depend on controlling or adjusting for confounding factors, require that all of these confounders are observed. This includes the methods of Bhatt et al. [26] and Weiss et al. [32], although this is not explicitly stated. Unfortunately, the unconfoundedness assumption is difficult to test using observational data, although the use of a negative control can highlight cases where the assumption may not hold [241] (this is discussed in more detail in Section 6.5).

To illustrate the bias that can be introduced when unobserved confounders are present, I simulated data in which the treatment and response variables were confounded and then fit this data with the same model structure in three ways:

1. Including all covariates used to generate the data in the model.
2. Withholding one covariate that affected the response but not the treatment assignment.
3. Withholding one covariate that affected both the response and treatment assignment, i.e. a confounder.

This allows us to investigate the effect of an unmeasured confounder on estimates of treatment effectiveness. The second model fit was included to illustrate the difference

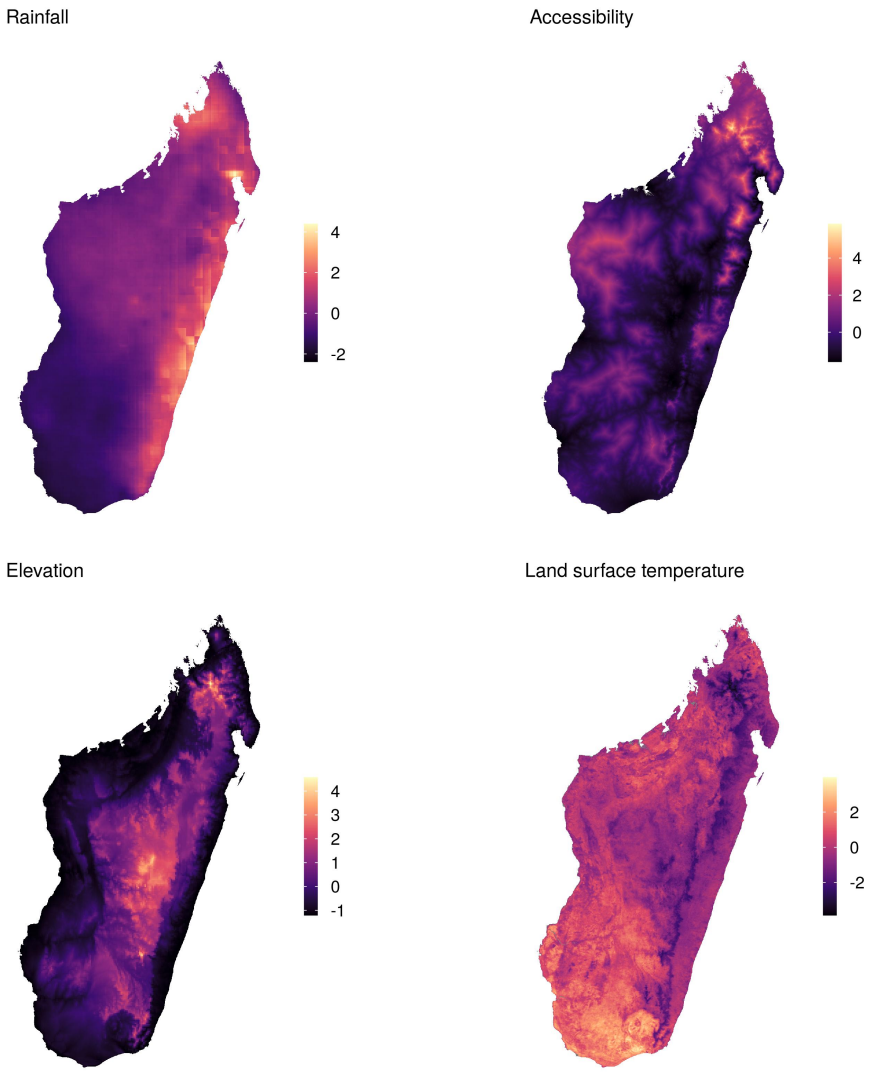


Figure 6.1: Covariates used for simulated prevalence datasets.

between the effect of an unmeasured confounder and an unmeasured variable that only affects the response. The data was designed to loosely resemble malaria prevalence in Madagascar with a binary treatment variable that could be thought of as indicating the use of IRS in different locations. The data was simulated as follows.

- **Baseline prevalence.** The baseline prevalence (prevalence in each location if IRS was not used) was a function of four covariates: accessibility to cities (access), rainfall, elevation and land surface temperature (LST). Note that accessibility to cities (before being standardised) is quantified by the travel time

to the nearest city above a certain size, and therefore a higher value represents a more remote location. Logit-transformed baseline prevalence was simulated as a linear combination of these covariates. Real covariate values from Madagascar were used, with 2013 annual averages used for the temporal variables (rainfall and LST). These covariate values are shown in Figure 6.1. The baseline prevalence, p_{baseline}, at a given location was

$$\text{logit}(p_{\text{baseline}}) = \beta_0 + \beta_{\text{access}} X_{\text{access}} + \beta_{\text{elevation}} X_{\text{elevation}} + \beta_{\text{rain}} X_{\text{rain}} + \beta_{\text{LST}} X_{\text{LST}} \tag{6.3}$$

where $X_{\text{access}}, X_{\text{elevation}}, X_{\text{rain}}, X_{\text{LST}}$ were the corresponding covariate values at this location. The following coefficient values were chosen,

$$\begin{aligned} \beta_0 &= -1.5 \\ \beta_{\text{access}} &= 0.5 \\ \beta_{\text{elevation}} &= -0.5 \\ \beta_{\text{rain}} &= 0.7 \\ \beta_{\text{LST}} &= 0.8 \end{aligned}$$

in order to reflect fairly realistic associations (at least in terms of sign) between these variables and baseline prevalence. These values were chosen to produce a plausible range of prevalence values but the patterns in the results are otherwise not particularly sensitive to the specific choice of these values.

- **IRS distribution.** IRS was treated as a binary variable, with a value of 1 for a location meaning that buildings in this location had been sprayed, and a value of 0 meaning they had not. The probability of IRS being conducted in a location was simulated as dependent on accessibility to cities, with a higher probability for locations nearer to cities. This probability, π_{IRS}, was modelled

as

$$\pi_{\text{IRS}}(p_{\text{baseline}}) = \frac{1}{1 + \exp(X_{\text{access}})}$$

so as X_{access} increased (meaning distance from the nearest city increased), the probability of IRS being conducted decreased. In real-world situations such a distribution pattern may well occur, as accessing remote locations in order to conduct IRS (or deliver other interventions) could prove difficult. The binary variable representing whether IRS had been carried out at this location, denoted X_{IRS}, was sampled as a Bernoulli random variable with probability π_{IRS} of success.

- **IRS effect.** The effect of IRS was simulated as dependent on the baseline prevalence, with a greater effect in areas where baseline prevalence was higher. This was achieved by simulating IRS as having a fixed effect of -0.5 on logit-transformed prevalence. Therefore, the prevalence, p, at a given location was simulated as

$$\text{logit}(p) = \text{logit}(p_{\text{baseline}}) - 0.5X_{\text{IRS}} \tag{6.4}$$

where $\text{logit}(p_{\text{baseline}})$ was calculated using (6.3).

- **Observed data.** Prevalence was simulated on a grid of 1km-by-1km pixels. Of these pixels, 600 were randomly sampled to represent survey sites and at each site the number of individuals in the survey was an integer sampled uniformly between 10 and 40. Let p_i be the prevalence at location i, calculated using (6.4). Then the number of individuals with infection observed, Y_i, out of the total tested, N_i, was a realisation of binomial variable,

$$Y_i \sim \text{Binomial}(N_i, p_i).$$

Of these 600 surveys, 300 were randomly selected to be 'observed' and used to

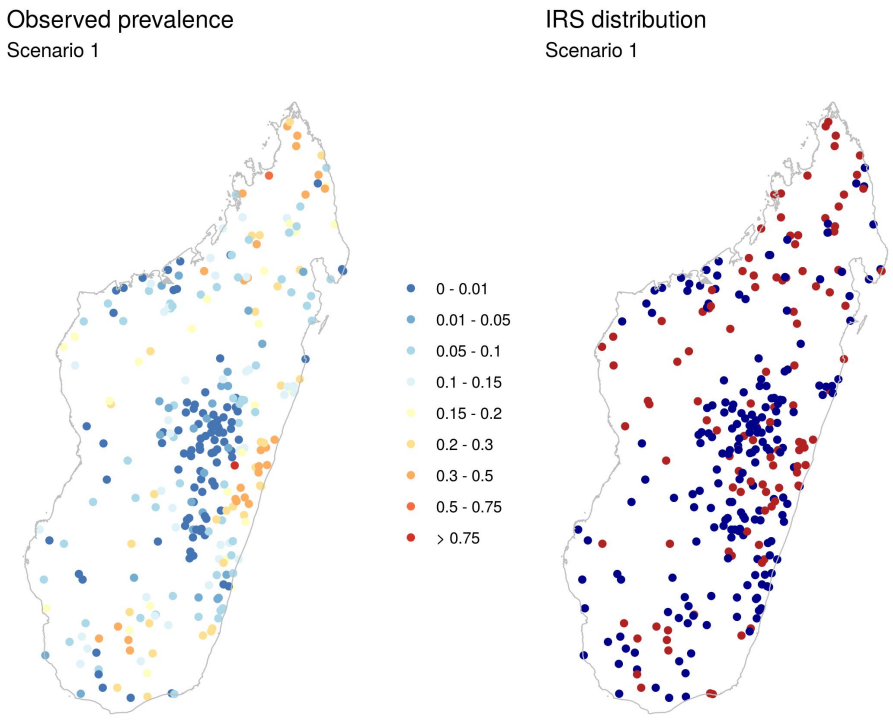


Figure 6.2: Observed prevalence (left) and IRS (right) in simulated data example 1. For IRS distribution, blue points represents locations where IRS was used and red where it was not.

fit the model. The remaining 300 surveys were used for model validation. The observed prevalence (the ratio Y_i/N_i) and IRS coverage at the surveys used to fit the model is shown in Figure 6.2.

The survey data from the 300 observed locations was fit with a Bayesian geospatial model, where prevalence at location i, p_i, was modelled (on the logit scale) as a linear combination of covariates and spatially correlated noise,

$$\text{logit}(p_i) = \beta_0 + \beta^T X_i + f(s_i) \tag{6.5}$$

where β_0, β were parameters, X_i were the covariate values at this location, s_i was the location and f was a Gaussian process with Mátern covariance. Observations at each

location were then modelled as binomial random variables based on this prevalence,

$$Y_i \sim \text{Binomial}(N_i, p_i). \tag{6.6}$$

As discussed above, this model was fit three times. The variables included in each fit were:

1. All five variables—accessibility, elevation, rainfall, LST and IRS use.
2. All variables except rainfall.
3. All variables except accessibility.

Recall that accessibility was a cause of both prevalence and IRS use, so in this third case (and only in this case) there was an unmeasured confounding variable. The fitted coefficient values in each of these cases are shown in Table 6.1. When accessibility is included, the value of the coefficient for IRS use is similar to the true size of the causal effect of IRS on logit prevalence. When accessibility is not included, however, the learned coefficient of IRS is more than twice the true causal effect. To understand why this is, consider the DAG shown in Figure 6.3 which represents the causal relationships between these variables. The association between IRS and observed prevalence in the data is a combination of the causal effect of IRS on prevalence (the path IRS $\rightarrow$ pr) and the non-causal path (sometimes called a 'backdoor path') IRS $\leftarrow$ access $\rightarrow$ pr. This backdoor path represents the fact that a location where IRS was carried out is more likely to be accessible than a location where it was not, and a more accessible location is more likely to have lower prevalence. When accessibility was not included in the model, the coefficient for IRS represented the association resulting from both these paths.

Figure 6.4 shows the true effect of IRS (defined as the difference between baseline prevalence and prevalence if IRS was conducted everywhere) and the effect predicted

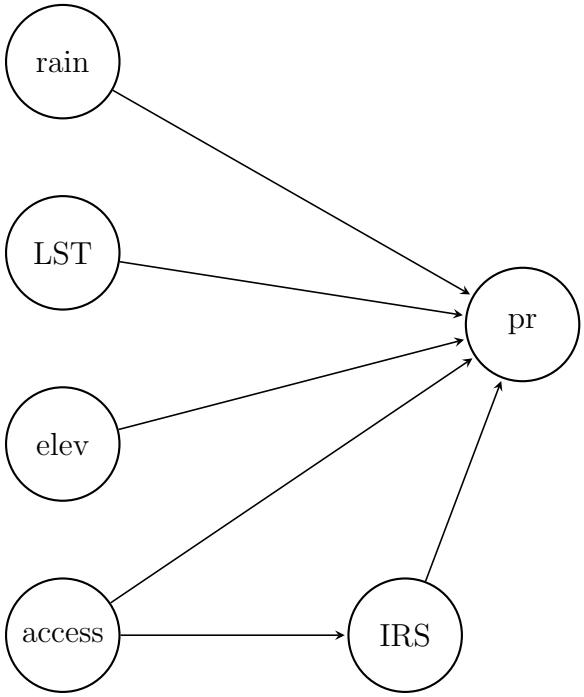


Figure 6.3: A DAG illustrating the causal relationships between variables in simulated data example 1. Here 'elev' and 'pr' represent elevation and prevalence (observed in a survey), respectively.

Table 6.1: True covariate coefficients from the data generating process and posterior mean values in the three model fits.

	True	All covariates	No rainfall	No accessibility
Accessibility	0.5	0.650	0.461	
Elevation	-0.5	0.049	-0.399	-0.086
Rainfall	0.7	0.837		0.651
LST	0.8	1.076	0.252	0.894
IRS use	-0.5	-0.544	-0.648	-1.056

by this model (in the three different fits) if the coefficient of IRS was assumed to represent the causal effect. In each case, this predicted effect was calculated by first predicting baseline prevalence and prevalence if IRS was carried out everywhere by setting X_{IRS} to 0 and 1, respectively, and using observed values for all other covariates. The predicted effect was then the difference between these surfaces. This is similar to how predictions were made by Bhatt et al. [26] and Weiss et al. [32]. We can see that the pattern of the predicted effect is very accurate in the model fit with all the covariates. When rainfall was withheld, the scale of the effect was somewhat overstated and the spatial trends were less accurate. This is expected, since the true size of the effect depended on baseline prevalence and this baseline prevalence was predicted less accurately without rainfall. When accessibility was withheld, the predicted effect was much higher across the country, as expected given the fitted coefficient value.

In contrast, when predicting actual prevalence (rather than the counterfactual scenarios of no IRS use or complete coverage) the model consistently performed fairly well. Figure 6.5 shows predicted and observed prevalence at the 300 survey locations that were not used in the model fitting process. Note that observed prevalence here, as before, means the proportion of individuals in the sample that were infected, not the underlying prevalence, p, at this location. In all three model fits these predictions were largely accurate. The correlation between predicted and observed values was

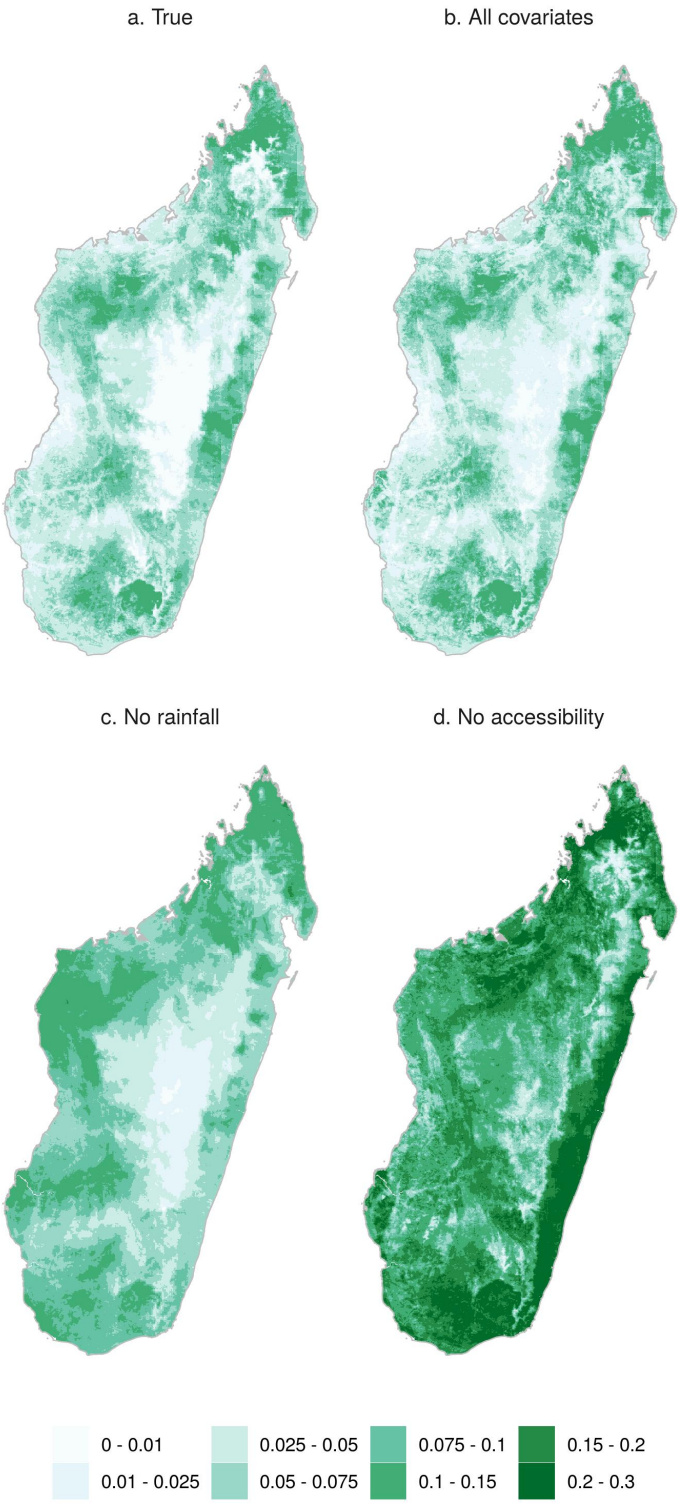


Figure 6.4: (a) True effect of IRS on prevalence, and predicted effect when the model was fit with (b) all covariates, (c) without rainfall, and (d) without accessibility.

