Improving geospatial models of risk for vector-borne, zoonotic diseases

Freya M. Shearer

St. John's College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

Abstract

Public health surveillance data are often incomplete, particularly where resources are lacking, but geospatial models can help to fill the gaps by providing estimates where data are sparse. By combining information on locations where diseases have been recorded with geographic data on environmental and socioeconomic covariates known to affect disease transmission using machine-learning models (such as boosted regression trees), niche modelling can generate fine-resolution, evidence-based risk maps for a variety of diseases of public health importance.

This thesis investigates the geographical distribution of two vector-borne, zoonotic diseases of public health importance: *Plasmodium knowlesi* malaria and yellow fever (YF). A number of new methodological approaches to niche modelling are developed for: mapping diseases whose distributions are impacted by multiple host and vector species, ameliorating spatial bias in disease reporting rates, and accounting for human vaccination coverage.

Chapter 2 investigates spatial variation in risk of human *P. knowlesi* infection across Southeast Asia. The infection risk model for *P. knowlesi* malaria is based on improvements to a standard niche modelling approach, and incorporates a novel joint distribution model to leverage data from a number of host species.

Chapter 3 estimates YF vaccination coverage through time across all age cohorts in every district/municipality of countries at risk of YF, globally. These estimates are used to estimate the additional vaccination coverage needed to prevent further YF outbreaks, and they provide information needed to account for population immunity when estimating YF infection risk.

Chapter 4 describes the development of a novel Poisson point process niche model, which is then used to predict YF infection risk in humans and demonstrates how vaccination coverage can be efficiently accounted for in disease niche models.

The disease risk maps of *P. knowlesi* malaria and YF produced through this thesis will act as resources to improve the targeting, implementation and evaluation of disease prevention, surveillance and control strategies. Methods developed to account for vaccination coverage, reporting rate biases, and complex transmission systems will be applicable to risk mapping for a range of vector-borne, zoonotic diseases of public health importance.

Improving geospatial models of risk for vector-borne, zoonotic diseases



Freya M. Shearer

St. John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

Acknowledgements

I would to thank my supervisors, Catherine Moyes, Nick Golding, and Simon Hay for all the support, advice, and technical help they have given me over the past three years. I could not have completed a thesis in a field that, three years ago was so new to me, without their patience and enthusiasm.

I am very grateful to the Rhodes Trust for providing me with the opportunity to study at the University of Oxford, and for funding the work in this thesis.

I would also like to acknowledge Brydee, Ryan, Maria, and Mahesh for being stringent proof-readers of this document.

The work in this thesis would not have been possible without the help and support of my brilliant officemates – David, Oli, and Moritz in the early days, and Maria, Josh, Toni, Laurence, Helena and Annie during the later stages. Thanks also to Ani, Jeff, Bobby and Jon, at IHME, University of Washington, and everyone at QAECO and Trinity College, University of Melbourne, for hosting me, and for our collaborations.

Thank you to my friends, teammates and coaches at OUAC and OUARFC, as well as the staff at Rhodes House and St John's College, for providing me with much joy and support throughout these three years in Oxford, particularly during the gloomy winter months. Special thanks go to Perth friends – Sheen, Sab, Pearse, James, Mon, Rob, and Shannon – for their visits.

I must also thank Ken and Helen for their love, and the endless supplies of tea and biscuits.

Finally, thank you to my parents, and Brydee and Ryan – I would not have reached this point without their constant love and support.

Abstract

Public health surveillance data are often incomplete, particularly where resources are lacking, but geospatial models can help to fill the gaps by providing estimates where data are sparse. By combining information on locations where diseases have been recorded with geographic data on environmental and socioeconomic covariates known to affect disease transmission using machine-learning models (such as boosted regression trees), niche modelling can generate fine-resolution, evidence-based risk maps for a variety of diseases of public health importance.

This thesis investigates the geographical distribution of two vector-borne, zoonotic diseases of public health importance: *Plasmodium knowlesi* malaria and yellow fever (YF). A number of new methodological approaches to niche modelling are developed for: mapping diseases whose distributions are impacted by multiple host and vector species, ameliorating spatial bias in disease reporting rates, and accounting for human vaccination coverage.

Chapter 2 investigates spatial variation in risk of human *P. knowlesi* infection across Southeast Asia. The infection risk model for *P. knowlesi* malaria is based on improvements to a standard niche modelling approach, and incorporates a novel joint distribution model to leverage data from a number of host species.

Chapter 3 estimates YF vaccination coverage through time across all age cohorts in every district/municipality of countries at risk of YF, globally. These estimates are used to estimate the additional vaccination coverage needed to prevent further YF outbreaks, and they provide information needed to account for population immunity when estimating YF infection risk.

Chapter 4 describes the development of a novel Poisson point process niche model, which is then used to predict YF infection risk in humans and demonstrates how vaccination coverage can be efficiently accounted for in disease niche models.

The disease risk maps of *P. knowlesi* malaria and YF produced through this thesis will act as resources to improve the targeting, implementation and evaluation of disease prevention, surveillance and control strategies. Methods developed to account for vaccination coverage, reporting rate biases, and complex transmission systems will be applicable to risk mapping for a range of vector-borne, zoonotic diseases of public health importance.

List of publications

Original manuscripts contributing to thesis

Moyes CL, **Shearer FM**, Huang Z, Wiebe A, Gibson HS, Nijman V, Mohd-Azlan J, Brodie JF, Malaivijitnond S, Linkie M, Samejima H, O'Brien TG, Trainor CR, Hamada Y, Giordano AJ, Kinnaird MF, Elyazar IRF, Sinka ME, Vythilingam I, Bangs MJ, Pigott DM, Weiss DJ, Golding N, and Hay SI. (2016). Predicting the geographical distributions of the macaque hosts and mosquito vectors of *Plasmodium knowlesi* malaria in forest and non-forested areas. *Parasites and Vectors*. Epub 28 Apr 2016. doi:10.1186/s13071-016-1527-0.

Shearer FM, Huang Z, Weiss DJ, Wiebe A, Gibson HS, Battle KE, Pigott DM, Brady OJ, Putaporntip C, Jongwutiwes S, Lau YL, Manske M, Amato R, Elyazar IRF, Vythilingam I, Bhatt S, Gething PW, Singh B, Golding N, Hay SI, and Moyes CL. (2016). Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria. *PLOS Neglected Tropical Diseases*. Epub Aug 5 2016. doi: 10.1371/journal.pntd.0004915.

Shearer FM, Moyes CL, Pigott DP, Brady OJ, Marinho F, Deshpande A, Longbottom J, Browne AJ, Kraemer MUG, O'Reilly KM, Hombach J, Yactayo S, de Araùjo VEM, da Nóbrega AA, Mosser JF, Stanaway JD, Lim SS, Hay SI, Golding N, and Reiner RC (2017). Estimating global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis. *The Lancet Infectious Diseases*. Epub Aug 16 2017. doi: 10.1016/S1473-3099(17)30419-X.

Shearer FM, Longbottom J, Browne A, Pigott DM, Brady OJ, Kraemer MUG, Marinho F, Hombach J, Yactayo S, de Araùjo VEM, da Nóbrega AA, Fullman N, Ray SE, Mosser JF, Stanaway J, Lim SS, Reiner RC, Moyes CL, Hay SI, and Golding N. (2018). Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. *The Lancet Global Health*. Epub Jan 18 2018. doi: 10.1016/S2214-109X(18)30024-X.

Conference abstracts during DPhil

Shearer FM, Moyes CL, Golding N, and Hay SI. Improving risk mapping for vector-borne, zoonotic diseases. *American Society for Tropical Medicine and Hygiene Annual Meeting*.

November 2015. Philadelphia, United States of America. Poster.

Shearer FM, Moyes CL, Golding N, and Hay SI. Improving risk mapping for vector-borne, zoonotic diseases. *Royal Society for Tropical Medicine and Hygiene Research in Progress Meeting*. December 2015. London, United Kingdom. Poster.

Shearer FM, Golding N, Hay SI, and Moyes CL. Mapping the risk of *Plasmodium knowlesi* malaria. *The International Society for Neglected Tropical Diseases Bites Conference*. March 2016. London, United Kingdom. Oral presentation.

Shearer FM, Moyes CL, Hay SI, and Golding N. Predicting geographical variation in human infection risk for vector-borne, zoonotic pathogens. *British Ecological Society Annual Meeting*. December 2016. Liverpool, United Kingdom. Oral presentation.

Original manuscripts in parallel to thesis

Mylne AQ, Pigott DM, Longbottom J, **Shearer F**, Duda KA, Messina JP, Weiss DJ, Moyes CL, Golding N, and Hay SI (2015). Mapping the zoonotic niche of Lassa fever in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. Epub Jun 17 2015. doi: 10.1093/trstmh/trv047.

Kraemer MU, Sinka ME, Duda KA, Mylne AQ, **Shearer FM**, Barker CM, Moore CG, Carvalho RG, Coelho GE, Van Bortel W, Hendrickx G, Schaffner F, Elyazar IR, Teng HJ, Brady OJ, Messina JP, Pigott DM, Scott TW, Smith DL, Wint GR, Golding N, and Hay SI (2015). The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *eLife*. Epub Jun 30 2015. doi: 10.7554/eLife.08347.

Kraemer MU, Sinka ME, Duda KA, Mylne A, **Shearer FM**, Brady OJ, Messina JP, Barker CM, Moore CG, Carvalho RG, Coelho GE, Van Bortel W, Hendrickx G, Schaffner F, Wint GR, Elyazar IR, Teng HJ, and Hay SI. (2015) . The global compendium of *Aedes aegypti* and *Ae. albopictus* occurrence. *Scientific Data*. Epub Jul 7 2015. doi: 10.1038/sdata.2015.35.

Pigott DM, Millear AI, Earl L, Morozoff C, Han BA, **Shearer FM**, Weiss DJ, Brady OJ, Kraemer MU, Moyes CL, Bhatt S, Gething PW, Golding N, and Hay SI (2016). Updates to the zoonotic niche map of Ebola virus disease in Africa. *eLife*. Epub Jul 14 2016. doi: 10.7554/eLife.16412.

Messina JP, Kraemer MU, Brady OJ, Pigott DM, **Shearer FM**, Weiss DJ, Golding N, Ruktanonchai CW, Gething PW, Cohn E, Brownstein JS, Khan K, Tatem AJ, Jaenisch T, Murray

CJ, Marinho F, Scott TW, and Hay SI. (2016). Mapping global environmental suitability for Zika virus. *eLife*. Epub Apr 19 2016. doi: 10.7554/eLife.15272.

Kraemer MUG, Faria NR, Reiner Jr RC, Golding N, Nikolay B, Stasse S, Johansson MA, Salje H, Faye O, Wint GRW, Niedrig M, **Shearer FM**, Hill SC, Thompson RN, Bisanzio D, Taveira N, Nax HH, Pradelski BSR, Nsoesie EO, Murphy NR, Bogoch II, Khan K, Brownstein JS, Tatem AJ, de Oliveira T, Smith DL, Sall AA, Pybus OG, Hay SI, and Cauchemez S. (2016). Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015-2016: a modelling study. *The Lancet Infectious Diseases*. Epub Dec 22 2016. doi: 10.1016/S1473-3099(16)30513-8.

Wiebe A, Longbottom J, Gleave K, **Shearer FM**, Sinka ME, Massey CN, Cameron E, Bhatt S, Gething PW, Hemingway J, Smith DL, Coleman M, and Moyes CL. (2017). Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. *Malaria Journal*. Epub Feb 20 2017. doi: 10.1186/s12936-017-1734-y.

Longbottom J, Browne AJ, Pigott DM, Sinka ME, Golding N, Hay SI, Moyes CL, and **Shearer FM** (2017). Mapping the spatial distribution of the Japanese encephalitis vector, *Culex tritaeniorhynchus* Giles, 1901 (Diptera: Culicidae) within areas of Japanese encephalitis risk. *Parasites and Vectors*. Epub Mar 16 2017. doi: 10.1186/s13071-017-2086-8.

Statement of contributions

My supervisors, Dr Catherine Moyes, Dr Nick Golding and Professor Simon Hay provided guidance and technical help throughout the planning, analysis and writing process for each chapter. Additional contributions for each chapter are listed below.