0.793, 0.702 and 0.630 when the model was fit with all covariates, without rainfall and without accessibility, respectively. Although correlation was lowest when accessibility (the confounding variable) was not included, this value is still relatively high indicating good predictive performance. The difference between these correlations when fit without rainfall and without accessibility certainly does not seem to reflect the much greater extent to which the model overestimates effectiveness when fit without accessibility compared to without rainfall. Therefore, cross-validation would likely not highlight that the causal effect was not correctly estimated (this idea is discussed in more detail in Section 6.5). This illustrates that even when the confounder was unobserved the model was effective in learning the associations between the response and explanatory variables. The problem was not that the model learned poorly, but that when the confounder was unobserved the coefficient for IRS in the model did not represent the causal effect. In the language of manipulated distributions introduced in Section 6.1, when estimating intervention effectiveness the estimand we were interested in was difference between the expected prevalence when intervening to not use IRS and when intervening to use IRS,

$$\mathrm{E}[Y|\mathrm{do}(X_{\mathrm{IRS}}=0), X \setminus X_{\mathrm{IRS}}] - \mathrm{E}[Y|\mathrm{do}(X_{\mathrm{IRS}}=1), X \setminus X_{\mathrm{IRS}}]$$

and the estimate used was

$$\hat{\mathrm{E}}[Y|X_{\mathrm{IRS}}=0, X \setminus X_{\mathrm{IRS}}] - \hat{\mathrm{E}}[Y|X_{\mathrm{IRS}}=1, X \setminus X_{\mathrm{IRS}}].$$

When there was an unobserved confounder, the conditional distributions

$$\mathrm{E}[Y|X_{\mathrm{IRS}}=i, X \setminus X_{\mathrm{IRS}}] \quad i=1,2$$

were not equal to the manipulated distributions

$$\mathrm{E}[Y|\mathrm{do}(X_{\mathrm{IRS}} = i), X \setminus X_{\mathrm{IRS}}] \quad i = 1, 2$$

and so this resulted in a biased estimate of effectiveness regardless of how well the conditional distribution was inferred.

It is worth noting that the model was able to accurately infer the effect of IRS when accessibility was observed in part because the model structure was correctly specified (except for missing variables in some cases). This allowed the model to correctly distinguish the effect of accessibility (when observed) on prevalence from that of IRS, even though IRS distribution was biased towards more accessible areas. If either of these variables had a non-linear effect on logit prevalence then these effects would not be captured exactly by this model. Such an example is given in the next section. Importantly, however, if all confounders are measured then this is a problem of accurately estimating conditional distributions (or conditional expectations). This is a purely predictive (not causal) problem and therefore standard model validation methods can be used to assess whether this has been successfully achieved.

In general, if an unmeasured confounder exists then there is little that can be done to account for it. However, we can use this example to reason about when we would expect such a confounder to significantly bias results and in which direction. Recall that when accessibility was not observed, the learned association between IRS and prevalence was a combination of the causal effect of IRS on prevalence and the backdoor path, $\mathrm{IRS} \leftarrow \mathrm{access} \rightarrow \mathrm{pr}$. In this example, the intervention was preferentially distributed to locations which were already at lower risk, inflating the apparent protective effect of the intervention. If the intervention was instead more likely to be used in high risk areas (and the confounding variable was unmeasured), then learned associations would underestimate intervention effectiveness. Additionally, in

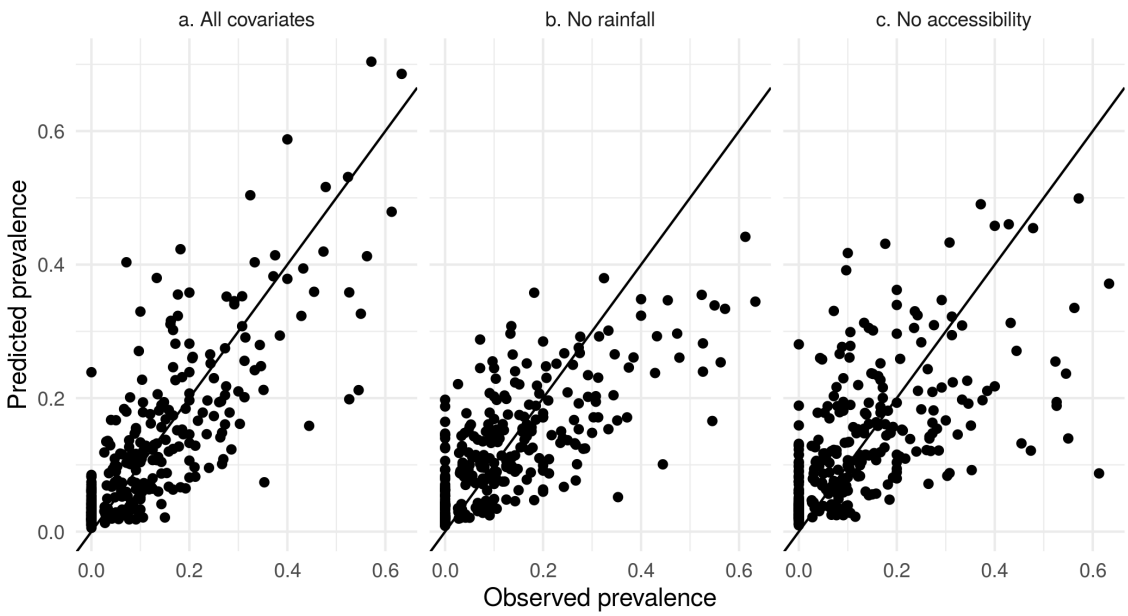


Figure 6.5: Observed and predicted prevalence at validation locations in simulated data example 1 when the model was fit with (a) all covariates, (b) without rainfall, and (c) without accessibility.

this example the strength of associations along each edge in this backdoor path were relatively strong, and overall this path made up a large proportion of the observed association. If instead the effect of the confounding variable on either the treatment or response variable was small then the learned association would be a less biased estimate of the causal effect. While this is inherently difficult to test (as the confounder is unmeasured), if it is expected that most of the factors that have a large impact on prevalence (or similarly IRS distribution) are included in the model then we may expect any bias due to confounding to be small.

6.3 Increasing model flexibility

In the model used in the previous section and in the work of Bhatt et al. [26] and Weiss et al. [32], the effect of an intervention was assumed to not interact with any other variables in the model (such as environmental or socioeconomic covariates or other

interventions when more than one was considered). This is a pragmatic choice which limits model complexity and in the context of continent-wide predictions (which are made on a fine-scale but aggregated to coarser levels for interpretation) these distinctions may be relatively unimportant. When using binomial regression or the empirical logit transformation (or Poisson regression if the response variable is incidence) this results in the effect of a fixed level of intervention coverage being dependent only on pre-intervention transmission levels. For example, when comparing two locations, the effect of increasing ITN coverage to 50% (compared to a baseline of no coverage) will always be greater in the location with the greater baseline transmission. While this is logical, predictions of the impact of interventions are therefore of similar utility to predictions of current risk (although including data on current interventions in the model may improve these predictions) when it comes to targeting interventions. It may instead be useful if these models could identify areas where an intervention may be particularly effective when compared to other areas with similar levels of transmission. These differences could be due to environmental conditions (which affect the relative abundance of different vector species, which may have different biting behaviour [242, 243]). For example, in an analysis of a recent IRS campaign in Northern Zambia—an area with two seasonal peaks in transmission, one in the dry season and one in the wet season—IRS appeared to have a strong effect in the rainy season but little to no effect in the dry season, despite high transmission occurring in both. This may be because of the behaviour of different species of vector, as the abundance of *Anopheles gambiae* peaks in the rainy season whereas *Anopheles funestus* abundance peaks in the dry season. Socioeconomic factors may also affect effectiveness (housing quality, for example, can affect the length of time indoor residual spraying has an effect [244, 245]). Interventions may also have a greater impact when combined [246, 247].

6.3.1 Simulated data example 2: treatment effect interactions

To provide a concrete example of the limitations of a model that does not allow for such interactions, I again simulated data where the true effect of an intervention was known, fit a model to this data and compared true and predicted effects. As in the previous example, the response variable was designed to represent malaria prevalence in Madagascar and the treatment was a binary variable representing whether IRS was carried out in a location. In this example, I considered two scenarios for the effect of IRS on prevalence. In the first scenario the size of the effect depended only on pre-intervention risk, while in the second it depended on both pre-intervention risk and rainfall.

As in the previous simulated example, the distribution of IRS was biased but this time the intervention was preferentially targeted to high baseline risk areas. Here we can imagine a study area where interventions have not previously been used, and therefore in the time period prior baseline risk and true risk were the same. Therefore, IRS distribution depending on baseline risk is consistent with the targeting of locations previously observed to be high prevalence or incidence. In this case, all the factors that influenced risk could be thought of as confounders (see the DAG in Figure 6.6). One of the causes of risk (accessibility) was not used in the model fitting process, so there was an unmeasured confounder. This more complex causal structure, where both measured and unmeasured confounders were present, was designed to more closely represent a real-world scenario than the previous simulated data example. Indeed, interventions are often deliberately targeted at presumed high risk areas and while many factors that may affect risk are measured (these are the covariates that have been included in risk mapping models throughout this thesis), these measured factors cannot account for all the variation in risk. Both these simulated datasets were fit with a model that did not include interactions between the effect of IRS and

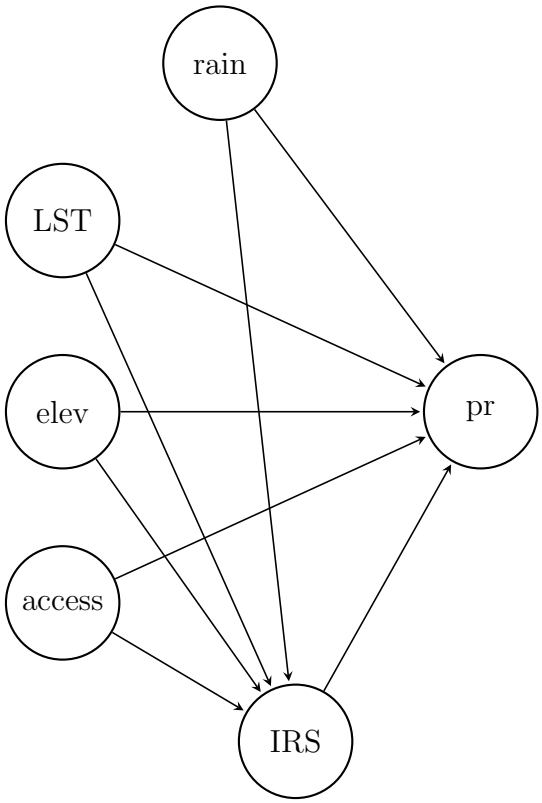


Figure 6.6: A DAG illustrating the causal relationships between variables in the second simulated data example. Here 'elev' and 'pr' represent elevation and prevalence (observed in a survey), respectively.

any other variable. The data was simulated as follows:

- **Baseline prevalence.** In both simulated datasets, the baseline prevalence was the same. As in the previous example, logit-transformed baseline prevalence was simulated as a linear combination of accessibility to cities, rainfall, elevation and land surface temperature, as in (6.3). In this case, the following coefficient

values were chosen.

$$\beta_0 = -1$$

$$\beta_{\text{access}} = 0.3$$

$$\beta_{\text{elevation}} = -0.5$$

$$\beta_{\text{rain}} = 1$$

$$\beta_{\text{LST}} = 1.5$$

The signs of these values were the same as in the previous example. The sizes of these effects were changed to reduce the relative effect of accessibility (the unmeasured confounder) on baseline prevalence. This ensured that the effect of unobserved factors on baseline risk was relatively small, which could be the case in real-world settings.

- **IRS distribution.** As in the previous example, IRS was treated as a binary variable. The probability of IRS use in a location was simulated as dependent on baseline prevalence. This probability, π_{IRS}, was modelled as

$$\pi_{\text{IRS}}(p_{\text{baseline}}) = \frac{1}{1 + \exp(-p_{\text{baseline}})}$$

so the probability of IRS use increased with baseline prevalence. The binary variable representing whether IRS had been carried out at this location, denoted X_{IRS}, was sampled as a Bernoulli random variable with probability π_{IRS} of success.

- **IRS effect.** In the first scenario, the effect of IRS was only dependent on baseline prevalence. This was modelled by assuming that IRS had a fixed effect of -0.5 on logit-transformed prevalence. Therefore, the prevalence, p, at a given

location was modelled as

$$\text{logit}(p) = \text{logit}(p_{\text{baseline}}) - 0.5X_{\text{IRS}}. \tag{6.7}$$

In the second scenario, the effect of IRS was dependent on both baseline prevalence and rainfall, with a higher effect where rainfall was higher. This was intended to simulate IRS effectiveness being moderated by environmental factors, rather than simply by baseline prevalence. This effect was modelled as,

$$\text{logit}(p) = \text{logit}(p_{\text{baseline}}) - \frac{X_{\text{rain}} - \text{rain}_{\text{min}}}{\text{rain}_{\text{max}} - \text{rain}_{\text{min}}} X_{\text{IRS}} \tag{6.8}$$

where rain_{max} and rain_{min} were the maximum and minimum values of rainfall across the country. This meant the effect of IRS increased from having no effect to reducing logit-transformed prevalence by 1 as rainfall increased. The maximum effect size of 1 was chosen so that the average effect of IRS was similar in both scenarios.

- **Observed data.** 300 pixels were randomly sampled to represent survey sites and at each site the number of individuals in the survey was an integer sampled uniformly between 10 and 40. Again, in a location i the number of individuals with infection observed, Y_i, out of the total tested, N_i, was a realisation of a binomial variable,

$$Y_i \sim \text{Binomial}(N_i, p_i)$$

where the underlying prevalence at this location, p_i, was calculated using (6.7) in scenario 1 and (6.8) in scenario 2. Figures 6.7 and 6.8 show the observed prevalence and IRS coverage in scenarios 1 and 2.

Before describing the model applied to these two datasets, it is worth considering the causal relationships between the different variables. The relationships between

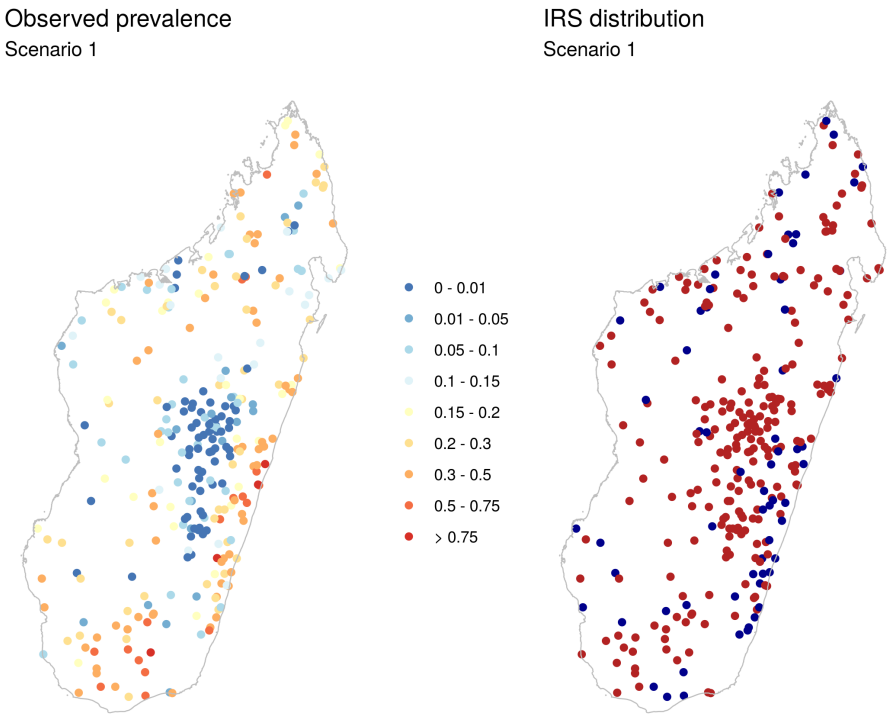


Figure 6.7: Observed prevalence (left) and IRS (right) in scenario 1, simulated data example 2. For IRS distribution, blue points represents locations where IRS was used and red where it was not.

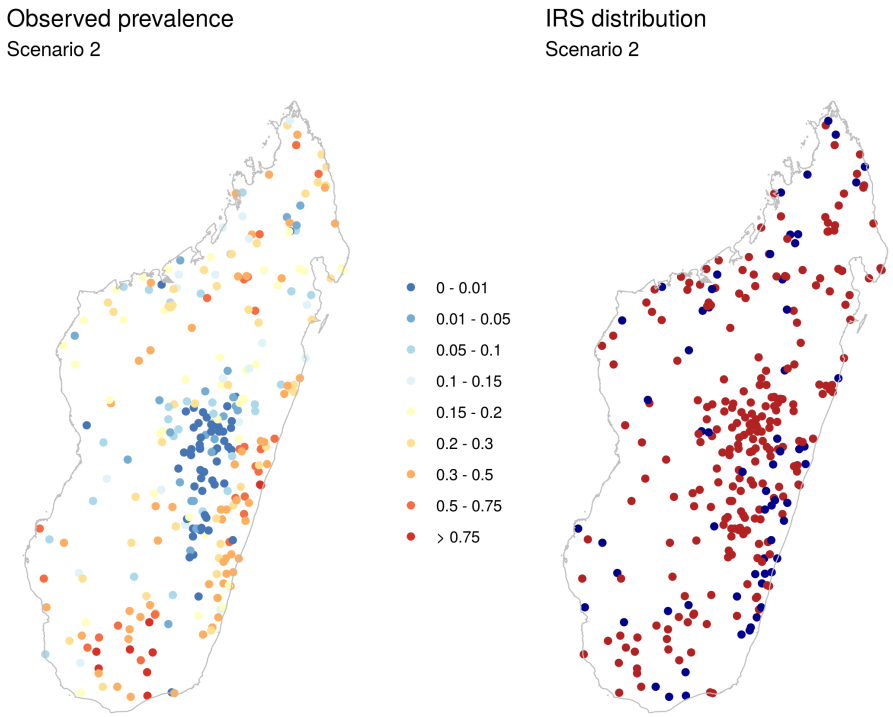


Figure 6.8: Observed prevalence (left) and IRS (right) in scenario 2, simulated data example 2. For IRS distribution, blue points represents locations where IRS was used and red where it was not.

variables in this example are shown in the DAG in Figure 6.6. Note that baseline prevalence is not included because it is a deterministic combination of other variables.

Clearly, to satisfy the unconfoundedness condition, all of accessibility, elevation, LST and rain must be observed as these are all causes of both IRS use and prevalence. In the subsequent model fitting stage accessibility is *not* included as an explanatory variable. We are therefore in a similar situation to the previous example, with an unobserved confounding variable present, and therefore we would expect any estimate of the causal effect of IRS to be biased. However, unlike in the previous example, here accessibility has a relatively small effect on baseline prevalence and therefore a small effect on both IRS distribution and observed prevalence. Therefore, we would expect the bias in the estimates to be smaller in this example. Hence, in addition to illustrating the effect of interactions between treatment effectiveness and other variables, this example also shows the effect of milder confounding (when compared to the previous example) on causal estimates.