Chapter 2: Dr Catherine Moyes provided occurrence datasets and species range maps. Zhi Huang, Dr Daniel Weiss, and Dr Catherine Moyes designed and constructed intact and disturbed forest data surfaces.

Chapter 3: Dr Robert Reiner Jr acted as a co-supervisor for this work, providing guidance throughout. Aniruddha Deshpande, Dr David Pigott and Maria Devine assisted with the extraction of outbreak response and preventive vaccination campaign data. Dr Valdelaine Araùjo, Dr Aglaêr da Nóbrega (Ministry of Health, Brazil), and Dr Sergio Yactayo (World Health Organization), provided unpublished vaccination data. Dr Fatima Marinho, Dr Valdelaine Araùjo, Dr Aglaêr da Nóbrega, Dr Jonathan Mosser, Dr Sergio Yactayo, Dr Joachim Hombach, Professor Stephen Lim, Dr Oliver Brady, Dr David Pigott, Dr Jeffrey Stanaway, and Dr Kathleen O'Reilly provided guidance during the analysis phase. Professor Stephen Lim also provided data for calculating the DTP3 bias correction terms.

Chapter 4: Annie Browne, Joshua Longbottom, and Maria Devine assisted with the extraction and collation of yellow fever occurrence data. Dr Sergio Yactayo (World Health Organization), Dr Valdelaine Araùjo, and Dr Aglaêr da Nóbrega (Ministry of Health, Brazil) provided expertise on yellow fever epidemiology. Dr Oliver Brady, Dr David Pigott, Dr Jeffrey Stanaway, Dr Robert Reiner Jr, Dr Moritz Kraemer and Dr Jonathan Mosser provided guidance during the planning and analysis phases.

Contents

Acknowledgements **ii**

Abstract **iii**

Contents **viii**

List of Figures **x**

List of Tables **xii**

List of Abbreviations **xiii**

INTRODUCTION **1**

1 Thesis introduction and overview **1**

1.1 Vector-borne, zoonotic pathogens and public health 1

1.2 Infectious disease risk mapping 2

1.3 Methods for infectious disease mapping 4

1.4 Niche mapping methods for occurrence-only data 7

1.5 Predicting suitability for disease transmission using boosted regression trees 8

1.6 Accounting for spatial bias in occurrence data 10

1.7 Incorporating intervention coverage into disease distribution mapping . 12

1.8 Aims and structure of the thesis 13

2 *Plasmodium knowlesi* malaria risk mapping
Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria **15**

2.1 Introduction 16

2.2 Methods 18

2.3 Results 30

2.4 Discussion 34

2.5 Conclusions 37

3 Yellow fever vaccination mapping
Estimating global yellow fever vaccination coverage from 1970 to 2016 38
3.1 Background to yellow fever . . . 38
3.2 Introduction . . . 40
3.3 Methods . . . 41
3.4 Results . . . 54
3.5 Discussion . . . 56
3.6 Conclusions . . . 62

4 Yellow fever risk mapping
Modelling the existing and potential global infection risk zones of yellow fever 63
4.1 Introduction . . . 64
4.2 Methods . . . 66
4.3 Results . . . 84
4.4 Discussion . . . 86
4.5 Conclusions . . . 90

DISCUSSION 93

5 General discussion 93
5.1 Recapitulation . . . 93
5.2 Methodological discussion . . . 94
5.3 Translating maps into policy . . . 101
5.4 General conclusions . . . 104

References 107

Appendices 121

A *Plasmodium knowlesi* malaria risk mapping 123
A.1 Plots of model covariates . . . 123
A.2 Species model uncertainty plots and top predictors . . . 127
A.3 *Plasmodium knowlesi* risk model uncertainty and top predictors . . . 137

B Yellow fever vaccination coverage 142
B.1 Yellow fever vaccination coverage by age group 1970-2010 . . . 142

C Yellow fever risk mapping 146
C.1 Supplementary plots of model inputs and outputs . . . 146

List of Figures

1.1 Nidus of pathogen transmission to humans . . . 3
1.2 Global YF risk map, 2010 . . . 4
1.3 Diagram of single decision tree . . . 10

2.1 Overview of *P. knowlesi* malaria risk mapping methods . . . 19
2.2 *P. knowlesi* malaria occurrence data . . . 22
2.3 *P. knowlesi* malaria vector distribution maps . . . 31
2.4 *P. knowlesi* malaria reservoir distribution maps . . . 32
2.5 *P. knowlesi* malaria risk map . . . 33

3.1 YF vaccination coverage from 1970 to 2016 . . . 55
3.2 YF vaccination coverage scenarios . . . 56
3.3 YF vaccination coverage by age group 2016 . . . 58

4.1 YF vaccination coverage versus *Ae. aegypti* . . . 64
4.2 Overview of YF risk mapping methods . . . 67
4.3 YF occurrence data . . . 68
4.4 YF candidate reservoir distribution map . . . 76
4.5 YF distribution within existing risk zones. . . . 91
4.6 Receptivity to YF transmission within potential risk zones. . . . 92

A.1 *P. knowlesi* land cover covariates . . . 124
A.2 *P. knowlesi* environmental covariates . . . 125
A.3 *P. knowlesi* socioeconomic covariates . . . 126
A.4 Uncertainty plots for *M. nemestrina* model . . . 130
A.5 Uncertainty plots for *M. leonina* model . . . 131
A.6 Uncertainty plots for *M. fascicularis* model . . . 132
A.7 Uncertainty plots for *An. dirus* model . . . 133
A.8 Uncertainty plots for the Dirus Complex model . . . 134
A.9 Uncertainty plots for the Leucosphyrus Complex model . . . 135
A.10 Uncertainty plots for the Leucosphyrus Group model . . . 136
A.11 Map of zoonotic *P. knowlesi* malaria transmission limits . . . 138
A.12 Uncertainty plot for *P. knowlesi* malaria risk model . . . 138

A.13 Evaluation of predictions outside MBS 139
A.14 Marginal effect plots for predictors of *P. knowlesi* malaria risk 140
A.15 Unmasked *P. knowlesi* malaria risk map. 141

B.1 YF vaccination coverage by age group in 1970 143
B.2 YF vaccination coverage by age group in 1980 143
B.3 YF vaccination coverage by age group in 1990 144
B.4 YF vaccination coverage by age group in 2000 144
B.5 YF vaccination coverage by age group in 2010 145

C.1 Environmental and biological covariates supplied to the YF risk model. . 147
C.2 Land cover covariates (and elevation) supplied to the YF risk model. . . 148
C.3 Environmental covariates supplied to the YF reservoir models. 149
C.4 YF model integration points. 150
C.5 Average and contemporary YF vaccination coverage. 151
C.6 Marginal effect plots for the most influential covariates for the YF risk model. 152
C.7 YF model uncertainty. 153

List of Tables

2.1 *P. knowlesi* model covariates 25

3.1 Online YF vaccination and demographic data sources 42
3.2 Processing of preventive and outbreak response YF vaccination data . . 43
3.3 Summary of preventive and outbreak response YF vaccination data . . 45
3.4 Number of people still requiring YF vaccination in 2016 57

4.1 YF risk model covariates 70
4.2 Non-human primate species with evidence of YFV infection 72

A.1 Relative influence of predictors for *M. nemestrina* model 127
A.2 Relative influence of predictors for *M. leonina* model 127
A.3 Relative influence of predictors for *M. fascicularis* model 128
A.4 Relative influence of predictors for *An. dirus* model 128
A.5 Relative influence of predictors for the Dirus Complex model 128
A.6 Relative influence of predictors for the Leucosphyrus Complex model . 129
A.7 Relative influence of predictors for the Leucosphyrus Group model . . . 129

List of Abbreviations

ACTs . . . . . artemisinin-based combination therapies

AUC . . . . . area under the receiver operator curve statistic

BRT . . . . . boosted regression trees

CAR . . . . . Central African Republic

CDC . . . . . Centers for Disease Control and Prevention

DALYs . . . . . disability-adjusted life years

DONs . . . . . Disease Outbreak News reports

DRC . . . . . Democratic Republic of the Congo

DTP3 . . . . . diphtheria-tetanus-pertussis-containing vaccine

EVI . . . . . enhanced vegetation index

GBD . . . . . Global Burden of Diseases, Risk Factors and Injuries Study

GBIF . . . . . Global Biodiversity Information Facility

GPW . . . . . Gridded Population of the World

IRS . . . . . indoor residual spraying

ITNs . . . . . insecticide-treated bed nets

IUCN . . . . . International Union for Conservation of Nature

JRF . . . . . Joint Reporting Form

LST . . . . . land surface temperature

MBS . . . . . Malaysia, Brunei, and Singapore

MESS . . . . . multivariate environmental similarity surface

MODIS . . . . . moderate resolution imaging spectrometer satellite

PPM . . . . . Poisson point process model

SDM . . . . . species distribution modelling

TCB . . . . . tasselled cap brightness

TCW . . . . . tasselled cap wetness

UNICEF . . . . . United Nations International Children's Emergency Fund

UNWPP . . . . . United Nations World Population Prospects

WERs . . . . . Weekly Epidemiological Records

WHO . . . . . World Health Organization

YF . . . . . yellow fever

YFV . . . . . yellow fever virus

1

Thesis introduction and overview

1.1 Vector-borne, zoonotic pathogens and public health

Vector-borne, zoonotic pathogens impose a substantial global burden on public health, with nearly 2.2 million disability-adjusted life years (DALYs) attributed to leishmaniasis, trypanosomiasis (African and American), and yellow fever (YF) alone in 2015 [1]. This burden is most heavily borne by people living in low-resource settings serviced by comparatively weaker public health systems. Understanding the spatial variation in risk of human infection and/or identifying areas at risk of disease establishment and spread are important areas of public health research.

Vector-borne, zoonotic disease systems are complex. They involve a minimum of four species (animal reservoir host, human host, vector, and pathogen), and sometimes many more. Vector-borne, zoonotic pathogens are typically maintained and amplified within a nonhuman host-vector cycle, before tangential transmission to domestic animals or humans, often involving additional 'bridge' vectors. A wide variety of vector classes exist, and they transmit pathogens between vertebrate hosts via a range of mechanisms. Biological arthropod vectors, such as mosquitoes or ticks, become infected when they blood feed from an infectious vertebrate host. After a period of incubation (where the pathogen replicates inside the vector), the vector becomes capable of transmitting the infection to another vertebrate host – including susceptible human hosts [2]. In

some cases, humans act as 'dead end' hosts, where sufficient volumes of the infectious agent are not reached within peripheral blood to re-infect vectors, for example in Japanese encephalitis and West Nile virus infections [3, 4]. In other cases, human-vector-human transmission can occur, such as when YF virus (YFV) is introduced into human populations [5].

To persist, vector-borne, zoonotic pathogens not only require suitable environmental conditions for their own biological processes, but also to support the presence of susceptible host and vector species in sufficient abundance [6]. The spatial distribution of such diseases is restricted to where all species involved in the transmission cycle coincide (Figure 1.1), and thus tends to be spatially heterogeneous. The epidemiology of vector-borne, zoonotic diseases can also be strongly impacted by anthropogenic activities such as deforestation, urbanisation and agriculture [6, 7]. These activities can create fragmented landscapes that increase interaction between reservoir hosts, vectors and humans, and the risk of tangential transmission to humans [8]. For spatial epidemiology to capture the full complexity of disease transmission systems with multiple vector and/or reservoir species, it is important to consider the current and potential distributions of all component species, as well as environmental and socioeconomic drivers of disease transmission. Developing the most accurate mapping approaches to vector-borne, zoonotic diseases is an important area of public health research, not only because they can help to reduce existing disease burden from these types of pathogens, but because many are emerging infections, and are thus recognised as increasing threats to human and animal health, as well as global economic and social stability [9].