In each scenario, the observed data was fit with a Bayesian geospatial model as described in (6.5) and (6.6) in Section 6.2.1. Here the covariates available to the model in both scenarios were elevation, rainfall, land surface temperature and IRS use. The accessibility variable was not used in the fitting process. After fitting the model, predictions for the effectiveness of IRS were made in each scenario in the same way as in the previous simulated data example.

Figure 6.9 shows the true and predicted effect of IRS in the two scenarios. We can see that in scenario 1 (where, apart from missing the accessibility variable, the model is well specified), the model predicts the spatial pattern of effectiveness very well, although the size of the effect is slightly underestimated. In scenario 2, however, we can see a clear difference between the true and predicted surfaces. The true effect was, by construction, largest in areas where both rainfall and baseline prevalence was high, predominantly on the east coast. In other high transmission areas, such as

in the southwest, the true effect was much smaller due to these areas experiencing less rainfall. The model was unable to capture these patterns and instead predicts a similar effect of IRS for all similar prevalence regions, which overestimates the effectiveness when rainfall is low and underestimates effectiveness when it is high. These results are expected, as the model has no way to incorporate interactions between IRS effectiveness and any other variable but this example illustrates that if there are variables that significantly moderate intervention effectiveness, then more flexible models will be needed to accurately capture variation in effectiveness.

6.3.2 Gaussian process models

As we have seen, if we suspect that the effectiveness of an intervention might be dependent on or moderated by other measured variables, we may need a model that allows these interactions. In the absence of knowledge about which variables impact intervention effectiveness or the nature of this relationship, a non-parametric approach would be to model the effect of covariates and interventions using a Gaussian process. This allows for arbitrary interactions between variables and functional forms, although this comes at the cost of being fairly data intensive.

I fit a model of this form to the simulated data from the previous section and investigated how well intervention effectiveness was inferred in scenarios 1 and 2. Again, let $(Y_1, N_1), ...(Y_{300}, N_{300})$ be the number of individuals infected and the number tested at the different survey locations. In this case, the prevalence at the location, s_i, of survey i was modelled as the sum of a Gaussian process over the covariates (again not including accessibility) and a spatial Gaussian process,

$$\text{logit}(p_i) = \beta_0 + g(X_i) + f(s_i) \tag{6.9}$$

where g, f were Gaussian processes and X_i were the values of elevation, rainfall, LST

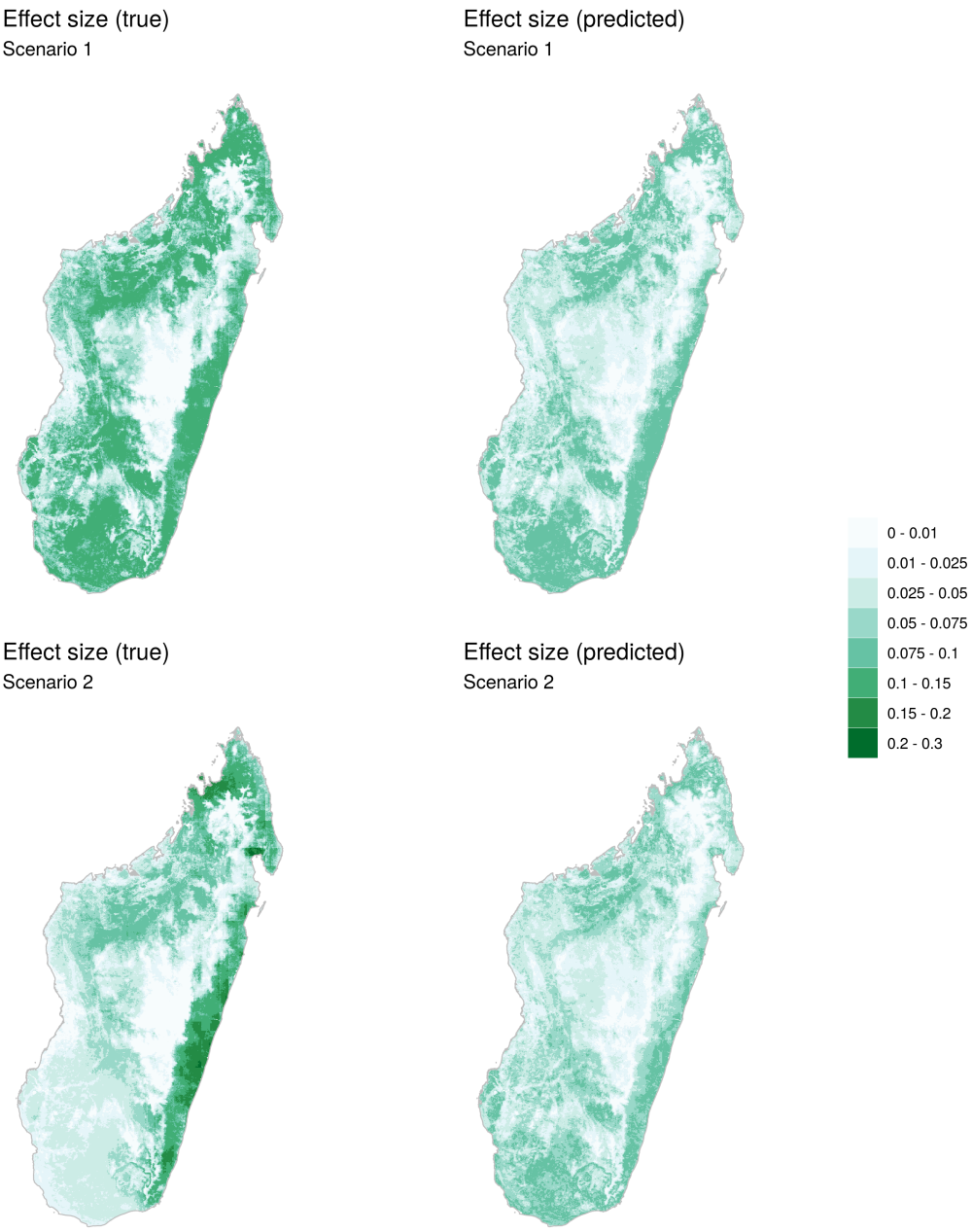


Figure 6.9: True (left) and predicted (right) effect of IRS in scenarios 1 and 2 of simulated data example 2.

and IRS use at this location. The spatial Gaussian process f had Mátern covariance structure and g had a radial basis function kernel. Figure 6.10 shows the true and predicted effects in these two scenarios when using this model. In both scenarios, the model captures the variation in the effect fairly well. In scenario 2, in particular, the model accurately predicted that the effect was higher on the east coast compared to the west and southwest (in contrast to the model in the previous section). However, we also see areas on the southwest coast where the predicted effect is very small (underestimating the true effect). This could be the effect of the Gaussian process reverting to the mean when there are few observations nearby to inform it.

These results suggest that using a Gaussian process over covariate space may be a useful approach for mapping intervention effectiveness. However, applying these ideas to real world data could be challenging for a number of reasons. Firstly, while I have considered a binary intervention variable in these examples, in practice these variables are more likely to be continuous, representing the coverage of that intervention in a location (for example, the proportion of buildings sprayed with insecticide or individuals with access to ITNs). Making predictions about the effect of continuous treatment variables is more challenging, as the same number of data points would have to inform potential outcomes for a range of possible treatment values, rather than just the outcomes with and without the intervention. The Gaussian process model is also more reliant, compared to a parametric model, on observations of similar locations (in terms of covariate values) with different intervention coverage in order to estimate effect sizes. Where intervention coverage is strongly associated with baseline risk, there may not always be enough variation in coverage to inform such estimates. For example, Figure 6.11 shows the proportion of individuals with access to an ITN at survey sites from the 2013 MIS in Madagascar [178]. We can see that in the highlands, where prevalence is low, ITN coverage is also consistently low. Therefore it would likely be difficult for a Gaussian process model to learn the impact

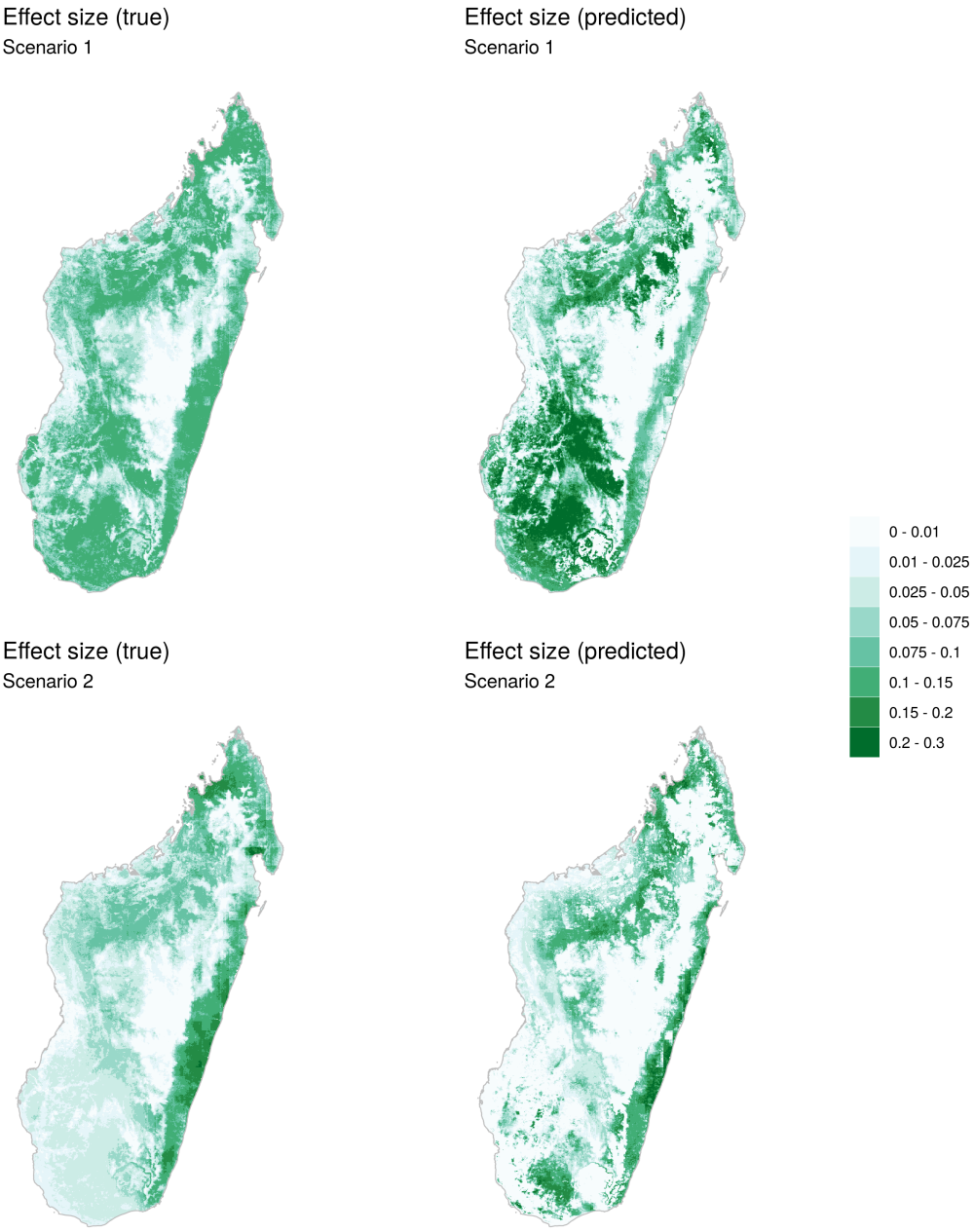


Figure 6.10: True (left) and predicted (right) effect of IRS when the data from scenarios 1 and 2 in simulated data example 2 were instead fit with the Gaussian process model.

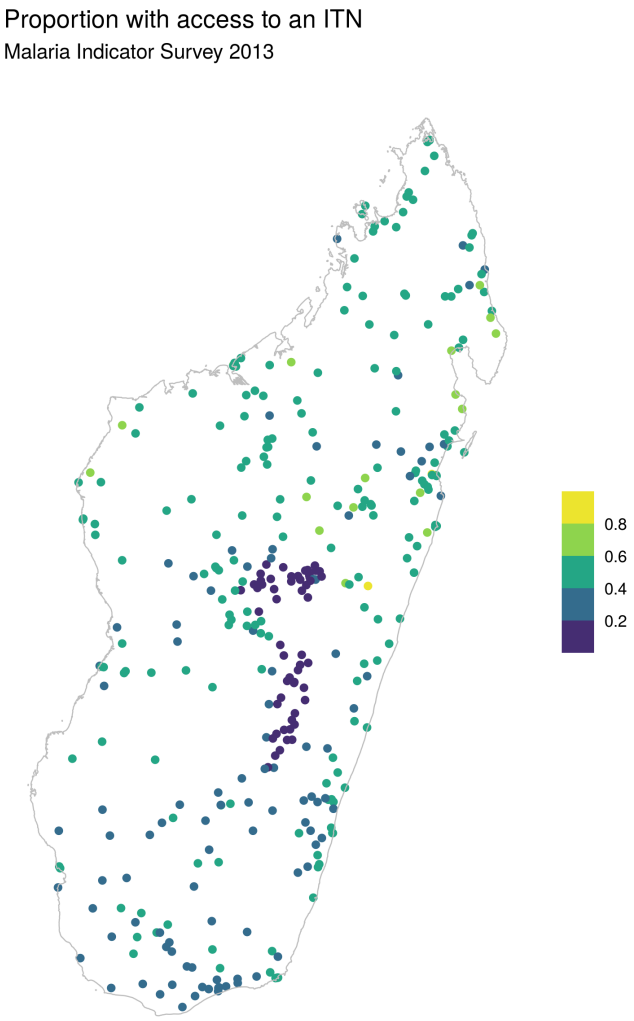


Figure 6.11: Proportion of individuals with access to an ITN in the 2013 MIS in Madagascar [178]

of high ITN coverage in areas of high elevation. In contrast, a model that simply assumes a linear effect of ITN coverage on logit prevalence (more similar to the model in the previous section) would be able to use all observations to inform this effect. This may be less of a problem, however, if policy-relevant questions concern smaller changes in coverage—for example estimating the impact of increasing coverage from 60% to 80% in a higher prevalence area.

In general, methods relying on Gaussian processes or other non-parametric approaches are also likely to suffer from the curse of dimensionality, where the number

of data points required grows quickly with the number of dimensions. In this example, with only four variables (the three covariates used in the fitting process and the binary IRS variable), the model was largely able to accurately learn the effect of IRS. However, with a larger number of covariates this performance is likely to decrease substantially. To investigate the performance of the model used in this section when more covariates are available, I refit the model to the same response data while including 5 additional covariates—an aridity index, potential evapotranspiration, night lights, distance to water and a topographic wetness index. Details of these covariates are in Table 4.1 in Chapter 4. The true and predicted effect of IRS is shown in Figure 6.12 for scenario 1 and 2 in this case. We can see that the inclusion of these additional covariates (which have no effect on either baseline prevalence or intervention effectiveness) greatly reduces the accuracy of the predictions. Very broad trends are somewhat identified by the model (for example a lower effectiveness in the central highlands) but in other areas predictions are very inaccurate. Here the causal discovery procedure outlined in the previous chapters could again be useful. Causal discovery could be used to identify confounding variables, by identifying which variables are causes of both intervention coverage and prevalence in the discovery algorithm output. These variables could then be used in a Gaussian process model or another non-parametric approach.

This approach also assumes that the variation in intervention effectiveness depends on measurable factors, such as environmental or socioeconomic variables. However, issues such as insecticide resistance could have a large effect on effectiveness but have no clear connection to these variables. Allowing intervention effectiveness to interact with location (in space or space-time) may allow these effects to be inferred but would likely require even more data.

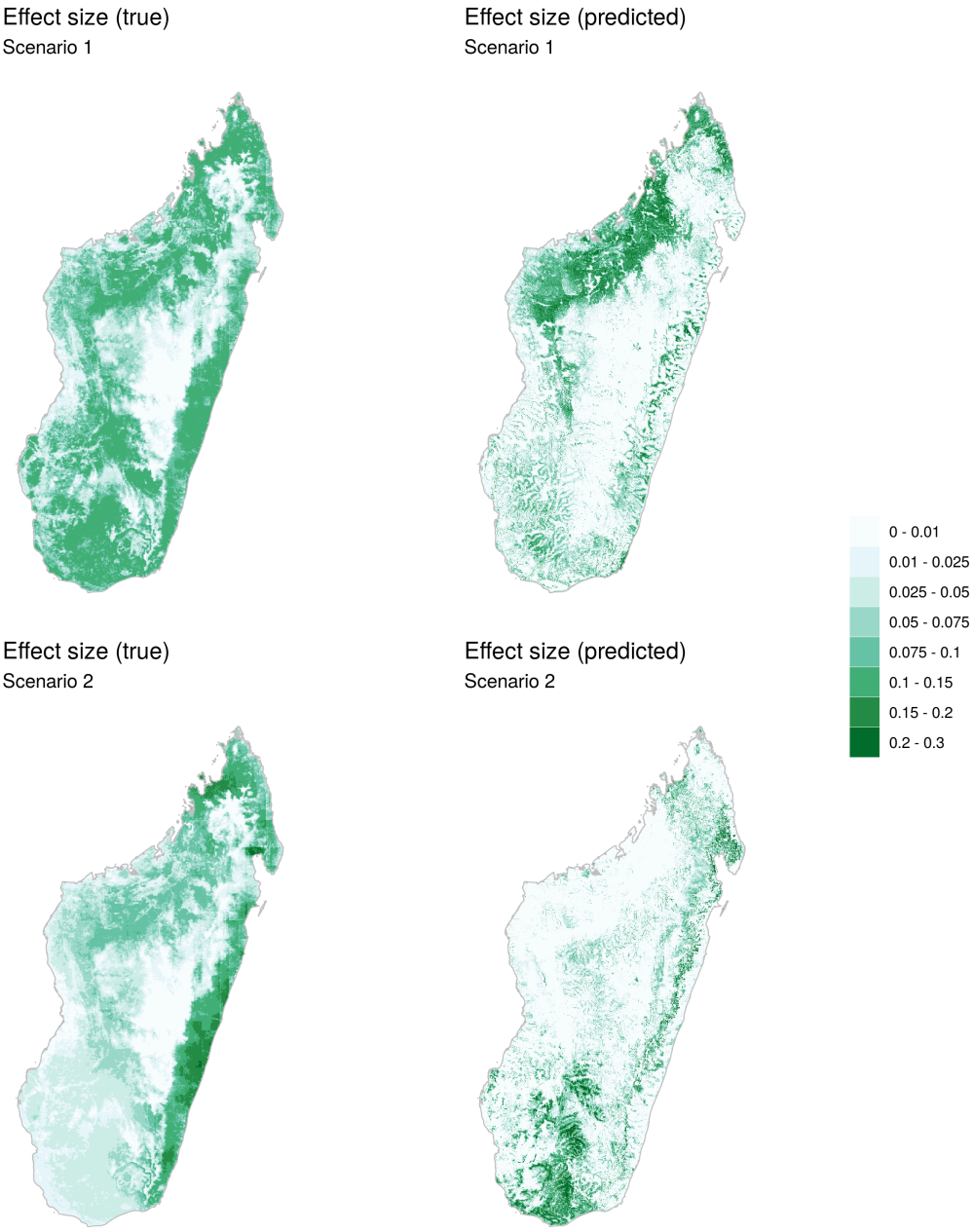


Figure 6.12: True (left) and predicted (right) effect of IRS when the data from scenarios 1 and 2 in simulated data example 2 were instead fit with the Gaussian process model with additional covariates.