1.2 Infectious disease risk mapping

Infectious disease risk mapping is the process of characterising geospatial variation in disease risk, often using statistical modelling techniques. The ultimate goal is to reduce the impact of disease on human health by providing an evidence-base for public health decision-making. Maps of infectious disease are used in many areas of public health; they provide baseline estimates of disease limits [10], they can be used to estimate the population at risk of infection [11], and the clinical burden of those infections [12]; and they

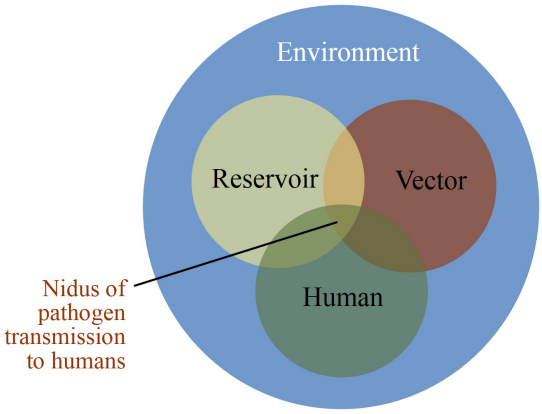


Figure 1.1: Vector-borne, zoonotic pathogen transmission to humans. How vector, reservoir, and human populations intersect within a permissive environment to enable pathogen transmission to humans. Figure adapted from Reisen [6].

are helpful tools for the spatial targeting of surveillance and interventions [13]. Disease risk mapping can also be used to identify areas at risk of disease introduction or spread [14, 15], and to understand how disease distributions may change under future environmental or socioeconomic conditions. For example, numerous mapping exercises have sought to predict the impact of climate change on the future distribution of dengue [16–18].

Over recent decades, the field of infectious disease mapping has been facilitated by advancements in computing and statistical modelling, as well as the increasing availability of high-resolution, large-scale data on environmental and socioeconomic risk factors for disease transmission. Models initially relied heavily on environmental variables, which in part, was due to the development of remote-sensing technology that enabled high-resolution (*e.g.* 100×100 m) global mapping of environmental variables, such as temperature and rainfall [19]. For many vector-borne and/or zoonotic pathogens, whose distributions depend on the limits of suitable habitats for reservoir and/or vector species, such models can be very informative. Disease distributions are also driven, however, by socioeconomic risk factors, such as population density and economic development [12, 20]. Whilst it is more difficult to quantify spatial variation in these factors accurately, such datasets are increasingly being developed and incorporated into geospatial models [21]. Most recently, disease modellers have become optimistic about the potential of 'big data' for infectious disease surveillance [22, 23], and how such data might feed into evolving risk maps, updated in real-time *via* automated mapping systems [24].

1.3 Methods for infectious disease mapping

A variety of approaches have been developed for mapping infectious diseases. At their most basic, disease maps can display raw epidemiological data (such as counts of reported cases), typically aggregated to political units [25]. More elaborate 'data maps' have been constructed that take into account a range of epidemiological information. Brady and colleagues refined the global spatial limits of dengue virus transmission by assigning each country a dengue status score based on available evidence of disease presence/absence, including case data and expert opinion [26]. The World Health Organization's (WHO) traveller vaccination-advice map for YF (Figure 1.2) categorised geographic areas as 'endemic', 'transitional', 'low risk' or 'no risk' based on various evidence, including expert opinion that selected specific thresholds for environmental variables thought to influence transmission [27].

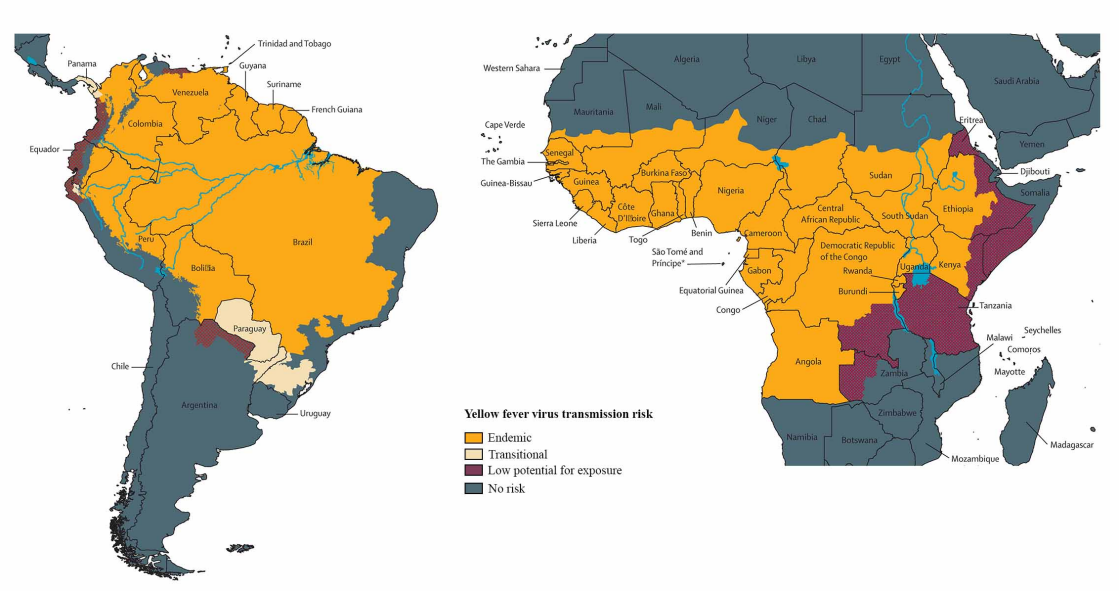


Figure 1.2: Global yellow fever (YF) risk map, 2010. Areas with risk of YF virus transmission in Africa and South America in 2010, generated for the Informal WHO Working Group on Geographic Risk for Yellow Fever. Figure adapted from Jentes and colleagues [27].

Since the distributions of infectious diseases are rarely spatially homogeneous across national or sub-national boundaries, these range maps are most useful when combined with modelling approaches [12]. For example, the dengue 'evidence consensus' map produced by Brady and colleagues [26] was used to define the geographic limits of a

dengue risk model [12]. Additionally, model-derived maps tend to be more powerful tools for supporting public health decision-making because: they can account for various sources of bias in the data, provide information about uncertainty, quantify relationships between disease transmission and potential risk factors, and help to fill gaps where surveillance data are incomplete. Two groups of models are widely employed for predicting and understanding patterns of infectious disease transmission: mechanistic and statistical models.

1.3.1 Mechanistic disease models

Mechanistic (or biological) models mathematically describe the biological processes underpinning disease transmission [28]. Sometimes these models can be linked to spatial environmental data and mapped to geographic areas. This has been done for the well-studied relationship between temperature and vector-borne disease transmission. For vector-borne disease transmission to occur, temperatures must be within a suitable range for the persistence of host and vector species, and for pathogen replication within the poikilothermic vector. Recent studies have used mechanistic models that capture these two mechanisms by which temperature mediates the intensity of dengue and malaria transmission, and combined them with remotely-sensed temperature datasets to produce spatially and temporally gridded estimates of temperature suitability for transmission [26, 29]. It has also been demonstrated that an index of temperature suitability is a better predictor of malaria risk than temperature alone [30].

Whilst there are many advantages to mechanistic models in terms of understanding the underlying processes determining a species' range and abundance (*i.e.* highly interpretable ecologically), they are difficult to construct, in part because they require very specific data on a species' traits and ecological requirements, as well as extensive field and laboratory validation [31]. Unfortunately, the biological knowledge required to mechanistically model all of the transmission process, in all existing scenarios, is not available for most pathogen systems, which would be required to produce a global map. Thus statistical models tend to be a more viable option because they implicitly

incorporate these processes by relating data (representing current disease distributions) to environmental conditions.

1.3.2 Statistical disease models

Statistical approaches to infectious disease mapping vary depending on the quantity and type of data available (*i.e.*, whether they represent disease prevalence, incidence, or occurrence), the eco-epidemiology of the disease, and the type of information required from the output. Rarely is sufficient data available for maximally useful epidemiological metrics to be mapped, for example spatial variation in prevalence [32, 33]. When such high quality data do exist, model-based geostatistics is considered the preferred method for infectious disease mapping [34]. These models can not only combine various disease, covariate and demographic data types, but also consider spatial relationships between data points, and if Bayesian approaches are employed, they can robustly quantify uncertainty in the resulting predictions. Prevalence of *P. falciparum* malaria has been mapped using model-based geostatistics and parasite rate surveys that comprehensively locate the disease in time and space, as well as report the infected fraction of the sampled local population [32].

Most often, high-quality epidemiological data are only available for limited locations (*i.e.* not spatially representative) and times, and the information within these data must be simplified to match more abundant data types. For many diseases of public health importance, the most abundant data available for mapping consist of locations and times where the disease has been reported, without corresponding information about where it is absent, or more or less prevalent [34]. This is referred to as occurrence, occurrence-only or presence-only data. Whilst disease presence/absence data arising from planned surveys (*i.e.* where the selection of survey sites are not biased by probability of disease presence) are highly informative, they are rare because they require significantly greater effort and expenditure to obtain.

Disease occurrence records are typically collected from a variety of sources, including manual searches of published scientific articles and less formal records such as those reported by ProMed mail archives [35] and Disease Outbreak News reports (DONs) [36,

37]. These sources tend to report disease occurrence on a range of spatial scales, and based on variable diagnostic evidence. Disease occurrences determined by PCR-based techniques are preferred for mapping, as diagnoses based on symptoms or serological evidence may confound closely related pathogens. Unfortunately, due to limitations of health care systems in many countries, costly PCR-based diagnoses are relatively rare.

For diseases of public health importance where the only data available for mapping are occurrence data, including those studied in this thesis, the most viable option for mapping is a niche modelling approach.

1.4 Niche mapping methods for occurrence-only data

Niche mapping, also known as species distribution modelling (SDM), is a type of correlative modelling. Co-opted from ecology (as it is the practical, analytical application of ecological niche theory [38–40]), niche mapping has become widely applied in epidemiology to produce predictive maps of reservoir, vector and disease distributions from reports of species/disease occurrence [41].

Niche models search for similarities between the environments associated with sets of observed species presences and the environmental (and socioeconomic) conditions associated with the rest of the geographic area under consideration [42]. In the same way that species distribution maps characterise habitats suitable for a given species, disease distribution maps describe environmental suitability for human infection based upon the climatic, biological and socioeconomic characteristics of those locations where the disease has been reported. The success of this approach generally requires that pathogens fall within transmission categories inherently linked to the environment, such as vector-, soil- or water-borne pathogens.

Niche models are typically fit using regression techniques derived from statistical or machine learning algorithms. The simplest are generalised linear models (GLMs), but these have limited flexibility due to their parametric form. More elaborate regression techniques have been developed to allow the modelling of complex ecological responses, including generalised additive models (GAMs) and multivariate adaptive regression

splines (MARS), as well as machine learning techniques, such as maximum entropy models (MAXENT) and boosted regression trees (BRTs) [43].

When applied to presence-absence data (such as the results of a systematic survey), these models generally use a binomial likelihood to represent the probability of a species being present at a given observation. Since most data available for species and infectious disease mapping are occurrence-only, standard practice in the SDM and disease mapping fields since their inception has been to supplement occurrence records with background or pseudo-absence points to represent areas where the species or disease has not been reported, and then model the presence/background label for each point as a draw from a Bernoulli likelihood. A variety of approaches have been proposed for generating these background points. Some studies use a large number of sites (*e.g.* 10,000) drawn randomly from within the study region or based on a minimum distance to occurrence data [44]. A number of disease mapping studies have randomly selected background data at a density inversely proportional to a disease status score at a given location (based on evidence for disease presence/absence), excluding sites a minimum distance from any occurrence site [12, 45]. Most recently, background data selection methods have been developed to account for reporting rate bias in the occurrence data, which will be discussed in more detail in Section 1.6.

Once a model has been fit to the occurrence and background data, it can be combined with continuous spatially gridded covariate data to predict suitability for disease transmission for every grid square within the geographic area of interest.

1.5 Predicting suitability for disease transmission using boosted regression trees

Elith and colleagues published a comprehensive comparison of 11 niche modelling methods, and found that machine learning approaches out-performed other regression techniques with regards to predictive performance, and of these, BRTs were marginally superior at fitting complex biological responses [43]. The BRT modelling framework has since been adopted to map the distribution of a number of diseases, including dengue

[12], the leishmaniases [45], Ebola virus disease [46], as well as disease vectors [47–50]. I therefore elected to use the BRT approach for the analyses performed in this thesis.