6.4 Estimating individual-level effects

So far we have considered the situation where community-level data is available, such as the prevalence and intervention coverage in a location, from which we can aim to identify community-level effects. For example, estimating what the effect would be on prevalence in a location if the intervention coverage in that location is increased. We can also think about individual-level intervention effects, for example the effect of an individual using an ITN on the risk to that individual. Individual and community-level effects are clearly connected, although this relationship is not necessarily straightforward. The community-level effect of, for example, ITNs distributed to individuals in a community will include both the direct effect for those individuals and the potential indirect effects on others due to reduced vector abundance and lower transmission. In this section, I propose a method for mapping individual-level effects of interventions, although I do not implement this method.

The idea underlying this approach is that if effect sizes, rather than measures of prevalence or incidence, were observed at different locations and time points, then standard statistical approaches could be used to model these effects and predict effect sizes at new locations and times. An example of such an approach is the work of Rathmes et al. [248] on mapping anti-malarial drug effectiveness using estimates of anti-malarial drug effectiveness from trials carried out primarily in sub-Saharan Africa between 1991 and 2019. These estimates were used as the response variable in a Bayesian geospatial model, which was used to make global estimates of effectiveness in this time period.

Utilising trial data in this way avoids the problem of confounding that may be present in observational data but is naturally limited by the number of trials conducted. Rathmes et al. [248] used 232 trials carried out across multiple continents over a 29 year period. While this may be sufficient to map global and long term trends, in the absence of highly informative explanatory variables any model informed by a

dataset of this size is unlikely to capture subnational or short term (such as seasonal) trends accurately. The number of clinical trials on other interventions are likely to be similar, with two systematic reviews of ITNs and IRS effectiveness identifying 333 and 134 potentially relevant studies, of which 22 and 6 were included in the reviews, respectively [229, 228].

Instead of using trial data, I therefore propose a two step approach. The first step is to infer the individual-level effects at each location using separate analyses. Unlike in the previous section, the aim of each analysis would be to calculate an average treatment effect. This could be done in a number of ways, such as inverse probability weighting [34], propensity score matching [36, 226] or using a similar (non-spatial) Bayesian hierarchical model to those used previously and averaging individual effects. The inferred effect sizes in this step can then be used as the response variable for a geospatial model. The data to calculate individual-level effects could come from routine surveillance data or from health surveys. For example, MIS and DHS surveys typically include information about the individuals in each household, including presence of a malaria infection, whether they have access to and use an ITN, and demographic information (such as age, gender, measures of socioeconomic status and education level). If all or most of the important confounding variables are measured then this may be sufficient information to estimate an average treatment effect at each location.

There are some subtleties to consider when applying this method. For example, the definition of average treatment effect depends on the population being averaged over. Therefore, there is a question of whether it is more suitable to use the population demographics in each location to calculate this effect or to use the same national demographics for all estimates. Which one is more suitable depends on the causal question being asked. If we simply want to predict the potential effect of implementing an intervention in different areas, we would likely want to take the effect

of demographics in each area into account. On the other hand, we might be interested in the associations learned in the second step, for example whether any of the explanatory variables can explain differences in effectiveness when controlling for demographics of different areas. In this case, using the demographic distributions in the entire population may be more suitable.

There are a number of benefits to this method when compared to estimating community-level effects as discussed in the previous section. Firstly, treatments applied to the individual (such as drug treatment, ITN access for an individual or IRS for a household) can often be treated as binary variables (rather than coverage proportion as on the community-level), which are generally more straightforward to model. Furthermore, even in locations with particularly high or low intervention coverage, it is likely that individuals with and without the intervention will be observed. This method also allows for learning spatiotemporal trends in treatment effectiveness that do not correlate with any explanatory variables (through the spatial or spatiotemporal autocorrelation term in the model in the second step). For example, growing insecticide resistance in a location could be seen in such a model as a reduction in intervention effectiveness, with this reduction captured by the spatiotemporal autocorrelation term. Indeed, in theory if enough locations were observed then the model in the second step would not necessarily need any explanatory variables.

One limitation of this model is the number of observations required at each location to accurately estimate the average treatment effect. The average sample size in each location in the 2013 MIS in Madagascar was 25.0. Larger sample sizes may be needed to estimate effectiveness at each location with enough precision to be useful. However, routine surveillance data may be a suitable source of information if intervention access and other demographic variables are recorded with cases. Additionally, this method assumes that all confounding variables are observed, as unobserved confounders could significantly bias effect estimates. Mapping the effect of

interventions on individuals could be useful for operational planning as it can identify where increases in intervention coverage could have the biggest impact. However, individual-level effects likely depend on current coverage and do not reflect the indirect protection of these interventions.

6.5 Model validation

Given the challenges of estimating causal effects, validating the results of these models is crucial. If effect estimates from RCTs at different locations and time points are available then these can be compared to the predicted effects. However, as we have discussed the number of RCTs on malaria control methods is fairly small. Indeed, in application where many such trials had been carried out, these effect sizes could be modelled directly instead, as in the work of Rathmes et al. [248]. We are therefore usually in the situation where primarily observational data is available. Validating estimates of causal effects with this data is less straightforward.

An important point to note is that predictions of causal effects cannot be completely validated using observational data alone. To understand why, consider estimating the effect of treatment T on response variable Y given explanatory variables X. Any regression model fit to this data is learning the associations between variables and so the prediction made by this model for $T = t, X = x$ is some function of the estimated conditional distribution $Y|T = t, X = x$. This is typically the full distribution in a Bayesian model or the conditional expectation, $\mathrm{E}[Y|T = t, X = x]$, otherwise. This is a statement about the values of Y in the subset of the population in which $T = t$ and $X = x$. As in any predictive model, we can use techniques such as cross-validation to assess the accuracy of these models, in other words how well the model predicts this conditional distribution. However, these techniques cannot assess whether the conditional distribution is equal to the manipulated distribution

$Y|\text{do}(T = t), X = x$, which is the distribution needed to predict the effect of interventions or counterfactual scenarios. As we saw in Section 6.2.1, if an unmeasured confounder exists then the conditional distribution may be learned effectively but this will not necessarily represent the manipulated distribution. In this case, no matter how well the conditional distribution was learned, it could not be used to make causal predictions.

Detecting the existence of unmeasured confounders using only observational data is challenging. Given only joint observations of the treatment, response and explanatory variables, detecting an unmeasured confounder is in general not possible. However, with the addition of contextual information it may be possible to identify situations where the existence of unmeasured confounders is likely. For example, suppose a drug designed to treat high blood pressure is known or strongly believed to have at worst no effect, that is it cannot increase blood pressure. If a model predicts that this drug will increase blood pressure, and the model performs well when using standard validation methods, then we may suspect that this model formulation does not correctly capture the causal effect, perhaps due to the existence of unmeasured confounders. However, this reasoning will only detect unmeasured confounders that substantially bias estimates (to the point where the inferred effect is not plausible) and requires a strong prior belief about the nature of the causal effect. A somewhat similar approach (which again makes use of contextual information) that may detect more subtle confounding is the use of a negative control [241]. A negative control is an alternative outcome variable, Y', which is similar to the response variable of interest, Y, but is known to be unaffected by the treatment, T. The analysis is then repeated using this negative control as the response variable. If a non-zero causal effect is inferred (e.g. the 95% credible interval does not contain zero) of the treatment on Y' then this suggests that there are confounding variables between Y' and T that have not been adequately controlled for. Since Y is similar to Y', it is plausible that these

variables also confound the relationship between T and Y. Figure 6.13 illustrates this idea. This approach has been used to investigate estimates of the effectiveness of flu vaccinations [249, 250]. Estimates of the reduction in mortality in the elderly due to these vaccinations are often unexpectedly high [251, 252]. However, Jackson et al. [249] showed that when these analyses were repeated with mortality *before* flu season replacing the original response variable, mortality during flu season, similar effects were inferred. That is, flu vaccinations appeared to substantially reduce pre-flu season mortality. This implausible result suggests that flu vaccination and pre-flu season mortality were confounded in these analyses (possibly by general health, as more healthy individuals were more able to travel to receive the vaccine) and similar confounding could have biased the original estimates. Mortality before flu season can be considered a negative control, as vaccinations would be expected to have no effect on mortality before flu season and we would expect there to be similar confounders between vaccination and both pre-season mortality and mortality during flu season. Identifying a suitable negative control can be challenging, although for treatments which have distinct start time (such as an IRS campaign), the response variable prior to the treatment may be suitable, as in the vaccination example. Negative controls are useful for highlighting the possibility of unmeasured confounding variables. However, even if the analysis finds no effect of the treatment on the negative control, this does not necessarily guarantee that relationship between the treatment and true response is not confounded.

Another important tool for validating effect estimates is fake data simulation. By simulating data, the true effect of treatment is known and we can investigate whether the model correctly recovers this effect—as was done in the simulated data examples in this chapter. Fake data simulation is commonly used in Bayesian modelling to investigate the inference of latent parameters, uncertainty calibration and limits of the model under misspecification [253]. These simulations are even more

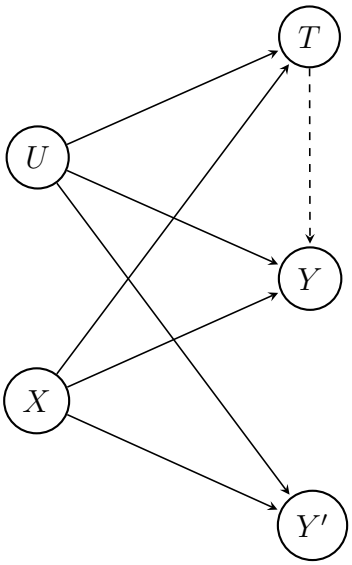


Figure 6.13: A DAG illustrating the concept of a negative control. The relationship of interest is the effect of T on Y, and Y' is a negative control. X are observed confounders that are controlled for and U are unobserved confounding variables that may or may not exist. If a non-zero causal effect of T on Y' is inferred it suggests variables U may exist.

important when estimating causal effects due to the limitations of observational data previously discussed. Therefore, fake data should be simulated in a wide variety of scenarios to understand the limitations of these inferences. Mechanistic models, such as EMOD [239], may be useful for simulating data without relying on the same or a similar data generating process to that assumed in the model.

Finally, in some cases 'natural experiments' may allow specific treatment effects to be estimated, which could be used to validate models. These could be instances where an intervention was applied to some individuals (or communities) in a way that is close to being random or unconfounded. For example, Fetzer and Graeber [254] estimated the effectiveness of contact tracing for COVID-19 in the UK using observational case data after a data processing error resulted in the many cases not being immediately referred to contact tracing. The cases that were not referred were geographically clustered but these areas were essentially random, and therefore the

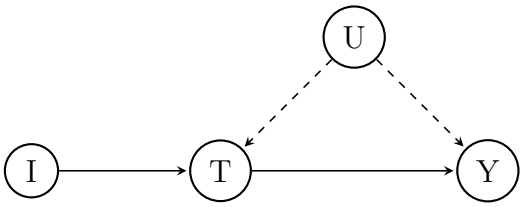


Figure 6.14: A DAG illustrating the causal relationships required to use the instrumental variables method. Here U represents unobserved confounders which may exist.

difference between subsequent case counts in areas where cases were or were not referred could be argued to represent the causal effect of contact tracing. Similar examples may occur when intervention campaigns are disrupted for other reasons. Another approach is the instrumental variables method [240]. Here some information about the effect of a treatment T on response Y can be learned even in the presence of unmeasured confounders if there exists a variable I, often referred to as the instrument, which is a cause of T, has no causal effect on Y except via T, and has no causes in common with Y. This required structure is shown in Figure 6.14. Since the relationship between I and Y is unconfounded, if there is an association between I and Y then this must reflect the causal effect of I on Y. Moreover, since I only affects Y through T, if there is a causal effect of I on Y then T must have a causal effect on Y, the latter being the effect of interest. Under further assumptions about the nature of these effects (such as linearity), the effect of T on Y can be identified. The main application of this approach in epidemiology is when I is a genetic feature, since such a variable and a disease outcome cannot be confounded by any environmental factors. These approach is known as Mendelian randomisation [255]. For example, the sickle cell trait, which is protective against malaria, has been used to estimate the effect of malaria on a number of other conditions [256, 257, 258, 259]. Estimates of effectiveness at specific locations and time points from natural experiments could supplement any effect sizes from RCTs as validation data points.

6.6 Conclusion

In this chapter I discussed methods for estimating spatial variation in community and individual-level treatment effects and the strengths and limitations of different approaches. I generally considered the situation where there is a single intervention (such as IRS) and the response variable is measured a suitable time later, or where the intervention and response could be considered constant over some time frame (such as ITN use and transmission). In practice, these concepts are complicated by a number of factors, including multiple interventions being used and the reduction of intervention effectiveness over time (due to ITN degradation, for example). Moreover, estimating causal effects from observational data requires all confounding variables to be measured (or at least all that have relatively large effects on the treatment and response variables, as discussed). This may not always be the case and it can be difficult to know when unmeasured confounding variables are present.

It is therefore perhaps unsurprising that there have been relatively few attempts to estimate spatially-varying intervention effectiveness from observational data. However, understanding the variation in these effects has potential to be an important part of operational planning and efficient allocation of limited resources in many countries. With causal inference methods, estimates can be made using routine surveillance and health survey data, meaning they can be made more frequently and over a greater area than estimates that rely on RCTs. Frequently updated estimates of ITN or IRS effectiveness could also be useful in surveillance of insecticide resistance, as reduced effectiveness could signal emerging or growing resistance in the local vector population. Similarly, reduced effectiveness of ACTs could suggest growing drug resistance. The approaches that I have discussed in this chapter are largely limited by data availability but as surveillance systems strengthen in many countries these methods will become increasingly applicable. With robust validation using simulated data and estimates from RCTs or natural experiments where available, making reliable

predictions of both community and individual-level effectiveness could be possible, although these predictions may be better suited to identifying broader trends than high resolution patterns on the scale of current prevalence and incidence maps.

Chapter 7

Conclusion

In this thesis, I developed novel statistical methods for use in spatial disease mapping, focusing on applications that combined data sources (such as prevalence data from national health surveys, routinely collected incidence data, environmental and socioeconomic covariates, and intervention coverage) in order to answer causal questions or improve estimates of malaria burden. In many cases, non-parametric methods were required to accurately model complex relationships between variables. I used kernel-methods extensively, which generate non-parametric approaches that are both computationally efficient and effective for high-dimensional problems. My main contribution is the work on causal discovery with spatial and spatiotemporal data. This method brings together many existing ideas, including constraint-based causal discovery algorithms, kernel based-independence tests, random Fourier feature approximations, and kernel-based approaches to learning relationships between variables observed at different resolutions. The result is a scalable, non-parametric causal discovery procedure that accounts for many of the idiosyncrasies of spatial and epidemiological data. The output is a graph that represents the causal structure between variables, with a measure of certainty for each causal relationship obtained through bootstrapping.

I applied this causal discovery method as a form of feature selection ('causal feature selection') that retained features that were inferred to have a direct causal relationship with the response variable. This method was non-parametric and model-free.

The results in Chapter 5 suggest this has promise as a feature selection method for Bayesian geospatial models of disease risk. An important aspect of these results was the stability of the causal feature selection method—the ranking of variables in terms of importance by the causal discovery procedure was highly consistent across different training sets. In contrast, the classical feature selection method implemented, spike-and-slab regression, was much less consistent. Many other classical selection methods for linear and logistic regression are also known to be unstable [216, 219, 214]. The stability of the causal selection method suggests that the features were selected based on meaningful underlying relationships (rather than chance associations in each training set), in which case we would expect more reliable predictions from models using this selection strategy compared to using less stable selection methods. Indeed, the predictive performance of the two models considered in Chapter 5 was as good or better when using causal feature selection compared to spike-and-slab regression. When simply including all features (i.e., performing no feature selection), in higher data settings the linear model performed better than when using causal selection. This suggests that in low complexity, high data settings it may be advantageous to use all features and allow the model to identify important variables during the fitting process. However, the more complex model performed better when using causal selection, as did the linear model in low data settings. Considering the other benefits of feature selection, including improved interpretability, reduced costs for collecting and maintaining covariate information, and lower computational burden, these results provide a strong case for using causal feature selection in these types of models.

This feature selection method, which is independent of the choice of model, could be used in combination with other more complex models that allow for non-linear effects or interactions between explanatory variables. Selecting features for complex models can otherwise be challenging due to the computational burden of repeatedly fitting these models with different potential feature sets, while including all available

features in such models is likely to lead to overfitting. Moreover, attempting to instead learn these effects in simpler models by creating new features via transformations of variables and subsequently performing selection can lead to enormous numbers of features to choose from (Weiss et al. [21], for example, consider more than 50 million potential covariate terms). The model-free aspect of the causal selection method would also allow for an alternative model-building workflow, in which features are selected first and a suitable model is then built.

Another application for the causal discovery procedure in spatial disease mapping is in the identification of variables that confound or moderate the relationship between disease control methods and transmission. As discussed in Chapter 6, including as many confounding variables as possible is crucial for accurately estimating intervention effectiveness. The examples in Chapter 6 also show that to fully understand spatial variation in intervention effectiveness it may be necessary to include variables that moderate the effect of the intervention and use models that are sufficiently complex to capture these effects. Including all available features in such models is likely to lead to overfitting, and in some cases could even induce collider bias. Causal discovery could be used to identify which variables to include. Other causal questions might require the identification of different causal structures—for example, investigating the direct and indirect contributions of malaria to overall mortality might involve distinguishing direct effects from those mediated by other variables. Again, this could be approached by first applying the causal discovery procedure and then using other causal inference tools (such as do-calculus) that make use of this inferred structure. Such two-step approaches have been used in parametric settings (e.g. assuming linear effects and Gaussian distributions) [217, 260]. The results of the causal discovery procedure may also be useful outputs themselves, helping to identify drivers of transmission. While the fundamental mechanisms of malaria transmission are known, the factors that drive transmission or lead to outbreaks in different set-

tings, and on what timescale these effects occur, are not necessarily clear and are of interest [261, 262, 263].

I also developed a novel method that combines routine surveillance incidence data and health survey prevalence data to make estimates of prevalence at a high temporal resolution, while accounting for potential sources of bias in the incidence data. Rates of treatment-seeking for fever and health facility catchment areas were modelled and the relative trends in the estimated incidence rates were calibrated to the more reliable prevalence estimates. This allowed for systematic biases in the incidence data (for example, due to under reporting or modelled treatment seeking rates that were consistently too high or low) to be accounted for and did not require a prescribed incidence-prevalence relationship. This method effectively utilised routine surveillance data, which has up to now rarely been incorporated into risk mapping work due to potential biases [17]. This was used to make estimates of monthly prevalence in Madagascar between 2013 and 2016, where previously only annual estimates for years in which national surveys were carried out (most recently 2011, 2013 and 2016) were available [53]. These new estimates provided more insight into both long term and seasonal trends in malaria risk across the country. Perhaps most notably, while observed prevalence was similar in the 2013 and 2016 malaria indicator surveys, the monthly estimates over the four years show substantial variation in the intervening years, with a reduction in transmission in 2014 followed by a large increase in 2015 in most of the country. Seasonal trends were also widely observed, including in the lower prevalence highland areas where transmission dynamics are often thought to be driven more by infrequent outbreaks. Understanding these trends could aid the evaluation of the effectiveness of control measures and targeting of future efforts. A more precise understanding of seasonal patterns in transmission, in particular, could help with timing the implementation of control methods such as IRS and seasonal chemoprevention to have the greatest impact. This method also allows estimates to

be updated as more routine surveillance data is collected, and therefore up-to-date risk maps can be produced more frequently than those based only on national health survey data. It is also likely that future risk mapping efforts will aim to incorporate more routine surveillance data in increasingly complex models (see [186, 189] for examples of method development in this area). My work could serve as a starting point and an example of an approach that learns useful information from potentially biased data sources.

In Chapter 6, I discussed the problem of estimating the effectiveness of malaria control methods, particularly in order to understand sub-national variation in effectiveness. Although I suggested possible approaches, this may be a difficult problem given the data currently available, as there may be confounding variables that are not measured or insufficient data to inform the flexible models needed to capture heterogeneity in effectiveness. Therefore these methods may be most applicable in countries where routine surveillance systems are strong, primarily outside of sub-Saharan Africa [264, 96, 17]. The relative effectiveness of different interventions may be of particular interest in areas where *P. vivax* is endemic (primarily in Asia and the Americas), as the distinct biology of this parasite species (in particular the dormant liver stage) is believed to impact the effectiveness of control methods and the efforts needed to achieve elimination [265, 266]. These methods may also become more feasible in sub-Saharan Africa as surveillance systems continue to strengthen [17]. In such analyses, and whenever causal questions are of interest (such as in Bhatt et al. [26], Weiss et al. [32]), it may be beneficial to describe these problems using causal language, for example by specifying the estimand in terms of potential outcomes, target trials or manipulated distributions (as discussed in Chapter 6) or describing the presumed causal structure between variables using a DAG. This could help clarify what exactly the effect of interest is (for example community or individual-level, or whether it is a population average or focused on some subset) and what assumptions

have been made to derive an estimate. Robust validation, including fake data simulation, negative controls and comparisons to estimates from clinical trials or natural experiments, is also important to assess whether these analyses have learned causal effects.

Over the past 20 years, model-based estimates of malaria burden have become increasingly sophisticated and are now valuable tools for decision-making on both a national and global-scale [14, 96, 27, 29, 267, 13]. In this thesis, I both developed new methodology and demonstrated how these techniques can be applied to answer policy-relevant questions. Therefore, I hope that this work can be used to aid disease mapping and decision-making in the short term as well as being the basis for more method development.

Appendix

Features sets found in the first step of the causal feature selection procedure described in Chapter 4 are listed below. A number following a dynamic covariate indicates a time lag in months (for example, Rainfall 1 indicates the Rainfall covariates with a one month time lag). "dtw" represents distance to water and "AI" the aridity index.

- Accessibility, dtw, Rainfall 0
- Accessibility, AI, dtw, Rainfall 0
- Accessibility, dtw, LST night1, Rainfall 0
- Accessibility, AI, dtw, Slope, Rainfall 0
- Accessibility, AI, dtw, LST night1, Rainfall 0
- Accessibility, dtw, LST night 1, Rainfall0, TCB 2
- Accessibility, AI, dtw, Slope, LST night 1, Rainfall 0
- Accessibility, AI, dtw, LST night 1, Rainfall 0, TCB 2
- Accessibility, dtw, LST night 1, Rainfall 0, TCB 2, TSI Pf 2
- Accessibility, AI, dtw, Slope, LST night 1, Rainfall 0, TCB 2
- Accessibility, AI, dtw, LST night 1, Rainfall 0, TCB 2, TSI Pf 2
- Accessibility, AI, dtw, Night lights, PET, Slope, TWI, Rainfall 0

– Accessibility, AI, dtw, Slope, LST night 1, Rainfall0, TCB 2, TSI Pf 2

– Accessibility, AI, dtw, Night lights, PET, Slope, TWI, LST night 1, Rainfall 0

– Accessibility, AI, dtw, Night lights, PET, Slope, TWI, LST night 1, Rainfall 0, TCB 2

– Accessibility, AI, dtw, Night lights, PET, Slope, TWI, LST night 1, Rainfall 0, TCB 2, TSI Pf 2

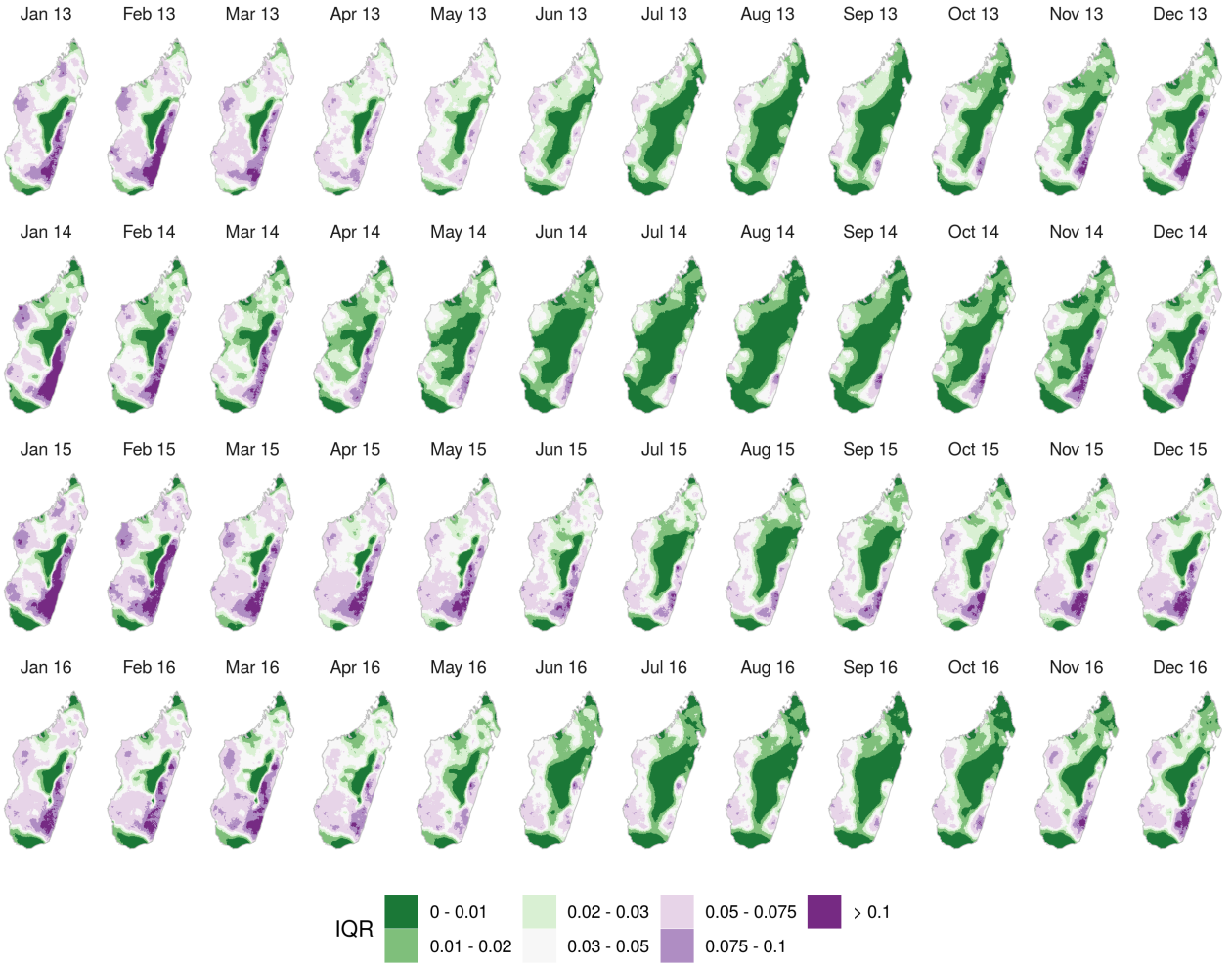


Figure A1: Monthly interquartile ranges.

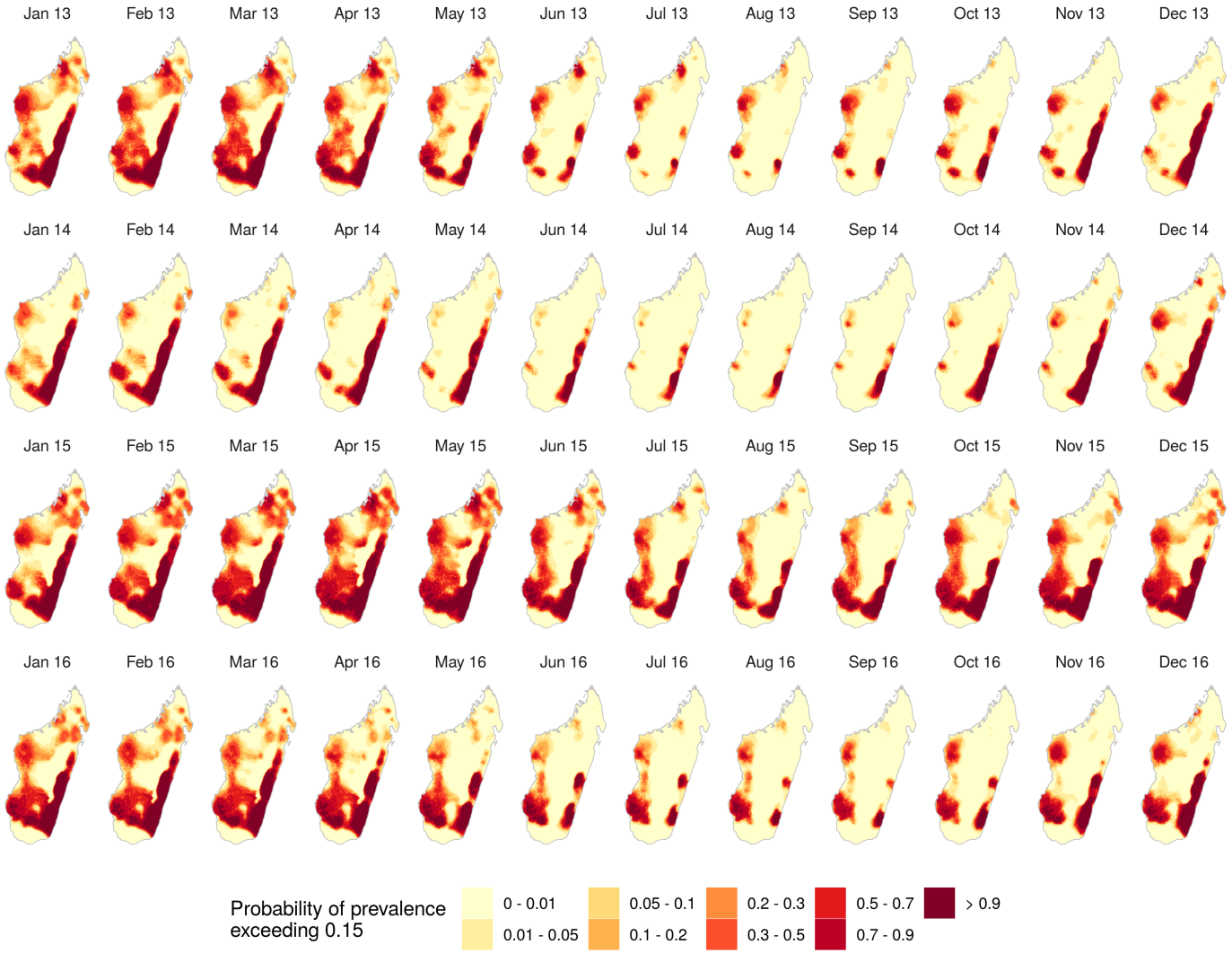


Figure A2: Probability of prevalence exceeding 0.15 each month.

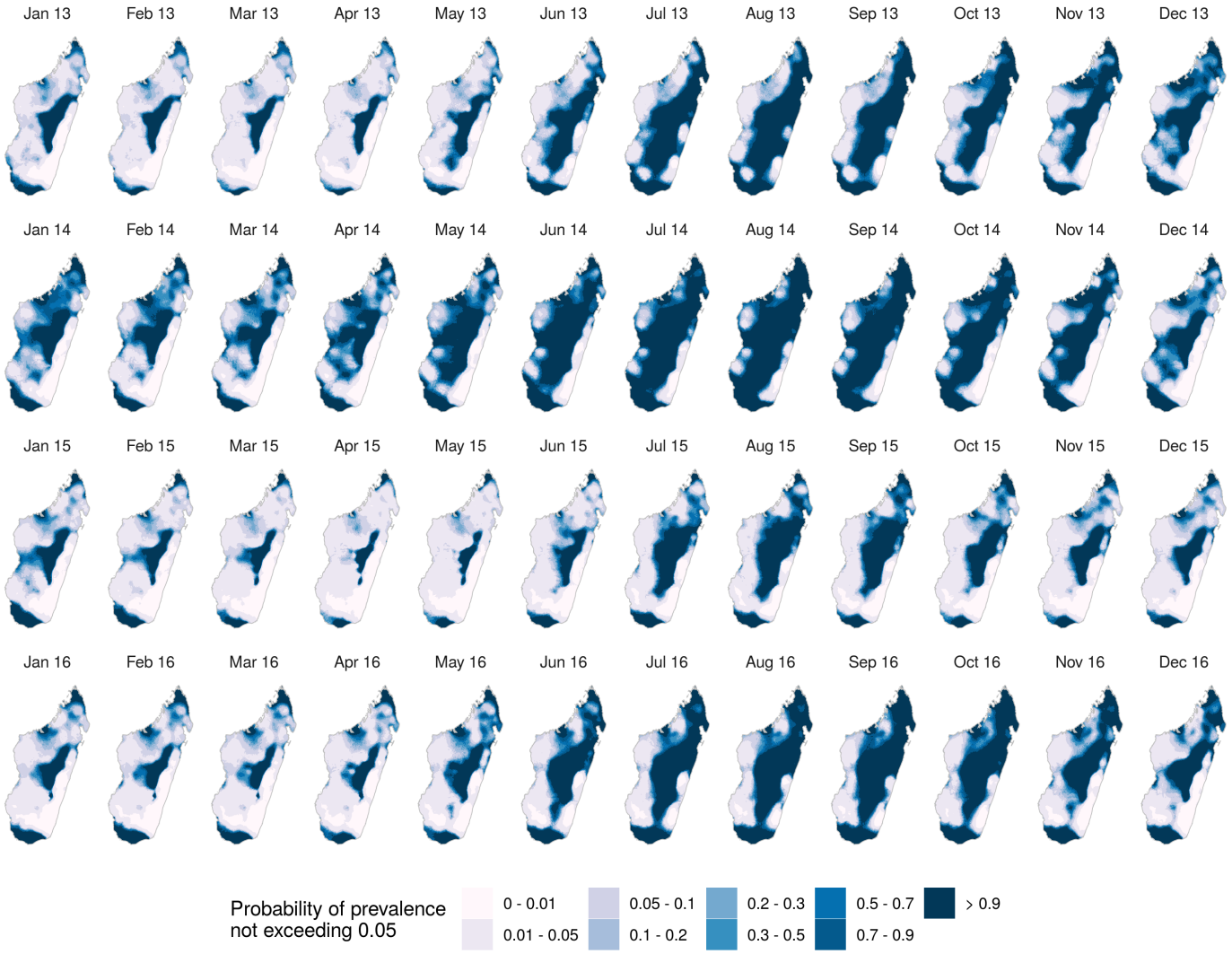


Figure A3: Probability of prevalence not exceeding 0.05 each month.

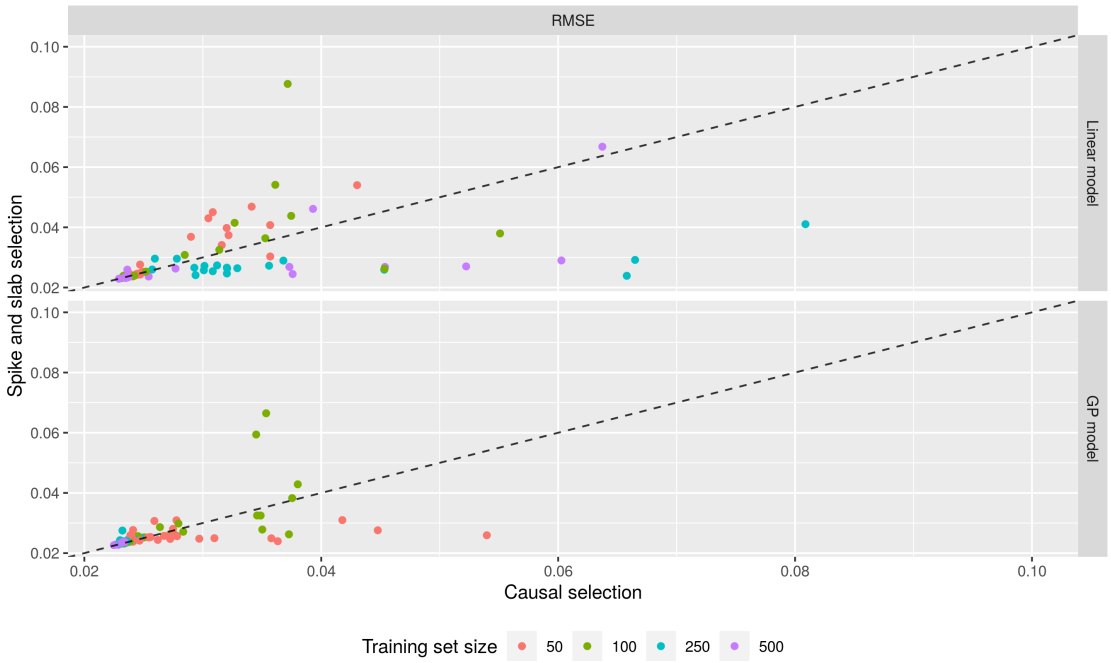


Figure A4: Comparison of root mean square error between predicted and true incidence rates for the linear (top) and Gaussian process (bottom) models when using spike-and-slab and causal selection.