Boosted regression trees combine two algorithms – regression trees and boosting – to examine the relationship between the predictor variables and the response data. The regression-tree algorithm repeatedly splits the data into two groups using randomly selected predictor variables for each split (Figure 1.3). A single regression tree is likely to contain classification errors, so model accuracy is improved through the boosting algorithm, which iteratively calls the regression-tree algorithm, constructing a set or 'ensemble' of simple tree models and combining them for prediction. For BRT, boosting is a forward stage-wise process that, with each step, gradually prioritises subsets of the data that are modelled poorly by the existing set of trees. Boosting can be seen as an optimisation process, with each step minimising the loss in model predictive performance, measured by a 'loss function'. (Residual deviance was used as the loss function for the analyses performed in this thesis). The first tree is the one that maximally reduces the loss function, and for each subsequent step, a new tree is fitted to the residuals of the preceding set of trees (*i.e.* variation in the response that is not so far explained by the model). The final BRT model consists of a linear combination of many trees (usually hundreds to thousands) that can be thought of as an additive regression model in which each term is a simple tree model. To avoid overfitting, cross-validation is used during model-building to test predictive accuracy on withheld portions of the data and to identify the optimal number of trees [51].

The BRT approach has the ability to fit complex nonlinear responses and, due to the hierarchical structure, automatically identifies and models high-dimensional interactions between explanatory variables. They are also very flexible in terms of the number and type (*e.g.* continuous or discrete data) of predictors that can be accommodated to inform predictions, and since irrelevant predictors will not be selected, no prior assumptions need to be made regarding the effect of potential predictors on disease transmission.

Boosted regression trees provide a flexible, high-performing technique for fitting niche models, but this approach requires further extension in order to map diseases for which mapping is complicated by reporting rate biases, and public health interventions.

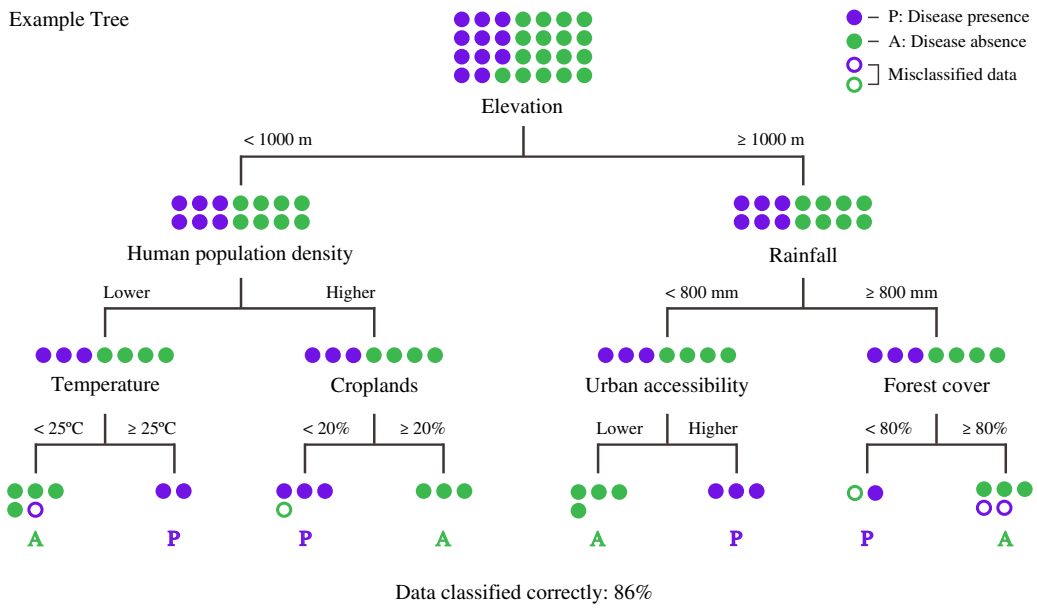


Figure 1.3: Simplified diagram of a single decision tree. Disease presence and absence data is repeatedly split into two groups, using a randomly selected variable (such as elevation) for each split. The objective is to separate the presences and absences at the terminal nodes of the tree. In this example, the regression-tree algorithm has classified 86% of the data correctly. Misclassified data will be prioritised as the algorithm builds the next tree. Figure adapted from illustration by Erik Vrielink [52].

1.6 Accounting for spatial bias in occurrence data

The vast majority of occurrence datasets used in species and disease distribution mapping are subject to spatial bias in reporting rates. For example, in ecological datasets, areas near roads and human population centers may be more likely to be surveyed than other sites [53]. The same is true for disease occurrence data, where occurrence of a disease may be more likely to be reported in more prosperous countries (with strong health systems) and/or major cities (increasing the probability of a least one case being reported). If unaccounted for, this bias can translate into environmental bias in the fitted model and result in a prediction map that is strongly skewed by survey/reporting effort rather than the true distribution of the species or disease, *i.e.* elevated risk/suitability predictions in the areas most likely to report [54, 55].

Thus choosing a method to handle this type of bias is a key step in the modelling process, which may have a significant impact on model predictions. A number of

approaches to the problem are described in the ecological literature [56]. One approach involves attempting to remove bias from the occurrence data, by either discarding or down-weighting records in over-sampled geographic regions [57]. This can also be achieved by using an environment filter to discard occurrence records clumped under similar environmental conditions [58]. In most cases, however, the discarding of records is not a viable option for already limited occurrence datasets. An alternative is to select background data that reflect the same bias as the occurrence data [55]. The aim is to achieve the same environmental bias in both datasets, so the resulting model should identify suitable environments for the species within the sampled space, rather than just areas that experience more sampling. Since information on the exact distribution of survey effort is generally incomplete or unavailable, a method for estimating survey effort is to collect occurrence records for a broad biological group (such as birds or mosquitoes), where the sampling methods are the same as those used for the target species. This approach does not eliminate reporting bias issues entirely but improved model performance has been demonstrated, when compared to using randomly selected background data [54, 55].

Spatial bias in reporting rates has only recently received consideration in disease distribution mapping and is yet to be comprehensively explored. For example, Escobar and colleagues used passive surveillance data to construct a sampling bias grid from which to generate background data for their model of rabies distribution in Chile, where active surveillance data were used as occurrence records [59]. Similarly, in their distribution model of Ebola virus disease, Pigott and colleagues generated background data by sampling 10,000 locations across their study area, biased towards more populous areas since, all else being equal, the probability of at least one case being reported will be higher in more populous areas [46]. Since the discussion of these issues first appeared in the literature, however, many papers have been published that do not address possible data biases at all [60–64]. Given the paucity of methods available, and a lack of published examples of their use, much scope exists for the development of novel approaches for dealing with disease reporting bias.

1.7 Incorporating intervention coverage into disease distribution mapping

In addition to environmental factors, the distributions of vector-borne, zoonotic pathogens are impacted by various public health interventions, such as insecticide spraying and vaccination. Due to limitations in data quality, volume and availability, these factors are more difficult to quantify spatially than environmental factors, and are rarely incorporated in geospatial models of disease risk [34]. Yet understanding the spatial distribution of interventions is important for producing the most accurate predictions of disease risk and for guiding public health decisions.

Maps of public health interventions have most frequently been used alongside disease risk maps. For example, subnational-level 'overlap' maps of interventions impacting lymphatic filariasis and onchocerciasis have been produced for individual countries and then considered alongside disease maps to identify areas of high disease risk that may not have received adequate intervention coverage [65, 66]. On a regional scale, data on the use of insecticide-treated bed nets (ITNs), access to artemisinin-based combination therapies (ACTs), and the provision of indoor residual spraying (IRS) have been used to produce time-series maps of coverage of these interventions across all malaria endemic African countries from 2000 to 2015. These were then used, in conjunction with models of *P. falciparum* infection prevalence and disease incidence, to quantify the role of ITNs, ACTs, and IRS in the decline of malaria from 2000 to 2015 [67].

More rarely has the impact of interventions been incorporated into actual model fitting, and to-date, not in niche models. Current niche modelling approaches predict disease risk primarily on the basis of environmental risk factors. Thus in areas where the environment is suitable, but high levels of disease control suppresses transmission, this approach may lead to over prediction of risk. Existing niche modelling approaches must be developed further in order to cater for diseases where public health interventions are expected to be key drivers of their distribution.

1.8 Aims and structure of the thesis

The aims of this thesis are to investigate the geographical distribution of two vector-borne, zoonotic diseases: *Plasmodium knowlesi* malaria and yellow fever (YF). To achieve these aims, there is a focus on the development of novel approaches for: mapping diseases whose distributions are impacted by multiple host and vector species, ameliorating spatial bias in disease reporting rates, and accounting for human vaccination coverage.

These diseases were selected for mapping for two primary reasons: their current importance to global public health, and the lack of existing knowledge of their geographical distributions.

Infection by the simian malaria parasite, *P. knowlesi*, can lead to severe and fatal disease in humans, and is the most common cause of malaria in parts of Malaysia. Despite being a serious public health concern, the geographical distribution of *P. knowlesi* malaria risk is poorly understood because the parasite is often misidentified as one of the human malarias. Human cases have been confirmed in at least nine southeast Asian countries, many of which are making progress towards eliminating the human malarias. Understanding the geographical distribution of *P. knowlesi* is important for identifying areas where malaria transmission will continue after the human malarias have been eliminated.

Yellow fever represents a substantial public health burden for 47 countries across Africa and the Americas, where it caused an estimated 5,800 deaths and 374,015 DALYs in 2016. During the course of this thesis, substantial YF outbreaks arose in Angola and Brazil, which, coupled with global shortages in vaccine stockpiles and the global expansion of YF's main urban vector, highlight a pressing need to assess present control strategies, and both the existing and potential geographical distributions of disease risk.

1.8.1 Structure

Chapter 2 investigates spatial variation in risk of human *P. knowlesi* infection across southeast Asia. The infection risk model for *P. knowlesi* malaria is based on a standard niche modelling approach, but incorporates a novel joint distribution model to leverage data from a number of host species. Additional improvements are made to previously

published approaches to enable the incorporation of temporally specific land cover data, and to deal with spatial reporting bias in the occurrence data.

Chapter 3 estimates YF vaccination coverage through time across all age cohorts in every district/municipality of countries at risk of YF, globally. These estimates are used to estimate the additional vaccination coverage needed to prevent further YF outbreaks, and they provide information needed to account for population immunity when estimating YF infection risk (*Chapter 4*).

Chapter 4 describes the development of a novel Poisson point process niche model, which is then used to predict human YF risk and demonstrates how vaccination coverage can be efficiently accounted for in disease niche models. The model incorporates information on the location of reports of YF infection, susceptibility of the population to infection, spatial bias in disease reporting and a range of environmental and biological variables thought to influence disease transmission.

Finally, *Chapter 5* provides a general discussion of the thesis and its implications for public health policy and the field of infectious disease mapping.

2

Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria

Contents

2.1 Introduction **16**
2.2 Methods **18**
2.3 Results **30**
2.4 Discussion **34**
2.5 Conclusions **37**

This chapter investigates the spatial variation in risk of human *P. knowlesi* malaria across southeast Asia. Using a standard niche modelling approach, I first produce species distribution models of known and putative reservoirs and vectors of *P. knowlesi* malaria. These maps are then used as explanatory variables, along with a range of environmental and socioeconomic factors, to model the spatial variation in risk of human *P. knowlesi* infection. The *P. knowlesi* reservoir and vector species distribution work is published in *Parasites and Vectors* [68] and the *P. knowlesi* malaria risk mapping work is published in *PLOS Neglected Tropical Diseases* [69].

2.1 Introduction

Plasmodium knowlesi, a simian malaria parasite transmitted to humans *via* anopheline mosquitoes, is a prime example of human disease involving multiple vector and animal reservoir species. Malaria cases caused by this parasite have been identified in at least nine southeast Asian countries. In many parts of Malaysia, *P. knowlesi* is the most common cause of malaria [70] and can cause severe and fatal disease [71–73].