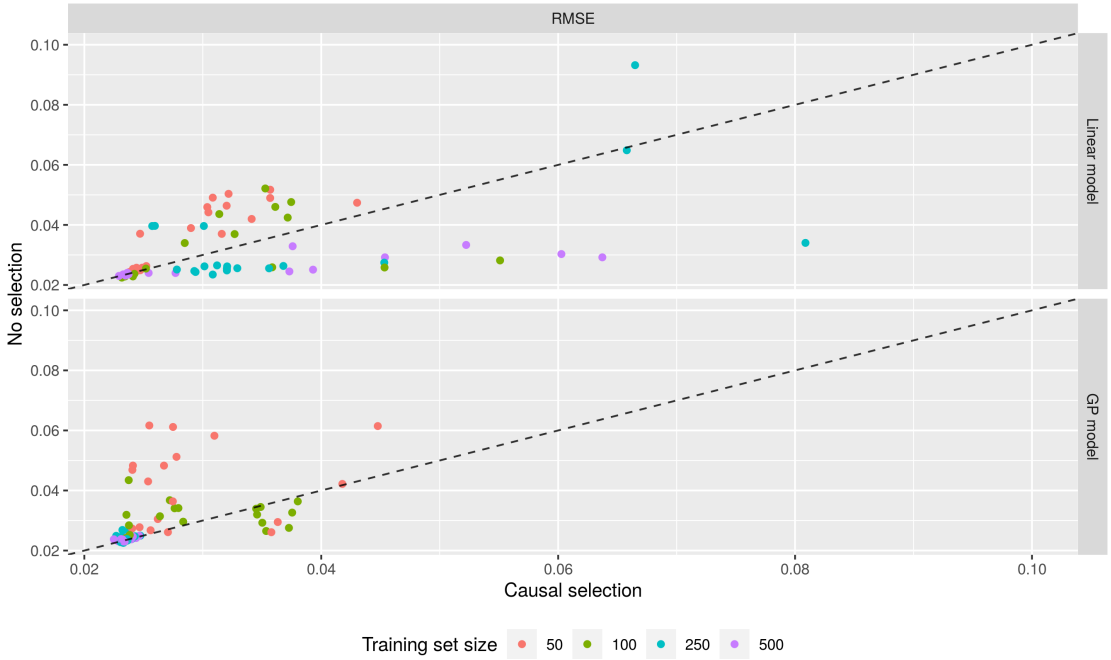


Figure A5: Comparison of root mean square error between predicted and true incidence rates for the linear (top) and Gaussian process (bottom) models when using no selection and causal selection.

Bibliography

[1] Rohan Arambepola, Peter Gething, and Ewan Cameron. Nonparametric causal feature selection for spatiotemporal risk mapping of malaria incidence in Madagascar. *arXiv preprint arXiv:2001.07745*, 2020.

[2] Rohan Arambepola, Suzanne H. Keddie, Emma L. Collins, Katherine A. Twohig, Punam Amratia, Amelia Bertozzi-Villa, Elisabeth G. Chestnutt, Joseph Harris, Justin Millar, Jennifer Rozier, Susan F. Rumisha, Tasmin L. Symons, Camilo Vargas-Ruiz, Mauricette Andriamananjara, Saraha Rabeherisoa, Arsène C. Ratsimbasoa, Rosalind E. Howes, Daniel J. Weiss, Peter W. Gething, and Ewan Cameron. Spatiotemporal mapping of malaria prevalence in Madagascar using routine surveillance and health survey data. *Scientific Reports*, 10(1):18129, Oct 2020.

[3] Tom Koch. *Cartographies of disease: maps, mapping, and medicine*. Esri Press Redlands, CA, 2005.

[4] SD Walter. Disease mapping: a historical perspective. *Spatial epidemiology: methods and applications*, pages 223–239, 2000.

[5] Lloyd G Stevenson. Putting disease on the map: the early use of spot maps in the study of yellow fever. *Journal of the History of Medicine and Allied Sciences*, pages 226–261, 1965.

[6] Donald Cameron and Ian G Jones. John Snow, the Broad Street pump and modern epidemiology. *International journal of epidemiology*, 12(4):393–396, 1983.

[7] P Stocks. Distribution in England and Wales of cancer of various sites. *Ann Rep Br Empire Cancer Campaign*, 13:239–280, 1936.

[8] P Stocks. Distribution in England and Wales of cancer of various organs. *16th Annual Report of the British Empire Cancer Campaign*, page 308, 1939.

[9] TJ Mason, FW McKay, R Hoover, WJ Blot, and JF Fraumeni. Atlas of cancer mortality for US counties: 1950–69. Department of Health, Education and Welfare, publication no. 75-789. *Washington, DC: US Government Printing Office*, 1975.

[10] Ursula Dalrymple, Bonnie Mappin, and Peter W Gething. Malaria mapping: understanding the global endemicity of *falciparum* and *vivax* malaria. *BMC medicine*, 13(1):1–8, 2015.

[11] AJ Lysenko and IN Semashko. Geography of malaria. a medico-geographic profile of an ancient disease. *Itogi Nauki: Medicinskaja Geografija*, 25:146, 1968.

[12] Simon I Hay, Carlos A Guerra, Andrew J Tatem, Abdisalan M Noor, and Robert W Snow. The global distribution and population at risk of malaria: past, present, and future. *The Lancet infectious diseases*, 4(6):327–336, 2004.

[13] Ludovica Ghilardi, George Okello, Linda Nyondo-Mipando, Chawanangwa Mahebere Chirambo, Fathy Malongo, Jenna Hoyt, Jieun Lee, Yovitha Sedekia, Justin Parkhurst, Jo Lines, Robert W Snow, Caroline A Lynch, and Jayne Webster. How useful are malaria risk maps at the country level? Perceptions

of decision-makers in Kenya, Malawi and the Democratic Republic of Congo. *Malaria Journal*, 19(1):1–17, 2020.

[14] Judy A Omumbo, Abdisalan M Noor, Ibrahima S Fall, and Robert W Snow. How well are malaria maps used to design and finance malaria control in Africa? *PloS one*, 8(1):e53198, 2013.

[15] President's Malaria Initiative. Madagascar malaria operational plan financial year 2018, 2017.

[16] President's Malaria Initiative. Madagascar malaria operational plan financial year 2019, 2018.

[17] Victor A Alegana, Emelda A Okiro, and Robert W Snow. Routine data for malaria morbidity estimation in Africa: challenges and prospects. *BMC Medicine*, 18(1):1–13, 2020.

[18] Hugh JW Sturrock, Adam F Bennett, Alemayehu Midekisa, Roly D Gosling, Peter W Gething, and Bryan Greenhouse. Mapping malaria risk in low transmission settings: challenges and opportunities. *Trends in parasitology*, 32(8): 635–645, 2016.

[19] Peter J Diggle, Jonathan A Tawn, and Rana A Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47 (3):299–350, 1998.

[20] Peter J Diggle and Paulo J Riberio Jr. *Model-based Geostatistics*. Springer, 2007.

[21] Daniel J Weiss, Bonnie Mappin, Ursula Dalrymple, Samir Bhatt, Ewan Cameron, Simon I Hay, and Peter W Gething. Re-examining environmental

correlates of *plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach. *Malaria journal*, 14(1):68, 2015.

[22] UNICEF MICS. `https://mics.unicef.org/`. Accessed: 2020-12-30.

[23] The DHS Program. Funded by USAID. `https://dhsprogram.com/`. Accessed: 2020-12-30.

[24] Simon I Hay, Carlos A Guerra, Peter W Gething, Anand P Patil, Andrew J Tatem, Abdisalan M Noor, Caroline W Kabaria, Bui H Manh, Iqbal RF Elyazar, Simon Brooker, David L Smith, Rana A Moyeed, and Robert W Snow. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med*, 6(3): e1000048, 2009.

[25] Samir Bhatt, Peter W. Gething, Oliver J. Brady, Jane P. Messina, Andrew W. Farlow, Catherine L. Moyes, John M. Drake, John S. Brownstein, Anne G. Hoen, Osman Sankoh, Monica F. Myers, Dylan B. George, Thomas Jaenisch, G. R. William Wint, Cameron P. Simmons, Thomas W. Scott, Jeremy J. Farrar, and Simon I. Hay. The global distribution and burden of dengue. *Nature*, 496 (7446):504–507, Apr 2013.

[26] S. Bhatt, D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. E. Battle, C. L. Moyes, A. Henry, P. A. Eckhoff, E. A. Wenger, O. Briët, M. A. Penny, T. A. Smith, A. Bennett, J. Yukich, T. P. Eisele, J. T. Griffin, C. A. Fergus, M. Lynch, F. Lindgren, J. M. Cohen, C. L. J. Murray, D. L. Smith, S. I. Hay, R. E. Cibulskis, and P. W. Gething. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526(7572):207–211, Oct 2015.

[27] Daniel J Weiss, Tim C D Lucas, Michele Nguyen, Anita K Nandi, Donal Bisanzio, Katherine E Battle, Ewan Cameron, Katherine A Twohig, Daniel A

Pfeffer, Jennifer A Rozier, Harry S Gibson, Puja C Rao, Daniel Casey, Amelia Bertozzi-Villa, Emma L Collins, Ursula Dalrymple, Naomi Gray, Joseph R Harris, Rosalind E Howes, Sun Yun Kang, Suzanne H Keddie, Daniel May, Susan Rumisha, Michael P Thorn, Ryan Barber, Nancy Fullman, Chantal K Huynh, Xie Kulikoff, Michael J Kutz, Alan D Lopez, Ali H Mokdad, Mohsen Naghavi, Grant Nguyen, Katya Anne Shackelford, Theo Vos, Haidong Wang, David L Smith, Stephen S Lim, Christopher J L Murray, Samir Bhatt, Simon I Hay, and Peter W Gething. Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet*, 394(10195):322–331, 2019.

[28] Peter W Gething, Iqbal RF Elyazar, Catherine L Moyes, David L Smith, Katherine E Battle, Carlos A Guerra, Anand P Patil, Andrew J Tatem, Rosalind E Howes, Monica F Myers, B Dylan George, Peter Horby, Heiman F L Werheim, Ric N Price, Ivo Müeller, J Kevin Baird, and Simon I Hay. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl Trop Dis*, 6(9):e1814, 2012.

[29] Katherine E Battle, Tim C D Lucas, Michele Nguyen, Rosalind E Howes, Anita K Nandi, Katherine A Twohig, Daniel A Pfeffer, Ewan Cameron, Puja C Rao, Daniel Casey, Harry S Gibson, Jennifer A Rozier, Ursula Dalrymple, Suzanne H Keddie, Emma L Collins, Joseph R Harris, Carlos A Guerra, Michael P Thorn, Donal Bisanzio, Nancy Fullman, Chantal K Huynh, Xie Kulikoff, Michael J Kutz, Alan D Lopez, Ali H Mokdad, Mohsen Naghavi, Grant Nguyen, Katya Anne Shackelford, Theo Vos, Haidong Wang, Stephen S Lim, Christopher J L Murray, Ric N Price, J Kevin Baird, David L Smith, Samir Bhatt, Daniel J Weiss, Simon I Hay, and Peter W Gething. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and

temporal modelling study. *The Lancet*, 394(10195):332–343, 2019.

[30] Miguel Angel Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004.

[31] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[32] Daniel J Weiss, Amelia Bertozzi-Villa, Susan F Rumisha, Punam Amratia, Rohan Arambepola, Katherine E Battle, Ewan Cameron, Elisabeth Chestnutt, Harry S Gibson, Joseph Harris, Suzanne Keddie, Justin J Millar, Jennifer Rozier, Tasmin L Symons, Camilo Vargas-Ruiz, Simon I Hay, David L Smith, Pedro L Alonso, Abdisalan M Noor, Samir Bhatt, and Peter W Gething. Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. *The Lancet Infectious Diseases*, 21(1):59–69, 2021.

[33] David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

[34] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[35] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, pages 465–472, 1990.

[36] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[37] Kenneth A Bollen. Structural equation models. *Encyclopedia of biostatistics*, 7, 2005.

[38] Judea Pearl. *Causality*. Cambridge university press, 2009.

[39] David Lewis. Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer, 1973.

[40] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[41] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[42] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[43] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[44] Peter Spirtes and Christopher Meek. Learning Bayesian networks with discrete variables from data. In *KDD*, volume 1, pages 294–299, 1995.

[45] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[46] Arthur Gretton, Alexander J Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos K Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119. Citeseer, 2005.

[47] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

[48] Bernhard Scholkopf. The kernel trick for distances. *Advances in neural information processing systems*, pages 301–307, 2001.

[49] Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[50] Ewan Cameron, Katherine E Battle, Samir Bhatt, Daniel J Weiss, Donal Bisanzio, Bonnie Mappin, Ursula Dalrymple, Simon I Hay, David L Smith, Jamie T Griffin, Edward A Wenger, Philip A Eckhoff, Thomas A Smith, Melissa A Penny, and Peter W Gething. Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nature communications*, 6(1):1–10, 2015.

[51] Rupam Tripura, Thomas J. Peto, Christianne C. Veugen, Chea Nguon, Chan Davoeung, Nicola James, Mehul Dhorda, Richard J. Maude, Jureeporn Duanguppama, Krittaya Patumrat, Mallika Imwong, Lorenz von Seidlein, Martin P. Grobusch, Nicholas J. White, and Arjen M. Dondorp. Submicroscopic Plasmodium prevalence in relation to malaria incidence in 20 villages in western Cambodia. *Malaria Journal*, 16(1):56, Jan 2017.

[52] Hamtandi Magloire Natama, Eduard Rovira-Vallbona, M. Athanase Somé, Serge Henri Zango, Hermann Sorgho, Pieter Guetens, Maminata Coulibaly-Traoré, Innocent Valea, Petra F. Mens, Henk D. F. H. Schallig, Luc Kestens, Halidou Tinto, and Anna Rosanas-Urgell. Malaria incidence and prevalence during the first year of life in Nanoro, Burkina Faso: a birth-cohort study. *Malaria Journal*, 17(1):163, Apr 2018.

[53] Su Yun Kang, Katherine E. Battle, Harry S. Gibson, Arsène Ratsimbasoa, Milijaona Randrianarivelojosia, Stéphanie Ramboarina, Peter A. Zimmerman, Daniel J. Weiss, Ewan Cameron, Peter W. Gething, and Rosalind E. Howes. Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to

assess *Plasmodium falciparum* endemicity trends between 2011 and 2016. *BMC Medicine*, 16(1):71, May 2018.

[54] Rosalind E Howes, Sedera Aurélien Mioramalala, Brune Ramiranirina, Thierry Franchard, Andry Joeliarijaona Rakotorahalahy, Donal Bisanzio, Peter W Gething, Peter A Zimmerman, and Arsène Ratsimbasoa. Contemporary epidemiological overview of malaria in Madagascar: operational utility of reported routine case data for malaria control planning. *Malaria Journal*, 15(1):502, 2016.

[55] Institut National de la Statistique (INSTAT), Programme National de lutte contre le Paludisme (PNLP), Institut Pasteur de Madagascar (IPM), and ICF International. *Madagascar Malaria Indicator Survey 2016 [Enquête sur les Indicateurs du Paludisme (EIPM)]*. INSTAT, PNLP, IPM and ICF International, Calverton, 2016.

[56] Tom G Farr, Paul A Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of geophysics*, 45(2), 2007.

[57] D J Weiss, A Nelson, HS Gibson, W Temperley, S Peedell, A Lieber, M Hancher, E Poyart, S Belchior, N Fullman, B Mappin, U Dalrymple, J Rozier, T C D Lucas, R E Howes, L S Tusting, S Y Kang, E Cameron, D Bisanzio, K E Battle, S Bhatt, and P W Gething. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336, 2018.

[58] Bernhard Lehner and Petra Döll. Global Lakes and Wetlands Database GLWD. *GLWD Documentation*, 2004.

[59] Milijaona Randrianarivelojosia. History of malaria treatment in Madagascar, 2010. URL https://www.mmv.org/sites/default/files/uploads/docs/artemisinin/2010_Madagascar/History_of_Malaria_Treatment.pdf. Artemisinin Conference.

[60] Justine I Blanford, Simon Blanford, Robert George Crane, Michael E Mann, Krijn P Paaijmans, Kathleen V Schreiber, and Matthew Brian Thomas. Implications of temperature variation for malaria parasite development across Africa. *Scientific reports*, 3(1):1–11, 2013.

[61] Nicholas J Arisco, Benjamin L Rice, Luciano M Tantely, Romain Girod, Gauthier N Emile, Hervet J Randriamady, Marcia C Castro, and Christopher D Golden. Variation in *Anopheles* distribution and predictors of malaria infection risk across regions of Madagascar. *Malaria journal*, 19(1):1–13, 2020.

[62] Vincent Robert, Kate Macintyre, Joseph Keating, Jean-Francois Trape, Jean-Bernard Duchemin, McWilson Warren, and John C Beier. Malaria transmission in urban sub-Saharan Africa. *The American journal of tropical medicine and hygiene*, 68(2):169–176, 2003.

[63] Jean-François Trape, Gilles Pison, André Spiegel, Catherine Enel, and Christophe Rogier. Combating malaria in Africa. *Trends in parasitology*, 18 (5):224–230, 2002.

[64] Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019.

[65] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approxima-

tions. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[66] Sara Martino and Andrea Riebler. Integrated nested laplace approximations (inla). *Wiley StatsRef: Statistics Reference Online*, pages 1–19, 2014.

[67] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[68] Anita K Nandi, Tim CD Lucas, Rohan Arambepola, Peter Gething, and Daniel J Weiss. disaggregation: An R package for Bayesian spatial disaggregation modelling. *arXiv preprint arXiv:2001.04847*, 2020.

[69] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623, 2014.

[70] Caroline O Buckee, Amy Wesolowski, Nathan N Eagle, Elsa Hansen, and Robert W Snow. Mobile phones and malaria: modeling human and parasite travel. *Travel medicine and infectious disease*, 11(1):15–22, 2013.

[71] Koen Peeters Grietens, Charlotte Gryseels, Susan Dierickx, Melanie Bannister-Tyrrell, Suzan Trienekens, Sambunny Uk, Pisen Phoeuk, Sokha Suon, Srun Set, René Gerrets, Sarah Hoibak, Joan Muela Ribera, Susanna Hausmann-Muela, Sochantha Tho, Lies Durnez, Vincent Sluydts, Umberto d'Alessandro, Marc Coosemans, and Annette Erhart. Characterizing types of human mobility to inform differential and targeted malaria elimination strategies in Northeast Cambodia. *Scientific Reports*, 5(1):16837, Nov 2015.

[72] Nick W Ruktanonchai, Patrick DeLeenheer, Andrew J Tatem, Victor A Alegana, T Trevor Caughlin, Elisabeth zu Erbach-Schoenberg, Christopher Lourenço, Corrine W Ruktanonchai, and David L Smith. Identifying malaria transmission foci for elimination using human mobility data. *PLoS computational biology*, 12(4):e1004846, 2016.

[73] Alessandro Sorichetta, Tom J. Bird, Nick W. Ruktanonchai, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Natalia Tejedor, Ian C. Waldock, Jason D. Sadler, Andres J. Garcia, Luigi Sedda, and Andrew J. Tatem. Mapping internal connectivity through human migration in malaria endemic countries. *Scientific Data*, 3(1):160066, Aug 2016.

[74] Malaria Atlas Project. Interactive map tool. `https://malariaatlas.org/explorer/`, 2021. [Online; accessed 15-November-2020].

[75] World Health Organization. *World malaria report 2018*. World Health Organization, 2019.

[76] Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[77] Katherine E Battle, Donal Bisanzio, Harry S Gibson, Samir Bhatt, Ewan Cameron, Daniel J Weiss, Bonnie Mappin, Ursula Dalrymple, Rosalind E Howes, Simon I Hay, and Peter W Gething. Treatment-seeking rates in malaria endemic countries. *Malaria Journal*, 15(1):20, 2016.

[78] Jian-Wei Xu, Qi-Zhang Xu, Hui Liu, and Yi-Rou Zeng. Malaria treatment-seeking behaviour and related factors of Wa ethnic minority in Myanmar: a cross-sectional study. *Malaria journal*, 11(1):417, 2012.

[79] Phyo Aung Naing, Thae Maung Maung, Jaya Prasad Tripathy, Tin Oo, Khin Thet Wai, and Aung Thi. Awareness of malaria and treatment-seeking be-

haviour among persons with acute undifferentiated fever in the endemic regions of Myanmar. *Tropical Medicine and Health*, 45(1):31, 2017.

[80] Antonio Trabucco and Robert J Zomer. Global aridity index (global-aridity) and global potential evapo-transpiration (global-PET) geospatial database. *CGIAR Consortium for Spatial Information*, 2009.

[81] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160): 850–853, 2013.

[82] Mark A Friedl, Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment*, 114(1):168–182, 2010.

[83] Christopher D Elvidge, Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21):5860–5879, 2017.

[84] NASA Earth Observations. Average land surface temperature. `http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD\_LSTD\_CLIM\_M`, 2017. [Accessed Sept 2017].

[85] NASA Earth Data. Land processes distributed active archive center. `https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd43b5`, 2017. [Accessed Sept 2017].

[86] NASA Earth Data. MODIS (MOD 13) - Gridded vegetation indices (NDVI

and EVI). http://modis.gsfc.nasa.gov/data/dataprod/dataproducts.php?MOD_NUMBER=13, 2017. [Accessed Sept 2017].

[87] Steven Piantadosi, David P Byar, and Sylvan B Green. The ecological fallacy. *American journal of epidemiology*, 127(5):893–904, 1988.

[88] Jon Wakefield and Gavin Shaddick. Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7(3):438–455, 2006.

[89] Jonathan Wakefield. Ecologic studies revisited. *Annu. Rev. Public Health*, 29: 75–90, 2008.

[90] A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7): 1025–1044, 1991.

[91] Melanie M Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of statistical planning and inference*, 121(2):311–324, 2004.

[92] Petr Keil, Jonathan Belmaker, Adam M Wilson, Philip Unitt, and Walter Jetz. Downscaling of species distribution models: a hierarchical approach. *Methods in Ecology and Evolution*, 4(1):82–94, 2013.

[93] Hugh JW Sturrock, Justin M Cohen, Petr Keil, Andrew J Tatem, Arnaud Le Menach, Nyasatu E Ntshalintshali, Michelle S Hsiang, and Roland D Gosling. Fine-scale malaria risk mapping from routine aggregated case data. *Malaria Journal*, 13(1):421, 2014.

[94] Benjamin M Taylor, Ricardo Andrade-Pacheco, and Hugh JW Sturrock. Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1125–1150, 2018.

[95] Rohan Arambepola, Tim CD Lucas, Anita K Nandi, Peter W Gething, and Ewan Cameron. A simulation study of disaggregation regression for spatial disease mapping. *arXiv preprint arXiv:2005.03604*, 2020.

[96] World Health Organization. *World malaria report 2019*. World Health Organization, 2020.

[97] Ministry of Health Directorate of Health Services and Central TB Division Family Welfare. National strategic plan for malaria elimination in India 2017-2022. 2017.

[98] Jai Prakash Narain and Lalit M Nath. Eliminating malaria in India by 2027: The countdown begins! *The Indian journal of medical research*, 148(2):123, 2018.

[99] POMPI INDIA. *Census of India 2011 Provisional Population Totals*. Office of the Registrar General and Census Commissioner New Delhi, 2011.

[100] Peter WG Tennant, Wendy J Harrison, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Claire Keeble, Lynsie R Ranker, Johannes Textor, Georgia D Tomova, Mark S Gilthorpe, and George TH Ellison. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. *medRxiv*, 2019.

[101] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[102] Daniel M Hausman and James Woodward. Independence, invariance and the causal Markov condition. *The British journal for the philosophy of science*, 50 (4):521–583, 1999.

[103] Daniel Steel. Indeterminism and the causal Markov condition. *The British journal for the philosophy of science*, 56(1):3–26, 2005.

[104] G Udny Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903.

[105] Naftali Weinberger. Faithfulness, coordination and causal coincidences. *Erkenntnis*, 83(2):113–133, 2018.

[106] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

[107] Dan Geiger and Judea Pearl. On the logic of causal models. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 3–14. Elsevier, 1990.

[108] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

[109] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[110] Jin Hyung Kim and Judea Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *International Joint Conference on Artificial Intelligence*, pages 0–0, 1983.

[111] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[112] Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

[113] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[114] Andrés Cano, Manuel Gómez-Olmedo, and Serafín Moral. A score based ranking of the edges for the PC algorithm. In *Proceedings of the Fourth European Workshop on Probabilistic Graphical Models*, pages 41–48, 2008.

[115] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1): 3741–3782, 2014.

[116] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[117] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[118] Taku Kudo and Yuji Matsumoto. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31, 2003.

[119] Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. *Machine learning*, 74(1):1–22, 2009.

[120] Shohei Hido and Hisashi Kashima. A linear-time graph kernel. In *2009 Ninth IEEE International Conference on Data Mining*, pages 179–188. IEEE, 2009.

[121] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

[122] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

[123] Walter Rudin. *Fourier analysis on groups*, volume 121967. Wiley Online Library, 1962.

[124] Jean-Francois Ton, Seth Flaxman, Dino Sejdinovic, and Samir Bhatt. Spatial mapping with Gaussian processes and nonstationary Fourier features. *Spatial statistics*, 28:59–78, 2018.

[125] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. In *International Conference on Machine Learning*, pages 3905–3914. PMLR, 2019.

[126] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[127] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[128] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[129] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[130] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperum-

budur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[131] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. *Advances in neural information processing systems*, 26:755–763, 2013.

[132] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[133] Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning*, pages 1742–1751. PMLR, 2017.

[134] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[135] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors-supplementary. *arXiv preprint arXiv:1205.4656*, 2012.

[136] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

[137] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[138] Louis De Branges. The Stone-Weierstrass theorem. *Proceedings of the American Mathematical Society*, 10(5):822–824, 1959.

[139] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20:489–496, 2007.

[140] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

[141] Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in neural information processing systems*, 22:1750–1758, 2009.

[142] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[143] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[144] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21:689–696, 2008.

[145] Dimitris Margaritis. Distribution-free learning of Bayesian network structure in continuous domains. In *AAAI*, volume 5, pages 825–830, 2005.

[146] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[147] Daniel Alpay. *The Schur algorithm, reproducing kernel spaces and system theory*. American Mathematical Soc., 2001.

[148] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. Multi-instance kernels. In *ICML*, volume 2, page 7, 2002.

[149] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957, 2015.

[150] Seth R Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported Obama in 2012? Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298, 2015.

[151] Seth R Flaxman, Daniel B Neill, and Alexander J Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–23, 2015.

[152] World Bank. The World Bank in Madagascar. `https://www.worldbank.org/en/country/madagascar/overview`. Accessed: 2020-07-14.

[153] Sima Barmania. Madagascar's health challenges. *The Lancet*, 386(9995):729–730, 2015.

[154] Institute for Health Metrics and Evaluation. Madagascar country profile. `http://www.healthdata.org/madagascar`. Accessed: 2020-07-14.

[155] Stéphane Laventure, Jean Mouchet, Sixte Blanchy, Laurence Marrama, Patrick Rabarison, Lala Andrianaivolambo, Edmond Rajaonarivelo, Ignace Rakotoarivony, and Jean Roux. Rice: a source of life and death on the highlands of Madagascar. *Cahiers d'études et de recherches francophones/Santé*, 6(2):79–86, 1996.

[156] J Mouchet, S Laventure, S Blanchy, R Fioramonti, A Rakotonjanabelo, P Rabarison, J Sircoulon, and J Roux. The reconquest of the Madagascar highlands by malaria. *Bulletin de la Societe de pathologie exotique (1990)*, 90 (3):162–168, 1997.

[157] J Mouchet, S Blanchy, A Rakotonjanabelo, G Ranaivoson, E Rajaonarivelo, S Laventure, M Rossella, and F Aknouche. Epidemiological stratification of malaria in Madagascar. *Archives de l'Institut Pasteur de Madagascar*, 60(1-2): 50–59, 1993.

[158] G Joncour. La lutte contre le paludisme à madagascar. *Bulletin of the World Health Organization*, 15(3-5):711, 1956.

[159] Milijaona Randrianarivelojosia, Andrianirina Raveloson, Arthur Randriamanantena, Jonathan J Juliano, Tahina Andrianjafy, Lucie A Raharimalala, and Vincent Robert. Lessons learnt from the six decades of chloroquine use (1945–2005) to control malaria in Madagascar. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 103(1):3–10, 2009.

[160] JP Lepers, JA Ramanamirija, MD Rason, and P Coulanges. In vitro sensitivity of *Plasmodium falciparum* to chloroquine in the highland region of Madagascar in 1987. *Bulletin de la Societe de pathologie exotique et de ses filiales*, 81(3): 338–344, 1988.

[161] Jean Paul Lepers, Philippe Deloron, Jean Claude Mouden, Jacques Le Bras, and Pierre Coulanges. Sudden increase in number of isolates of *Plasmodium falciparum* resistant to chloroquine in Madagascar. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 83(4):491–492, 1989.

[162] ML Willcox, E Rakotondrazafy, R Andriamanalimanana, D Andrianasolo, and P Rasoanaivo. Decreasing clinical efficacy of chloroquine in Ankazobe, Cen-

tral Highlands of Madagascar. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(5):311–314, 2004.

[163] Valérie Andriantsoanirina, Didier Ménard, Luciano Tuseo, and Rémy Durand. History and current status of *Plasmodium falciparum* antimalarial drug resistance in Madagascar. *Scandinavian journal of infectious diseases*, 42(1):22–32, 2010.

[164] JeanPaul Lepers, Philippe Deloron, Didier Fontenille, and Pierre Coulanges. Reappearance of *falciparum* malaria in central highland plateaux of madagascar. *Lancet (London, England)*, 1(8585):586–586, 1988.

[165] D Fontenille and I Rakotoarivony. Reappearance of *Anopheles funestus* as a malaria vector in the Antananarivo region, Madagascar. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 82(4):644–645, 1988.

[166] JP Lepers, P Deloron, MD Andriamagatiana-Rason, JA Ramanamirija, and P Coulanges. Newly transmitted *Plasmodium falciparum* malaria in the central highland plateaux of Madagascar: assessment of clinical impact in a rural community. *Bulletin of the World Health Organization*, 68(2):217, 1990.

[167] JEAN Mouchet, Sylvie Manguin, JACQUES Sircoulon, STEPHANE Laventure, OUSMANE Faye, AW Onapa, Pierre Carnevale, J Julvez, and Didier Fontenille. Evolution of malaria in Africa for the past 40 years: impact of climatic and human factors. *Journal of the American Mosquito Control Association*, 14(2):121, 1998.

[168] Anthony E Kiszewski and Awash Teklehaimanot. A review of the clinical and epidemiologic burdens of epidemic malaria. *The American journal of tropical medicine and hygiene*, 71(2_suppl):128–135, 2004.

[169] Milijaona Randrianarivelojosia, David A Fidock, Olivier Belmonte, Stephanie G Valderramos, Odile Mercereau-Puijalon, and Frédéric Ariey. First evidence of pfcrt mutant *Plasmodium falciparum* in Madagascar. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(9):826–830, 2006.

[170] M Albonico, F Giorgi De, J Razanakolona, A Raveloson, G Sabatinelli, V Pietra, and D Modiano. Control of epidemic malaria on the highlands of Madagascar. *Parassitologia*, 41(1-3):373–376, 1999.

[171] F Rakotomanana, Isabelle Jeanne, JB Duchemin, V Pietra, L Raharimalala, ML Tombo, and F Ariey. Geographic approach in malaria control in the central highlands of Madagascar. *Archives de l'Institut Pasteur de Madagascar*, 67(1-2): 27, 2001.

[172] R Romi, MC Razaiarimanga, R Raharimanga, EM Rakotondraibe, LH Ranaivo, V Pietra, A Raveloson, and G Majori. Impact of the malaria control campaign (1993-1998) in the highlands of Madagascar: parasitological and entomological data. *The American journal of tropical medicine and hygiene*, 66(1):2–6, 2002.

[173] President's Malaria Initiative. Madagascar malaria operational plan financial year 2008, 2007.

[174] The World Bank. Madagascar: Measuring the impact of the political crisis. 2013.

[175] Institut National de la Statistique (INSTAT), Programme National de lutte contre le Paludisme (PNLP), and ICF International. *Madagascar Malaria Indicator Survey 2011 [Enquête sur les Indicateurs du Paludisme (EIPM)]*. INSTAT, PNLP, IPM and ICF International, Calverton, 2012.

[176] Laurence Randrianasolo, Yolande Raoelina, Maherisoa Ratsitorahina, Lisette Ravolomanana, Soa Andriamandimby, Jean-Michel Heraud, Fanjasoa Rako-

tomanana, Robinson Ramanjato, Armand Eugène Randrianarivo-Solofoniaina, and Vincent Richard. Sentinel surveillance system for early outbreak detection in Madagascar. *BMC Public Health*, 10(1):31, 2010.

[177] Michele Nguyen, Rosalind E. Howes, Tim C.D. Lucas, Katherine E. Battle, Ewan Cameron, Harry S. Gibson, Jennifer Rozier, Suzanne Keddie, Emma Collins, Rohan Arambepola, Su Yun Kang, Chantal Hendriks, Anita Nandi, Susan F. Rumisha, Samir Bhatt, Sedera A. Mioramalala, Mauricette Andriamananjara Nambinisoa, Fanjasoa Rakotomanana, Peter W. Gething, and Daniel J. Weiss. Mapping malaria seasonality in Madagascar using health facility data. *BMC Medicine*, 18(1):26, Feb 2020.

[178] Institut National de la Statistique (INSTAT), Programme National de lutte contre le Paludisme (PNLP), Institut Pasteur de Madagascar (IPM), and ICF International. *Madagascar Malaria Indicator Survey 2013 [Enquête sur les Indicateurs du Paludisme (EIPM)]*. INSTAT, PNLP, IPM and ICF International, Calverton, 2013.

[179] Emmanuel Chanda, Michael Coleman, Immo Kleinschmidt, Janet Hemingway, Busiku Hamainza, Freddie Masaninga, Pascalina Chanda-Kapata, Kumar S Baboo, David N Dürrheim, and Marlize Coleman. Impact assessment of malaria vector control using routine surveillance data in Zambia: implications for monitoring and evaluation. *Malaria Journal*, 11(1):437, 2012.

[180] Adam Bennett, Joshua Yukich, John M Miller, Penelope Vounatsou, Busiku Hamainza, Mercy M Ingwe, Hawela B Moonga, Mulakwo Kamuliwo, Joseph Keating, Thomas A Smith, Richard W Steketee, and Thomas P Eisele. A methodological framework for the improved use of routine health system data to evaluate national malaria control programs: evidence from Zambia. *Population health metrics*, 12(1):30, 2014.

[181] Felana A Ihantamalala, Feno MJ Rakotoarimanana, Tanjona Ramiadantsoa, Jean Marius Rakotondramanga, Gwenaëlle Pennober, Fanjasoa Rakotomanana, Simon Cauchemez, Charlotte JE Metcalf, Vincent Herbreteau, and Amy Wesolowski. Spatial and temporal dynamics of malaria in Madagascar. *Malaria Journal*, 17(1):58, 2018.

[182] Madagascar: Standard DHS, 2020. `https://dhsprogram.com/methodology/survey/survey-display-560.cfm`. Accessed: 2021-01-02.

[183] National Malaria Control Programme of Madagascar. Plan strategique national de lutte contre le paludisme: Elimination progressive du paludisme à Madagascar, 2018.

[184] Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

[185] Craig Wang, Milo A Puhan, Reinhard Furrer, and SNC Study Group. Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, 23:72–90, 2018.