A number of monkey species have been identified as natural hosts of *P. knowlesi* parasites, including *Macaca fascicularis*, and *M. nemestrina* [74, 75]. *Macaca leonina*, a close relative of *M. nemestrina*, is considered a putative host species. It is distributed further north than either *M. fascicularis* or *M. nemestrina*, as far as Myanmar, where human *P. knowlesi* infections have been reported [76]. Several anopheline mosquito species from the Leucosphyrus Group have been implicated as vectors of *P. knowlesi*, including *Anopheles dirus* [77] and *An. cracens* [78] in the Dirus Complex, and *An. latens* [79], *An. balabacensis* [80], and *An. introlatus* [81] in the Leucosphyrus Complex.

Despite the potential severity of infection [70, 72, 82–85], diagnostics that identify *P. knowlesi* are not routinely used. Unless blood samples are tested using expensive nested PCR-based diagnostics, cases of *P. knowlesi* are often misdiagnosed by microscopy as one of the human malarias, principally *P. malariae* or *P. falciparum* [86–88]. *Plasmodium knowlesi* is routinely considered as a potential cause of malaria cases in three countries: Malaysia, Brunei and Singapore (hereafter a region referred to as MBS), the latter two having already eliminated the human malarias. The bulk of *P. knowlesi* malaria cases have been reported in this region. Cases have also been reported in Cambodia, Indonesia, Myanmar, the Philippines, Thailand and Vietnam [85, 89, 90], but sampling has been limited and the full geographical extent of disease risk across most of the region, including within these countries, is unknown.

Understanding the geographical distribution of *P. knowlesi* is important to identify areas where residual malaria transmission could remain once the human malarias have been eliminated [91]. Many countries in southeast Asia, including Malaysia, the Philippines, Thailand and Vietnam, are currently in the process of eliminating the human

malarias, namely *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale* [92]. These parasites are primarily transmitted between humans *via* mosquitoes and are not commonly transmitted from other animals to humans. Current control measures that reduce the human malarias include mass anti-malarial drug administration, the provision of ITNs and IRS. These control measures do not, however, target parasite populations in the *P. knowlesi* reservoir host species so *P. knowlesi* populations will not be eliminated. Further, ITNs and IRS are unlikely to offer the same degree of personal protection to humans, or community protection through reductions in mosquito longevity, since the vectors for *P. knowlesi* bite and rest outdoors [79]. If the presence of *P. knowlesi* is not considered when elimination strategies are developed, the impact of elimination measures and reduction in overall malaria cases in these areas will not match projections.

The aim of this chapter is to produce the first map of the geographical distribution of *P. knowlesi* malaria. To achieve this, I employ a standard niche modelling approach that has previously been applied to the mapping of diseases, such as dengue [12], as well as their vector and/or animal host species [50]. However, modifications are made to this approach to address various sources of bias in the available data, to improve handling of wide-area data, and to allow temporal changes in land cover to be incorporated.

Since the spatial distribution of vector-borne, zoonotic diseases is restricted to locations where all species required for transmission co-occur, species distribution maps are first produced for all known and putative reservoirs and vectors of *P. knowlesi*. These include the main macaque species identified as natural hosts, and several anopheline mosquito species, all from the *Anopheles leucosphyrus* group, that have been implicated as vectors of *P. Knowlesi*. The reservoir and vector distribution maps are then used as explanatory variables, along with a range of environmental and socioeconomic factors, to model the spatial variation in risk of human *P. knowlesi* infections across southeast Asia.

Previous work has assessed the evidence for the spatial limits of *P. knowlesi* transmission [90] and the work presented in this chapter refines those limits and investigates the variation in risk within them. The maps of reservoir and vector species alone are useful for defining the limits of zoonotic transmission, but an index of disease risk cannot be extrapolated directly from reservoir/vector maps. This is because co-occurrence of

reservoir and vector species is necessary for transmission, but it is not always sufficient, as many other factors contribute [6]. The disease model developed below incorporates the relationships between *P. knowlesi* infection and the distributions of reservoir and vector species, along with a range of other potential risk factors, to produce fine-scale evidence-based predictions of relative zoonotic *P. knowlesi* transmission risk.

The final output provides an initial map that aims to identify locations where disease surveillance and epidemiological investigations would be most informative to improve our understanding of *P. knowlesi* malaria risk.

2.2 Methods

2.2.1 Overview

Maps predicting environmental suitability for each reservoir and vector species of *P. knowlesi*, and a map predicting risk of *P. knowlesi* transmission to humans, were generated at a 5 km resolution by fitting BRT model ensembles.

Three host macaque species distribution models were produced for *M. fascicularis*, *M. nemestrina* and *M. leonina*. Four vector species distribution models were produced for the Leucosphyrus Group, Leucosphyrus Complex, Dirus Complex and *Anopheles dirus* species. Each species (or complex or group) model used a set of geo-positioned occurrence points for the target species, a sampling bias dataset and 20 environmental covariate data surfaces. Annual data surfaces for forest cover and eight other land cover classes were used to account for substantial changes in forest cover that have occurred during the study period from 1990 to 2014.

A schematic of the processed followed to map *P. knowlesi* malaria disease risk is provided in Figure 2.1. The occurrence database contained records of *P. knowlesi* infections in humans, non-human primates, and mosquitoes, collated and geo-positioned from a variety of sources. A novel joint BRT model was fitted in order to leverage occurrence data from all three host species. The model was also supplied 19 gridded datasets of environmental, socioeconomic and biological explanatory covariates including the species distributions for *M. nemestrina*, *M. fascicularis* and the Leucosphyrus Group.

Reporting rate bias is likely to be more pronounced for *P. knowlesi* malaria than other diseases since significant resources are required to accurately diagnose infection and *P. knowlesi* infection is not routinely considered a possible cause of malaria in the region, with the exception of MBS. A disease model was therefore fitted using data only from MBS (the model training region), where reporting bias could be accounted for through the selection of background data [55]. This model was then used to predict the human risk of *P. knowlesi* infection across southeast Asia. To assess the model's predictive capacity outside its training region, its performance was tested on a set of disease presence and absence records from locations outside MBS. A multivariate environmental similarity surface (MESS) was also generated to identify regions where the model was required to extrapolate to environments not found within MBS, and therefore evaluate the appropriateness of inferring risk in those regions.

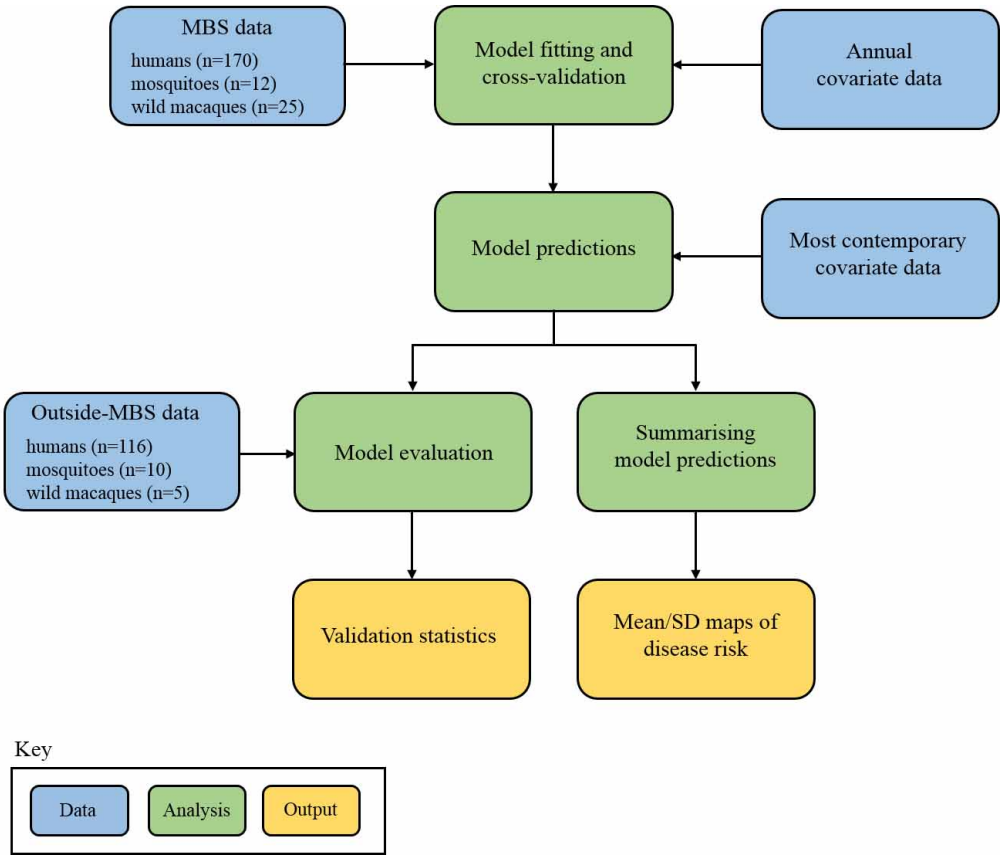


Figure 2.1: Schematic overview of the *P. knowlesi* malaria disease map methods. Blue boxes describe input data, green boxes denote analyses, and yellow boxes represent final outputs. MBS= Malaysia, Brunei and Singapore. SD = standard deviation.

2.2.2 Occurrence datasets

Reservoir and vector occurrences

The *M. fascicularis* and *M. nemestrina* datasets contained 1,115 and 1,025 occurrence records from 1999 to 2014, respectively. The *M. leonina* dataset was comprised of 450 records from 1992 to 2014. The Leucosphyrus Group dataset contained 571 occurrence records from 1982 to 2013. This included 523 records for the Dirus Complex (with 107 records for *An. dirus*) and 41 for the Leucosphyrus Complex. Each species occurrence record contained the coordinates in decimal degrees of a location where that species had been reported at least once in the specified year. All occurrence locations represented an area not greater than 5×5 km (points), with the exception of 33 records in the *M. leonina* dataset, which will be referred to hereafter as polygons.

The macaque sampling bias dataset contained 1,348 data points, representing the locations of all mammal surveys, within the study area, held by the Global Biodiversity Information Facility (GBIF) [93] from 1990 onwards. The anopheline sampling bias dataset contained 1,314 species level and 2,047 complex level data points, representing locations of malaria vector surveys, within the study area, held by the Malaria Atlas Project [94].

Plasmodium knowlesi infection occurrences

The *P. knowlesi* infection occurrence dataset contained records of *P. knowlesi* presence or absence for southeast Asia, obtained from reports in the published literature from 2004 to 2015. Each presence record contained the coordinates of a location less than 5×5 km (points) or an area larger than this (polygons) where a human, macaque or mosquito infection was confirmed using specific diagnostics that are able to distinguish *P. knowlesi* from the other *Plasmodium spp*. Each of the polygon records were either linked to a shapefile from the Food and Agriculture Organization (FAO) representing an administrative area (such as a province or district), or a bespoke polygon (such as a nature reserve). Presence records were excluded if *P. knowlesi* presence in the surrounding area (within 300 km radius) was not verified by a second, independent laboratory. Each absence record contained the coordinates for a site where an appropriate diagnostic for *P. knowlesi* was used, but no infections were found in a sample size of at least

500 individuals, or where no malaria cases were reported across an administrative division in 2012 [95].

The final dataset contained a total of 439 *P. knowlesi* occurrence records, consisting of 301 presence and 138 absence records. The occurrence dataset used for model fitting (records falling within MBS) totalled 198 records, corresponding to 62 unique point locations and 136 polygons (Figure 2.2B). The model fitting dataset consisted of human (166), macaque (23) and mosquito (9) occurrence records. The evaluation dataset (records falling outside MBS) totalled 66 presence and 65 absence records, comprising 29 points and 102 polygons (Figure 2.2A).

2.2.3 Explanatory covariates

Gridded data surfaces, at a 5 km resolution, of a variety of environmental, socioeconomic, and biological factors were supplied to the niche models as potential explanatory covariates. Table 2.1 lists each covariate, its data source and whether it was included in only the reservoir and vector models, only the disease model, or both. Covariates supplied to the reservoir and vector models included information on the mean and standard deviation values for each pixel for daytime land surface temperature (LST), enhanced vegetation indices (EVI), tasselled cap brightness (TCB) and tasselled cap wetness (TCW), as well as values for elevation, human population density and ten land cover classes (open shrublands, woody savannas, savannas, grasslands, permanent wetlands, croplands, and cropland/natural vegetation mosaic, urban and built-up, intact forest, and disturbed forest).

The LST, EVI, TCB and TCW data surfaces were derived from NASAs moderate resolution imaging spectrometer (MODIS) satellite imagery [96, 97]. They were computed from the original 1×1 km dataset and gap-filled using Weiss and colleagues' algorithm to model missing data caused by cloud cover [98].

Land cover covariates were derived from the International Geosphere-Biosphere land cover classification available within the MODIS MCD12Q1 dataset at a 500×500 m spatial resolution [99]. The 500×500 m categorical datasets were summarised to produce a value for proportional land cover of either open shrublands, woody savannas, savannas,

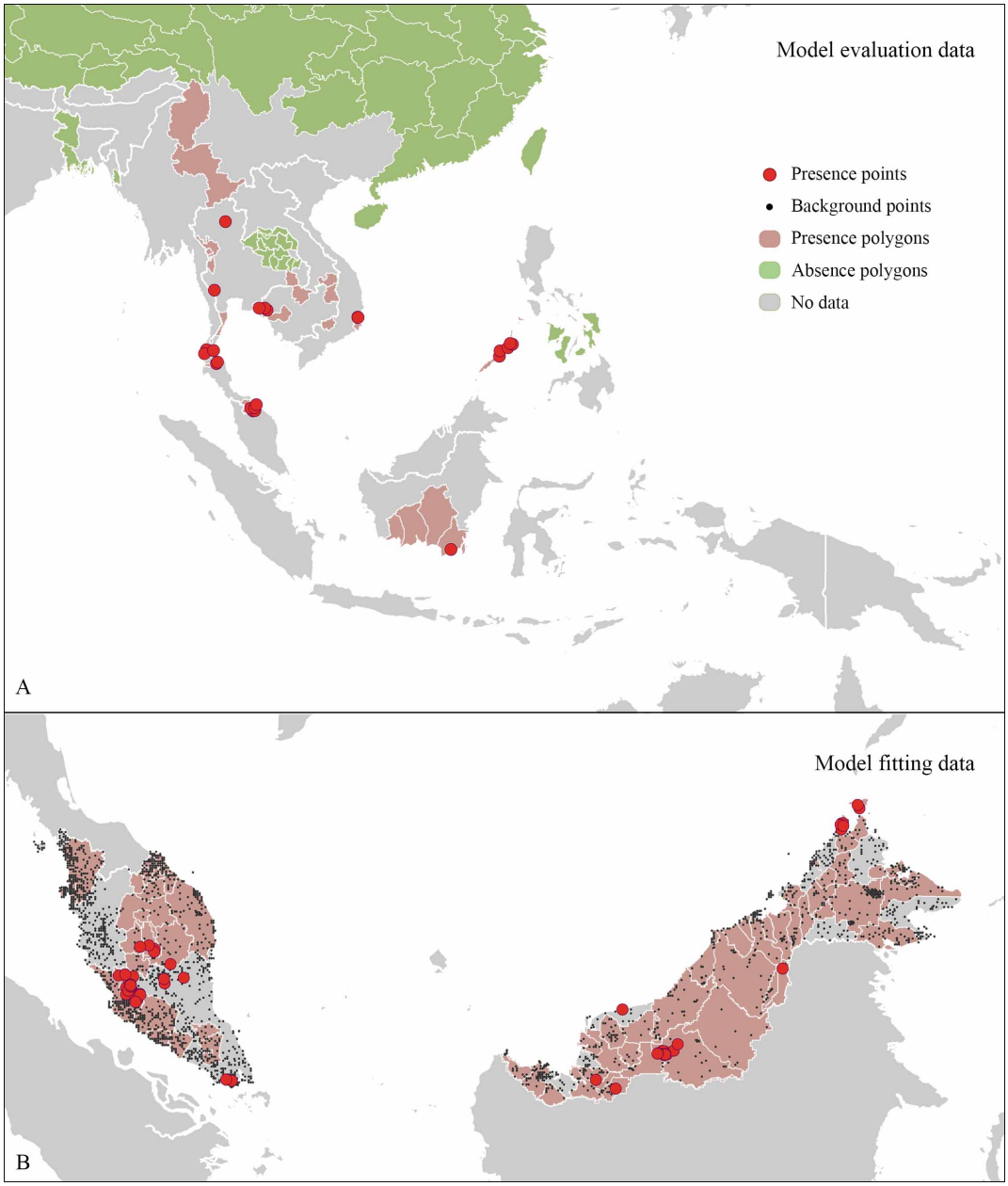


Figure 2.2: Occurrence data used for *Plasmodium knowlesi* malaria risk model fitting and evaluation A. Location of presence and absence points/polygons outside Malaysia, Brunei and Singapore used for model evaluation. **B.** Location of presence and absence points/polygons as well as background points from Malaysia, Brunei and Singapore used for model fitting.

grasslands, permanent wetlands, croplands, urban and built-up or cropland/natural vegetation mosaic within the larger 5×5 km grid cells. Annual data surfaces were calculated using data for each year from 2001 to 2012.

Separate intact forest and disturbed forest data surfaces were constructed for this project by Zhi Huang, Dr Dan Weiss and Dr Catherine Moyes. Annual data for the first five

forest classes in the International Geosphere and Biosphere land cover dataset were used to produce data surfaces that track the decline in forest cover and any areas of regrowth from 2001 to 2012. The values range from 0.0 (no forest cover) to 1.0 (complete forest cover). The development of these layers is described in detail by Moyes and colleagues [68] and additional information on covariates derived from satellite imagery and the subsequent processing of them, is also provided elsewhere [30].

Fine scale human population density data for 2010 was obtained from the WorldPop Project [21] and Gridded Population of the World (GPW) [100]. These two data sets were mosaicked together, with WorldPop data used in preference where they overlapped, as it is a more spatially dis-aggregated dataset. Then the sum of the 1×1 km pixels in each 5×5 km pixel was calculated to produce the final surface. The elevation covariate was derived from the Shuttle Radar Topography Mission dataset [101], with the original 90 m spatial resolution data summarised to produce a 5×5 km surface.

A number of the covariates listed above were also supplied to the disease model, including the following: elevation, tasselled cap wetness (mean and standard deviation), tasselled cap brightness (standard deviation), human population density and nine land cover classes (open shrublands, woody savannas, savannas, grasslands, permanent wetlands, croplands, cropland/natural vegetation mosaic, intact forest and disturbed forest). Covariates supplied to the disease model, but not to reservoir and vector models, included temperature suitability for *P. falciparum* transmission, urban accessibility, and the predicted species distributions for *M. nemestrina*, *M. fascicularis* and the Leucosphyrus Group.

It has been shown that an index of temperature suitability for malaria transmission that incorporates the relationships between temperature and both sporogeny and adult mosquito longevity is a better predictor of malaria risk than temperature alone [30]. For this reason a temperature suitability index for *P. falciparum* transmission [29] was included, in the absence of the biological data needed to develop a *P. knowlesi* transmission temperature suitability index. A *P. vivax* temperature suitability surface was also available, but the extrinsic incubation period for *P. knowlesi* is closer in duration to that of *P. falciparum* [102].

The urban accessibility index was calculated using data from the European Commission Joint Research Centre Global Environment Monitoring Unit giving the time required to travel from a given pixel to an urban settlement of 50,000 people or more *via* land or water-based transportation networks [103]. A reciprocal of this quantity was taken and re-scaled between 0 and 1 such that high values of urban accessibility index indicated less travel time to a city. More detailed information on the construction of the urban accessibility layer is provided elsewhere [12].

Annual data surfaces from 2001 to 2012 describing the distributions of *M. fascicularis*, *M. nemestrina*, and members of the Leucosphyrus Group, were included as biological covariates in the disease modelling procedure. The distribution maps for *M. leonina* were not included, since this species is not present within MBS.

I also explored whether covariates capturing net change in forest cover could be included to further account for the potential impact of deforestation on vector/reservoir distributions and disease risk. Data surfaces were constructed representing annual change in the proportion of intact and disturbed forest, *i.e.* for a hypothetical 5 x 5 km pixel, if the fraction of intact forest was 0.8 for 2001 and 0.6 for 2002, the value in the annual change surface for 2002 would be -0.2. These surfaces were supplied as covariates to preliminary reservoir and vector models, but were not selected as predictors, and were subsequently discarded from the final models.

Since the original data for the covariates came from a variety of sources, all covariate grids were standardised to ensure identical spatial resolution, extent and boundaries.

2.2.4 Model fitting

To carry out each niche mapping analysis, an ensemble of BRT models was fitted to each occurrence dataset using the R package 'gbm' [105]. The core setup used has been described previously [12, 45, 51]. For the work presented in this chapter, changes were made to previous methods to improve handling of absence data, sampling bias and polygon data, and to allow temporal changes in land cover to be incorporated. A novel joint distribution model was also fitted to the disease data to leverage additional infection data available from reservoir and vector species.

Table 2.1: *Plasmodium knowlesi* model explanatory covariates. List of explanatory covariates supplied to niche models, whether they were included in only the reservoir and vector models, only the disease model, or both, and their data source. Plots of each surface are provided in Appendices A.1, A.2, and A.3.

Covariates	Analyses	Date source
Open shrublands, woody savannas, savannas, grasslands, permanent wetlands, croplands, urban and built-up, and cropland/natural vegetation mosaic land cover classes (proportional)	Reservoirs and vectors Disease (except urban)	MODIS land cover product [99]
Intact forest cover (proportional)	Reservoirs and vectors Disease	MODIS land cover product [99] Intact Forest Landscape [104]
Disturbed forest cover (proportional)	Reservoirs and vectors Disease	MODIS land cover product [99] Intact Forest Landscape [104]
Elevation	Reservoirs and vectors Disease	Shuttle Radar Topography Mission [101]
Temperature suitability index for *Plasmodium falciparum* transmission	Disease	Gething *et al.*, 2011 [29]
tasselled cap wetness, a measure of surface moisture (mean and standard deviation)	Reservoirs and vectors Disease	Gap-filled MODIS satellite data [97, 98]
tasselled cap brightness, a measure of moisture on bare surfaces (mean and standard deviation)	Reservoirs and vectors Disease (standard deviation only)	Gap-filled MODIS satellite data [97, 98]
Land surface temperature (mean and standard deviation)	Reservoirs and vectors	Gap-filled MODIS satellite data [97, 98]
Enhanced vegetation index (mean and standard deviation)	Reservoirs and vectors	Gap-filled MODIS satellite data [97, 98]
Human population density	Reservoirs and vectors Disease	WorldPop [21] and Gridded Population of the World [100]
Urban accessibility	Disease	European Commission Joint Research Centre [103]
Species distributions for *Macaca fascicularis*, *M. nemestrina* and the Leucosphyrus Group	Disease	*Chapter 2*

Rather than exclusively using synoptic (averaged across time) covariate values for each of the occurrence locations irrespective of the occurrence date, annual land cover data surfaces were incorporated from 2001 to 2012. This was necessary to account for the substantial changes in land cover that have occurred in the region over the study period due to deforestation [106], which is hypothesised to have impacted the distribution of *P. knowlesi* reservoir and vector species [107]. Using the annual land cover data surfaces, occurrence data collected between 2001 and 2012 were matched with covariate values for the relevant year. Covariate values for 2001 were used for occurrence data prior to this date and covariate values for 2012 were used for post-2012 data. Final predictions of environmental suitability for each species were made for each set of explanatory covariates from 2001 to 2012. For the disease model, predictions were made using a complete set of the most contemporary covariate data surfaces.

Each occurrence dataset was supplemented with a large number of background records to represent areas where each reservoir/vector species or pathogen has not been reported as occurring. To cope with spatial bias in the occurrence data and thus improve model performance, background data were selected to reflect the same bias as in the occurrence data, as per Phillips and colleagues [55].

For the reservoir/vector species models, mammal and malaria vector occurrence records from within the study area were used as background data for the macaque and mosquito models, respectively. These species were chosen for the background datasets because the sampling methods are the same as those used for the target species.