[186] Tim CD Lucas, Anita K Nandi, Elisabeth G Chestnutt, Katherine A Twohig, Suzanne H Keddie, Emma L Collins, Rosalind E Howes, Michele Nguyen, Susan Rumisha, Andre Python, Rohan Arambepola, Amelia Bertozzi-Villa, Penelope Hancock, Punam Amratia, Katherine E Battle, Ewan Cameron, Peter W Gething, and Daniel J Weiss. Mapping malaria by sharing spatial information between incidence and prevalence datasets. *medRxiv*, 2020.

[187] Fei Liu, MJ Bayarri, and JO Berger. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

[188] Daniel J Weiss, Samir Bhatt, Bonnie Mappin, Thomas P Van Boeckel, David L Smith, Simon I Hay, and Peter W Gething. Air temperature suitability for *Plasmodium falciparum* malaria transmission in Africa 2000-2012: a high-resolution spatiotemporal prediction. *Malaria Journal*, 13(1):171, 2014.

[189] Tim C.D. Lucas, Anita K. Nandi, Suzanne H. Keddie, Elisabeth G. Chestnutt, Rosalind E. Howes, Susan F. Rumisha, Rohan Arambepola, Amelia Bertozzi-Villa, Andre Python, Tasmin L. Symons, Justin J. Millar, Punam Amratia, Penelope Hancock, Katherine E. Battle, Ewan Cameron, Peter W. Gething, and Daniel J. Weiss. Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence. *Spatial and Spatio-temporal Epidemiology*, page 100357, 2020.

[190] Chris C Funk, Pete J Peterson, Martin F Landsfeld, Diego H Pedreros, James P Verdin, James D Rowland, Bo E Romero, Gregory J Husak, Joel C Michaelsen, and Andrew P Verdin. A quasi-global precipitation time series for drought monitoring. *US Geological Survey Data Series*, 832(4):1–12, 2014.

[191] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[192] Alfred Pach, Michelle Warren, Irene Chang, Justin Im, Chelsea Nichols, Christian G Meyer, Gi Deok Pak, Ursula Panzner, Se Eun Park, Vera von Kalckreuth, Stephen Baker, Henintsoa Rabezanahary, Jean Philibert Rakotondrainiarivelo, Tiana Mirana Raminosoa, Rakotozandrindrainy Raphaël, and Florian Marks. A qualitative study investigating experiences, perceptions, and healthcare system performance in relation to the surveillance of typhoid fever in Madagascar. *Clinical Infectious Diseases*, 62(suppl_1):S69–S75, 2016.

[193] Victor A Alegana, Jim A Wright, Uusiku Pentrina, Abdisalan M Noor,

Robert W Snow, and Peter M Atkinson. Spatial modelling of healthcare utilisation for treatment of fever in Namibia. *International journal of health geographics*, 11(1):6, 2012.

[194] NASA Socioeconomic Data Center for International Earth Science Information Network CIESIN Columbia University and Applications Center (SEDAC). Gridded population of the world, version 4 (GPWv4): Administrative unit center points with population estimates, 2018. URL `http://dx.doi.org/10.7927/H4F47M2C`.

[195] Andrew J Tatem. Worldpop, open data for spatial demography. *Scientific data*, 4(1):1–4, 2017.

[196] Ministry of Health of Madagascar. Reference manual of principal health system indicators [in french]., 2014.

[197] K Kristensen, A Nielsen, CW Berg, H Skaug, and B Bell. Template model builder TMB. *J. Stat. Softw*, 70:1–21, 2015.

[198] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

[199] President's Malaria Initiative. Madagascar malaria operational plan financial year 2017, 2016.

[200] President's Malaria Initiative. Madagascar malaria operational plan financial year 2013, 2012.

[201] President's Malaria Initiative. Madagascar malaria operational plan financial year 2014, 2013.

[202] Mai Do, Stella Babalola, Grace Awantang, Michael Toso, Nan Lewicky, and Andrew Tompsett. Associations between malaria-related ideational factors and care-seeking behavior for fever among children under five in Mali, Nigeria, and Madagascar. *PloS one*, 13(1), 2018.

[203] Katie Wilson and Jon Wakefield. Pointless spatial modeling. *Biostatistics*, 21 (2):e17–e32, 2020.

[204] Joseph Friedman, Patrick Liu, Emmanuela Gakidou, IHME COVID, and Model Comparison Team. Predictive performance of international COVID-19 mortality forecasting models. *medRxiv*, 2020.

[205] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[206] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[207] Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

[208] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[209] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[210] Marilena Barbieri, James O Berger, Edward I George, and Veronika Rocková. The median probability model and correlated variables. *arXiv preprint arXiv:1807.08336*, 2018.

[211] Peter W Gething, Thomas P Van Boeckel, David L Smith, Carlos A Guerra, Anand P Patil, Robert W Snow, and Simon I Hay. Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. *Parasites & vectors*, 4(1):92, 2011.

[212] Kevin Dunne, Padraig Cunningham, and Francisco Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research*, pages 1–22, 2002.

[213] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[214] Willi Sauerbrei, Anika Buchholz, Anne-Laure Boulesteix, and Harald Binder. On stability issues in deriving multivariable regression models. *Biometrical Journal*, 57(4):531–555, 2015.

[215] Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

[216] Peter C Austin and Jack V Tu. Bootstrap methods for developing predictive models. *The American Statistician*, 58(2):131–137, 2004.

[217] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[218] Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

[219] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[220] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[221] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An introduction to g methods. *International journal of epidemiology*, 46(2):756–762, 2017.

[222] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[223] Moritz UG Kraemer, Robert C Reiner Jr, and Samir Bhatt. Causal inference in spatial mapping. *Trends in parasitology*, 35(10):743–746, 2019.

[224] David Roxbee Cox. *Planning of experiments.* Wiley, 1958.

[225] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[226] Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

[227] Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

[228] J Pryce, M Richardson, and C Lengeler. Insecticide-treated nets for preventing malaria. *Cochrane Database of Systematic Reviews*, (11), 2018.

[229] B Pluess, FC Tanser, C Lengeler, and BL Sharp. Indoor residual spraying for preventing malaria. *Cochrane Database of Systematic Reviews*, (4), 2010.

[230] CE Shulman, EK Dorman, AO Talisuna, BS Lowe, C Nevill, Robert W Snow, H Jilo, N Peshu, JN Bulmer, and K Graham, S Marsh. A community random-

ized controlled trial of insecticide-treated bednets for the prevention of malaria and anaemia among primigravid women on the Kenyan coast. *Tropical Medicine and International Health*, 3(3):197–204, 1998.

[231] Edmund Nii Laryea Browne, Gillian H Maude, and Fred N Binka. The impact of insecticide-treated bednets on malaria and anaemia in pregnancy in Kassena-Nankana district, Ghana: a randomized controlled trial. *Tropical Medicine & International Health*, 6(9):667–676, 2001.

[232] Olaf Müller, Manuela De Allegri, Heiko Becher, Justin Tiendrebogo, Claudia Beiersmann, Maurice Ye, Bocar Kouyate, Ali Sie, and Albrecht Jahn. Distribution systems of insecticide-treated bed nets for malaria control in rural Burkina Faso: cluster-randomized controlled trial. *PLoS One*, 3(9):e3182, 2008.

[233] Olaf Müller, Corneille Traoré, Bocar Kouyaté, Yazoumé Yé, Claudia Frey, Boubacar Coulibaly, and Heiko Becher. Effects of insecticide-treated bednets during early infancy in an African area of intense malaria transmission: a randomized controlled trial. *Bulletin of the World Health Organization*, 84:120–126, 2006.

[234] Pedro L Alonso, Steve W Lindsay, JRM Armstrong, A De Francisco, FC Shenton, BM Greenwood, M Conteh, K Cham, AG Hill, PH David, G Fegan, and AJ Hall. The effect of insecticide-treated bed nets on mortality of Gambian children. *The Lancet*, 337(8756):1499–1502, 1991.

[235] Tobias O Apinjoh, Judith K Anchang-Kimbi, Regina N Mugri, Delphine A Tangoh, Robert V Nyingchu, Hanesh F Chi, Rolland B Tata, Charles Njumkeng, Clarisse Njua-Yafi, and Eric A Achidi. The effect of Insecticide Treated Nets (ITNs) on *Plasmodium falciparum* infection in rural and semi-urban communities in the south west region of Cameroon. *PLoS One*, 10(2):e0116300, 2015.

[236] Frank M Smithuis, Moe Kyaw Kyaw, U Ohn Phe, Ingrid van der Broek, Nina Katterman, Colin Rogers, Patrick Almeida, Piet A Kager, Kasia Stepniewska, Yoel Lubell, Julie A Simpson, and Nicholas J White. The effect of insecticide-treated bed nets on the incidence and prevalence of malaria in children in an area of unstable seasonal transmission in western Myanmar. *Malaria journal*, 12(1):363, 2013.

[237] Jamie T Griffin, T Deirdre Hollingsworth, Lucy C Okell, Thomas S Churcher, Michael White, Wes Hinsley, Teun Bousema, Chris J Drakeley, Neil M Ferguson, María-Gloria Basáñez, and Azra C Ghani. Reducing *plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med*, 7(8):e1000324, 2010.

[238] Patrick GT Walker, Jamie T Griffin, Neil M Ferguson, and Azra C Ghani. Estimating the most efficient allocation of interventions to achieve reductions in *plasmodium falciparum* malaria burden and transmission in africa: a modelling study. *The Lancet Global Health*, 4(7):e474–e484, 2016.

[239] Anna Bershteyn, Jaline Gerardin, Daniel Bridenbecker, Christopher W Lorton, Jonathan Bloedow, Robert S Baker, Guillaume Chabot-Couture, Ye Chen, Thomas Fischle, Kurt Frey, Jillian S Gauld, Hao Hu, Amanda S Izzo, Daniel J Klein, Dejan Lukacevic, Kevin A McCarthy, Joel C Miller, Andre Lin Ouedraogo, T Alex Perkins, Jeffrey Steinkraus, Quirine A ten Bosch, Hung-Fu Ting, Svetlana Titova, Bradley G Wagner, Philip A Welkhoff, Edward A Wenger, Christian N Wiswell, and for the Institute for Disease Modeling. Implementation and applications of EMOD, an individual-based multi-disease modeling platform. *Pathogens and Disease*, 76(5), 07 2018.

[240] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

[241] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

[242] Fredrick G Kabbale, Anne M Akol, John B Kaddu, and Ambrose W Onapa. Biting patterns and seasonality of *anopheles gambiae sensu lato* and *anopheles funestus* mosquitoes in kamuli district, uganda. *Parasites & vectors*, 6(1):340, 2013.

[243] Oljira Kenea, Meshesha Balkew, Habte Tekie, Teshome Gebre-Michael, Wakgari Deressa, Eskindir Loha, Bernt Lindtjørn, and Hans J Overgaard. Human-biting activities of *anopheles* species in south-central ethiopia. *Parasites & vectors*, 9 (1):527, 2016.

[244] Khamis A Haji, Narjis G Thawer, Bakari O Khatib, Juma H Mcha, Abdallah Rashid, Abdullah S Ali, Christopher Jones, Judit Bagi, Stephen M Magesa, Mahdi M Ramsan, , Issa Garimo, George Greer, Richard Reithinger, and Jeremiah M Ngondi. Efficacy, persistence and vector susceptibility to pirimiphos-methyl (Actellic® 300cs) insecticide for indoor residual spraying in Zanzibar. *Parasites & vectors*, 8(1):1–7, 2015.

[245] Delenasaw Yewhalaw, Meshesha Balkew, Josephat Shililu, Sultan Suleman, Alemayehu Getachew, Gedeon Ashenbo, Sheleme Chibsa, Gunawardena Dissanayake, Kristen George, Dereje Dengela, Yemane Ye-Ebiyo, and Seth R Irish. Determination of the residual efficacy of carbamate and organophosphate insecticides used for indoor residual spraying for malaria control in Ethiopia. *Malaria Journal*, 16(1):1–9, 2017.

[246] Mary J Hamel, Peter Otieno, Nabie Bayoh, Simon Kariuki, Vincent Were, Doris Marwanga, Kayla F Laserson, John Williamson, Laurence Slutsker, and John

Gimnig. The combination of indoor residual spraying and insecticide-treated nets provides added protection against malaria compared with insecticide-treated nets alone. *The American journal of tropical medicine and hygiene*, 85(6):1080–1086, 2011.

[247] Nancy Fullman, Roy Burstein, Stephen S Lim, Carol Medlin, and Emmanuela Gakidou. Nets, spray or both? the effectiveness of insecticide-treated nets and indoor residual spraying in reducing malaria morbidity and child mortality in sub-Saharan Africa. *Malaria journal*, 12(1):1–10, 2013.

[248] Giulia Rathmes, Susan F. Rumisha, Tim C. D. Lucas, Katherine A. Twohig, Andre Python, Michele Nguyen, Anita K. Nandi, Suzanne H. Keddie, Emma L. Collins, Jennifer A. Rozier, Harry S. Gibson, Elisabeth G. Chestnutt, Katherine E. Battle, Georgina S. Humphreys, Punam Amratia, Rohan Arambepola, Amelia Bertozzi-Villa, Penelope Hancock, Justin J. Millar, Tasmin L. Symons, Samir Bhatt, Ewan Cameron, Philippe J. Guerin, Peter W. Gething, and Daniel J. Weiss. Global estimation of anti-malarial drug effectiveness for the treatment of uncomplicated *plasmodium falciparum* malaria 1991–2019. *Malaria Journal*, 19(1):374, Oct 2020.

[249] Lisa A Jackson, Michael L Jackson, Jennifer C Nelson, Kathleen M Neuzil, and Noel S Weiss. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International journal of epidemiology*, 35(2):337–344, 2006.

[250] Marc Lipsitch, Ayan Jha, and Lone Simonsen. Observational studies and the difficult quest for causality: lessons from vaccine effectiveness and impact studies. *International journal of epidemiology*, 45(6):2060–2074, 2016.

[251] Gregory A Poland. If you could halve the mortality rate, would you do it? *Clinical Infectious Diseases*, 35(4):378–380, 08 2002. ISSN 1058-4838.

[252] Kristin L Nichol, James D Nordin, David B Nelson, John P Mullooly, and Eelko Hak. Effectiveness of influenza vaccine in the community-dwelling elderly. *New England Journal of Medicine*, 357(14):1373–1381, 2007.

[253] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

[254] Thiemo Fetzer and Thomas Graeber. Does contact tracing work? quasi-experimental evidence from an excel error in England. 2020.

[255] Connor A Emdin, Amit V Khera, and Sekar Kathiresan. Mendelian randomization. *Jama*, 318(19):1925–1926, 2017.

[256] Hyunseung Kang, Benno Kreuels, Ohene Adjei, Ralf Krumkamp, Jürgen May, and Dylan S Small. The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International journal of epidemiology*, 42(5):1390–1398, 2013.

[257] Dylan S Small, Terrie E Taylor, Douglas G Postels, Nicholas AV Beare, Jing Cheng, Ian JC MacCormick, and Karl B Seydel. Evidence from a natural experiment that malaria parasitemia is pathogenic in retinopathy-negative cerebral malaria. *Elife*, 6:e23699, 2017.

[258] Ismail D. Legason, Ruth M. Pfeiffer, Krizia-Ivana Udquim, Andrew W. Bergen, Mateus H. Gouveia, Samuel Kirimunda, Isaac Otim, Eric Karlins, Patrick Kerchan, Hadijah Nabalende, Ariunaa Bayanjargal, Benjamin Emmanuel, Paul Kagwa, Ambrose O. Talisuna, Kishor Bhatia, Meredith Yeager, Robert J. Biggar, Leona W. Ayers, Steven J. Reynolds, James J. Goedert, Martin D. Ogwang,

Joseph F. Fraumeni, Ludmila Prokunina-Olsson, and Sam M. Mbulaiteye. Evaluating the causal link between malaria infection and endemic Burkitt lymphoma in northern Uganda: a Mendelian randomization study. *EBioMedicine*, 25:58–65, 2017.

[259] Anthony O Etyang, Sailoki Kapesa, Emily Odipo, Evasius Bauni, Catherine Kyobutungi, Marwah Abdalla, Paul Muntner, Solomon K Musani, Alex Macharia, Thomas N Williams, J Kennedy Cruickshank, Liam Smeeth, and J Anthony G Scott. Effect of previous exposure to malaria on blood pressure in Kilifi, Kenya: a mendelian randomization study. *Journal of the American Heart Association*, 8(6):e011771, 2019.

[260] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature methods*, 7(4):247–248, 2010.

[261] Thomas Kesteman, Solofoniaina A. Rafalimanantsoa, Harimahefa Razafimandimby, Heriniaina H. Rasamimanana, Vaomalala Raharimanga, Benjamin Ramarosandratana, Arsene Ratsimbasoa, Jocelyn Ratovonjato, Nohal Elissa, Laurence Randrianasolo, Alyssa Finlay, Christophe Rogier, and Milijaona Randrianarivelojosia. Multiple causes of an unexpected malaria outbreak in a high-transmission area in Madagascar. *Malaria Journal*, 15(1):57, Feb 2016.

[262] Alemayehu Midekisa, Belay Beyene, Abere Mihretie, Estifanos Bayabil, and Michael C Wimberly. Seasonal associations of climatic drivers and malaria in the highlands of Ethiopia. *Parasites & vectors*, 8(1):1–11, 2015.

[263] Karina Laneri, Brenno Cabella, Paulo Inácio Prado, Renato Mendes Coutinho, and Roberto André Kraenkel. Climate drivers of malaria at its southern fringe in the Americas. *PloS one*, 14(7):e0219249, 2019.

[264] Cara Smith Gueye, Kelly C. Sanders, Gawrie NL Galappaththy, Christina Rundi, Tashi Tobgay, Siv Sovannaroth, Qi Gao, Asik Surya, Garib D. Thakur, Mario Baquilod, Won-ja Lee, Alby Bobogare, Sarath L. Deniyage, Wichai Satimai, George Taleo, Nguyen M. Hung, Chris Cotter, Michelle S. Hsiang, Lasse S. Vestergaard, and Roly D. Gosling. Active case detection for malaria elimination: a survey among asia pacific countries. *Malaria Journal*, 12(1):358, Oct 2013.

[265] G Dennis Shanks. Control and elimination of *plasmodium vivax*. *Advances in parasitology*, 80:301–341, 2012.

[266] Ric N Price, Robert J Commons, Katherine E Battle, Kamala Thriemer, and Kamini Mendis. *Plasmodium vivax* in the era of the shrinking *P. falciparum* map. *Trends in parasitology*, 2020.

[267] David A Larsen, Anne Martin, Derek Pollard, Carrie F Nielsen, Busiku Hamainza, Matthew Burns, Jennifer Stevenson, and Anna Winters. Leveraging risk maps of malaria vector abundance to guide control efforts reduces malaria incidence in Eastern Province, Zambia. *Scientific reports*, 10(1):1–12, 2020.