For the disease model, the occurrence dataset was supplemented with a large number of background records to represent areas where *P. knowlesi* infection has not been reported within MBS. Six thousand background points were generated in total with the same proportion of human, macaque and mosquito background points as the presence dataset. Human infection background points were generated by randomly sampling across MBS, biased towards human population density, since more populous areas have a greater probability of reporting at least one case. This method was also used to generate background points for the mosquito infection data since all studies that looked for *P. knowlesi* infections in vector species, selected study locations based on the presence of human *P. knowlesi* cases in the immediate vicinity. Background points for the macaque infection data were randomly sampled from the *M. fascicularis* and *M. nemestrina* occurrence datasets, as well as the mammal survey dataset that reflected the bias in locations chosen for macaque studies. Covariate values for the specific times and locations of the background data were then extracted. Prior to covariate extraction, human and vector background points were assigned a year randomly sampled from the temporal distribution of presence points for each species.

To increase the robustness of model predictions and quantify model uncertainty, a number of sub-models (120 for the reservoir and vector models, 500 for the disease model) were fitted with each trained to a separate bootstrap dataset from the complete presence/background dataset. Each bootstrap had the same number of data points as the complete dataset and contained a minimum of ten presence and ten background points

(a minimum of five was used for reservoir and vector models). To incorporate uncertainty of the geographic location of occurrences linked to polygon locations, one 5×5 km pixel within each polygon was randomly selected for each bootstrap. This resulted in a Monte Carlo simulation to integrate across uncertainty in the spatial location of the samples, assuming that it is equally likely that the samples arose from each pixel. Likelihood weights were applied to the background points such that the sum of weights of the background data plus the number of absence records equalled the number of presence records for each bootstrap dataset. This has been shown to improve model predictive accuracy where unbiased presence-absence data are unavailable [44]. The algorithm hyper-parameters were set to the following values: cross-validation folds=10 (=5 for reservoir and vector models), tree complexity=4, learning rate=0.005, bag fraction=0.75, step size=10.

Since the disease occurrence dataset included data from humans, macaques and mosquitoes, a novel joint model was fitted for human, macaque and mosquito hosts that enabled all available infection data to be leveraged, whilst not constraining the model to assume that the distribution of infection risk would be identical for all three host organisms. As BRTs can fit high-dimensional interactions, the joint model is able to fit different environmental responses for each host organism, or if there is no difference in the signal, to fit the same response for all of them. To make predictions, an additional covariate data surface was generated for each host species, where the value of every pixel equalled one, two or three, to represent mosquito, macaque and human hosts, respectively. Then three separate predictions were made, each using the complete set of the most contemporary environmental and socioeconomic data surfaces, as well as the relevant host data surface.

The model did not detect appreciable differences in the environmental predictors of infection risk in humans, macaques or mosquitoes. The model of human infection risk was therefore used for final model outputs.

2.2.5 Model prediction and evaluation

Reservoir and vector maps

Species distribution maps were generated displaying the mean predicted value of environmental suitability for each 5×5 km pixel across the 120 sub-models, along with the 0.025 and 0.975 quantile values from these distributions for each pixel. Each species' model predictive performance was evaluated using the area under the receiver operator curve statistic (AUC), *i.e.* the area under a plot of the true positive rate versus false positive rate, reflecting the ability to discriminate between presence and background records [108]. The AUC ranges from 0 to 1, where a score of 1 indicates perfect discrimination between occurrence and background records, a score of 0.5 suggests that predictive capacity is no better than random, and less than 0.5 is worse than random. The AUC score has been extensively used in the species' and disease distribution modelling literature, and can be interpreted as the probability that any randomly sampled occurrence record will have a higher predicted value than a randomly sampled background record. For each sub-model, the mean AUC under 10-fold cross validation was calculated using a pairwise distance sampling procedure to remove potential spatial sorting bias in the cross validation datasets. This AUC was then averaged across all 120 sub-models in the ensemble to give an overall estimate of predictive performance, as well as uncertainty in this estimate.

Disease map

To generate the final disease prediction map, a mean predicted value of suitability for infection was calculated across the 500 sub-models (each fitted using occurrence and covariate data from within MBS) for each 5×5 km pixel within and outside MBS. To assess model uncertainty, the standard deviation of the mean was also calculated for each pixel.

The model's predictive performance was evaluated by calculating the mean of the AUCs for each of the 500 sub-models, calculated under 10-fold cross validation. While each sub-model in the BRT ensemble was fitted using occurrence and background data from MBS (the model training region), the goal of the analysis was to predict to a much broader study area from northeast Bangladesh to southwest Papua New Guinea. To assess the model's predictive performance outside its training region, a separate AUC

value was calculated for each sub-model using a validation dataset made up of presence and absence records from locations outside MBS. This AUC was calculated for each sub-model and then averaged across all 500 sub-models in the ensemble.

2.2.6 Disease model multivariate environmental similarity surface

The geographic regions outside MBS encompasses environments beyond the ranges of covariate values sampled within the training dataset. Model predictions to such environments are inherently less reliable than interpolations made to areas with environments within the range of covariate values in the training dataset. Thus it is important to assess the environmental similarities and differences between model training and prediction regions [109].

To investigate whether predictions to new geographic regions required extrapolation to covariate values beyond the range of the model training data, a multivariate environmental similarity surface (MESS) [110] was computed using R packages 'dismo' [111] and 'raster' [112]. This surface represents the similarity of the environment at each location to the covariate values at the presence and background locations in MBS (the reference data). The MESS calculation produces negative values for novel environments, locations where at least one covariate has a value that is outside the range of reference values (hereafter extrapolation), and positive values for locations within this range (hereafter interpolation). The raw MESS output was converted into a binary map indicating areas in which model predictions used interpolation versus extrapolation.

2.2.7 Disease map masking

The disease model output was restricted to areas within the range of species known and hypothesised to be required for zoonotic transmission (*i.e.* the overlap in range maps of at least one reservoir and vector species), using the species distribution maps produced in this chapter.

A high-resolution map for the zone of zoonotic transmission was also generated (Appendix A.11) using the reservoir and vector distribution maps and occurrence datasets. Threshold environmental suitability values for each of the species distribution maps for *M. fascicularis*, *M. nemestrina*, *M. leonina*, and the Leucosphyrus Group were used

to classify each continuous species map as either present or absent for that species at every 5×5 km pixel in the study region. Threshold values were chosen that encompassed 90% of presence points in each species respective occurrence database, then every 5×5 km pixel in the suitability map above that threshold value was classified as present for that species. These binary presence/absence maps were combined to produce a final binary output showing areas of spatial co-occurrence of all species required for zoonotic, vector-borne transmission to humans *i.e.* presence of at least one macaque species plus at least one member of the *Anopheles leucosphyrus* group.

2.3 Results

2.3.1 Maps of *P. knowlesi* reservoir and vector species

The predictive species distribution maps generated from the mean model outputs for each vector and reservoir species, masked by published range maps for each species, are displayed in Figures 2.3 and 2.4, respectively. These predictions were produced using the most contemporary covariate values available (*i.e.* values for land cover classes were from 2012). The AUC values for the predicted macaque distributions indicated high predictive performance, with 0.858 (SE ± 0.001) for *M. fascicularis*, 0.821 (SE ± 0.003) for *M. nemestrina*, and 0.830 (SE ± 0.002) for *M. leonina*. The AUC values for the predicted mosquito distributions were also high, with 0.860 (SE ± 0.005) for *An. dirus*, 0.885 (SE ± 0.002) for the Dirus Complex, 0.842 (SE ± 0.009) for the Leucosphyrus Complex, and 0.883 (SE ± 0.001) for the Leucosphyrus Group.

The 0.025 and 0.975 quantiles from the model ensemble, and the top predictor variables for each species are provided in Appendix A.

2.3.2 Map of *P. knowlesi* malaria risk

The model prediction of geographical variation in relative risk of *P. knowlesi* infection in humans is displayed in Figure 2.5A. Ten-fold cross validation statistics for the model ensemble (calculated using model training data) indicated high predictive performance with an AUC of 0.833 (SE ± 0.002). A map of model uncertainty (standard deviation of the mean value for each pixel) is provided in Appendix A.12.

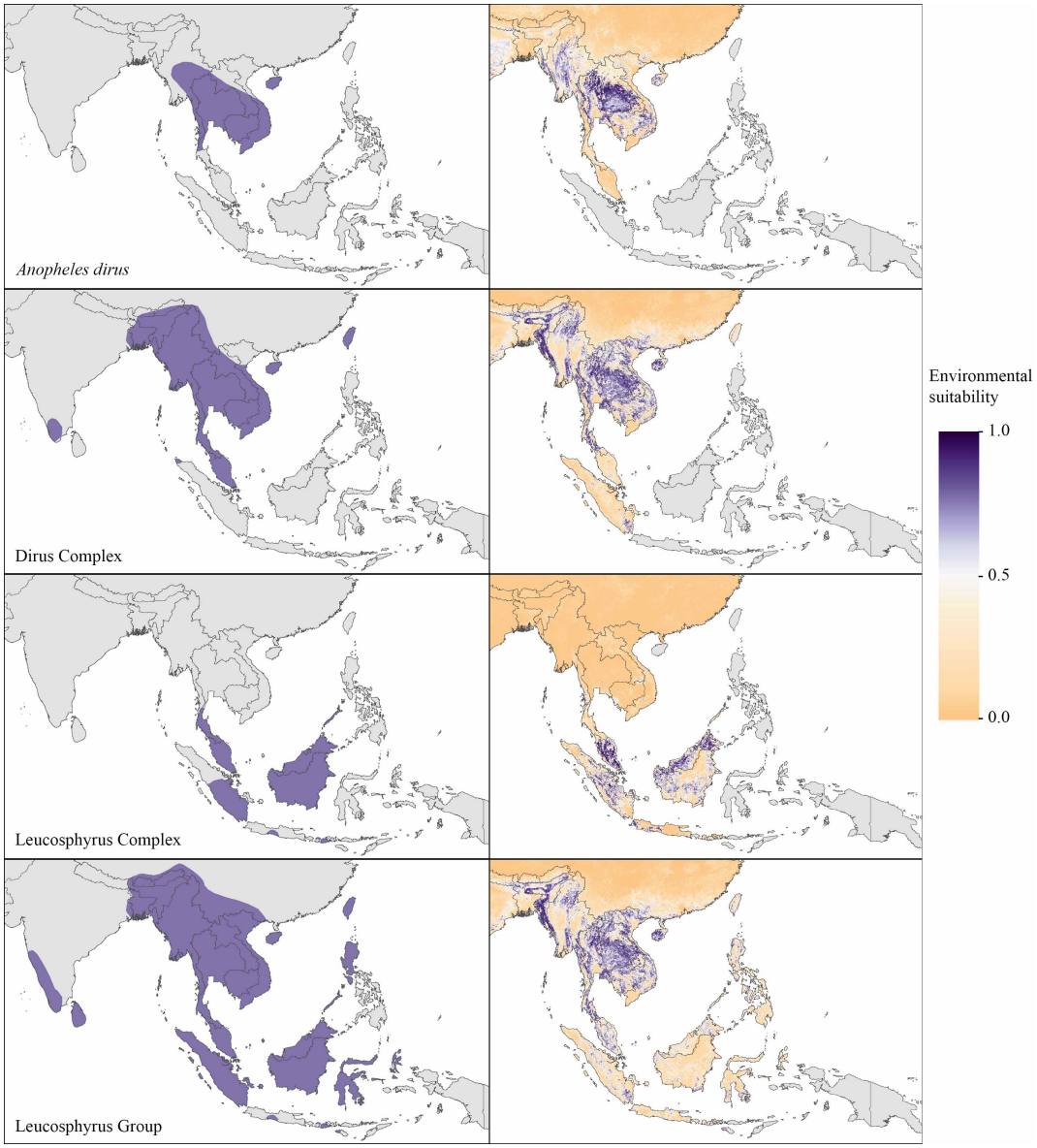


Figure 2.3: *Plasmodium knowlesi* malaria mosquito vector distribution maps. Predictive species distribution maps generated from boosted regression tree model outputs for vectors of *P. knowlesi* malaria; *Anopheles dirus*, Dirus Complex, Leucosphyrus Complex and the Leucosphyrus Group (right). Each map is masked by a previously published range map for that species, complex or group (left). The range maps were based on International Union for Conservation of Nature (IUCN) ranges for these species [113].

The model output was restricted to areas within the geographic range of the species required for zoonotic transmission. An unmasked version of the mean output, showing risk of zoonotic *P. knowlesi* malaria for a hypothetical scenario where known reservoir and vectors species have expanded their ranges to reach all suitable environments, is provided in Appendix A.15.

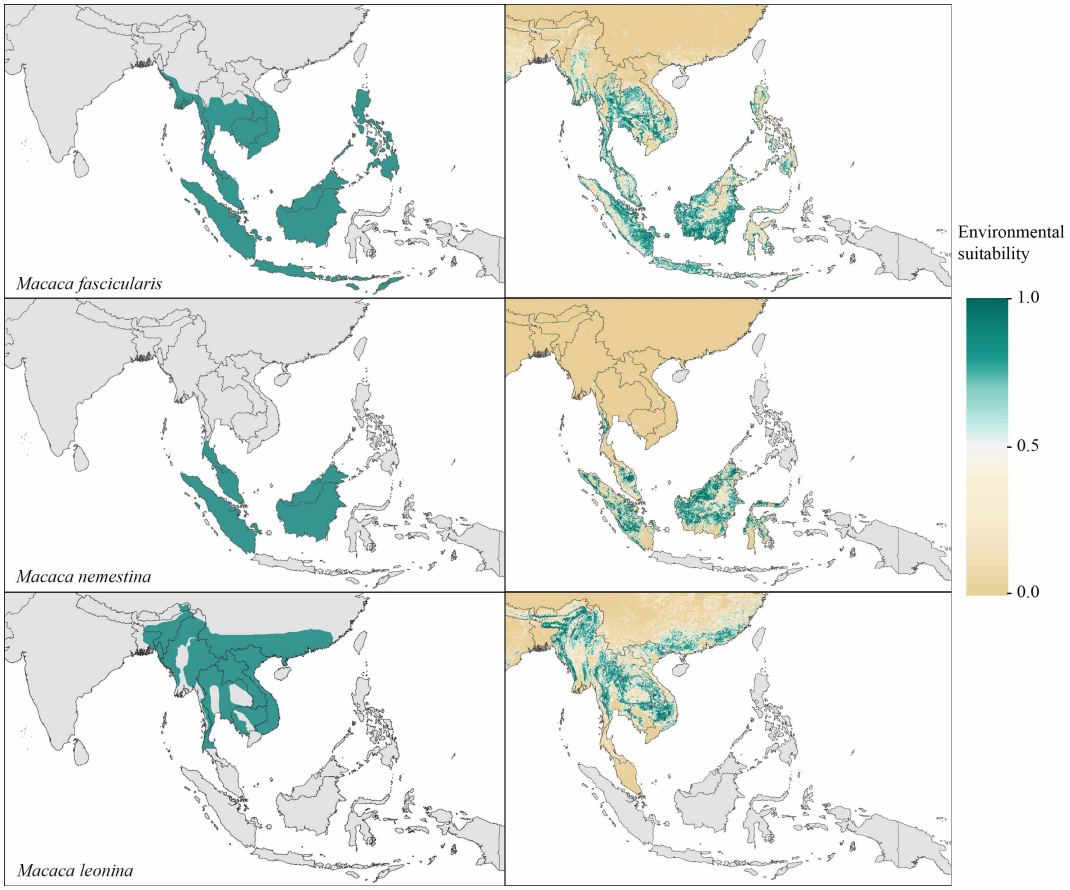


Figure 2.4: *Plasmodium knowlesi* macaque reservoir distributions maps. Predictive species distribution maps generated from boosted regression tree model outputs for *P. knowlesi* host macaque species: *Macaca fascicularis*, *M. leonina* and *M. nemestrina* (right). Each map is masked by a previously published range map for that species (left). The range maps were based on International Union for Conservation of Nature (IUCN) ranges for these species [113].

The predicted map is presented alongside a projection of malaria eliminating countries in the year 2025 (Figure 2.5B) and together the two maps show countries where *P. knowlesi* transmission may persist after the human malarias are eliminated. The elimination projections, generated by the University of California San Francisco Global Health Group's Malaria Elimination Initiative, are based on current national and regional goals as well as recent epidemiological trends for the human malarias, principally *P. falciparum* and *P. vivax* [114].

Within MBS, the model predicted considerable spatial variation in risk of *P. knowlesi* infection, with areas ranging from relatively low risk to high risk predicted within Peninsular Malaysia, and both Sabah and Sarawak States of Malaysian Borneo (Figure 2.5A).

The model also predicts areas of high risk for *P. knowlesi* infection in a number of

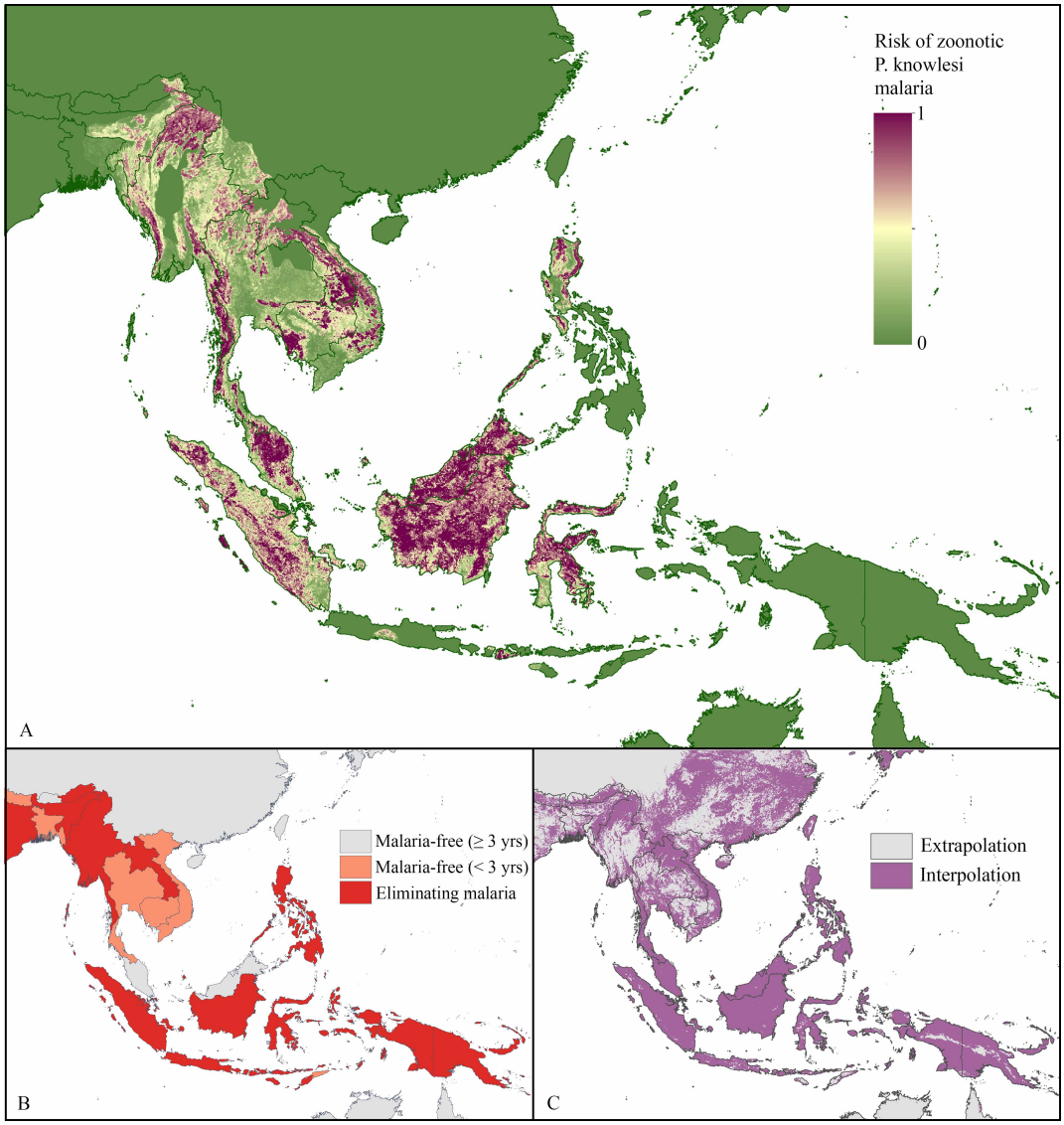


Figure 2.5: Maps of relative *Plasmodium knowlesi* malaria risk, human malaria elimination status, and model extrapolation versus interpolation. A. Predicted relative risk of *P. knowlesi* malaria ranging from low to high risk. **B.** Countries projected to be malaria-free, eliminating malaria, or controlling malaria by 2025 (Map sourced from the University of California San Francisco Global Health Group's Malaria Elimination Initiative) **C.** Comparison of environments in Malaysia, Brunei and Singapore (the model training region) with those across the rest of southeast Asia, using all covariates and the multivariate environmental similarity surface methods. The map distinguishes between areas of model interpolation and areas where the model was required to extrapolate to novel environments.

countries that are forecast to be malaria-free by 2025 (Malaysia, Cambodia, Thailand and Vietnam) as well as countries projected to be eliminating malaria (Myanmar, Laos, Indonesia and the Philippines) (Figure 2.5B). Large areas of high risk were predicted in Myanmar, Laos, Cambodia and Indonesia, with smaller areas predicted in Vietnam

and Thailand. Human cases of *P. knowlesi* infection have been reported across this broad area (Figure 2.5A).

Regions for which there is no field data include areas of high predicted-risk, for example, eastern and western parts of Indonesia, and far eastern parts of India, although the predictions for the latter depend on whether *M. leonina*, included in the range of zoonotic transmission, is indeed a reservoir species (Figure 2.5A).

The model predictions outside MBS are a result of both interpolation within the environmental space and extrapolation. The binary MESS map (Figure 2.5C) shows that model extrapolation to novel environments occurred in large regions in Cambodia, Vietnam, Thailand, Myanmar, India, and the Andaman and Nicobar Islands, indicating that predictions in these areas should be interpreted with caution. The model did, however, demonstrate high predictive performance at sites outside the model-training region, with an AUC of 0.796 (SE ± 0.003), calculated using 131 presence/absence locations from the evaluation dataset (Figure 2.5A).

The main predictors for *P. knowlesi* infection risk were urban accessibility, human population density, elevation, proportional cover of land with croplands and environmental suitability for *M. nemestrina*. Marginal effect plots for each of these covariates are displayed in Appendix A.14.

2.4 Discussion

Characterising the distribution of all component species is an important first step in understanding the distribution of a vector-borne, zoonotic disease, especially where human infection data are scarce. In this chapter, species distribution maps were produced for known and putative reservoirs and vectors of *P. knowlesi* malaria. These maps were then used as explanatory variables, along with a range of environmental and socioeconomic factors, and a spatial database of *P. knowlesi* infections, to model the spatial variation in relative risk of human *P. knowlesi* infections. Empirical data on *P. knowlesi* presence or absence are lacking for most of southeast Asia and this map provides an initial evidence base to prioritise areas for disease surveillance and future epidemiological investigations.

The predictive performance of the disease model fitted within MBS was high and it also had a high capacity to predict suitability for infection in regions outside MBS. The latter result should, however, be treated with caution as data for model evaluation were only available from a limited number of locations outside MBS, and the selection of locations for which *P. knowlesi* has been tested is likely to be biased.

Another important caveat is the large area to which model predictions were made, relative to the model training region, since this required the model to extrapolate to some novel environments (see Figure 2.5C). Extrapolated predictions are inherently less reliable than those made in areas of interpolation and include large parts of continental Asia. Sampling for *P. knowlesi* infections in areas of extrapolation is likely to have the biggest impact on improving the disease risk predictions.

The final map therefore shows the risk of zoonotic *P. knowlesi* transmission from known reservoirs (specifically *M. nemestrina* and *M. fascicularis*) and vectors of the *Anopheles leucosphyrus* group. If human-to-human transmission were occurring, this form of the disease is likely to have a different niche to the zoonotic disease, *i.e.* a different relationship with environmental, socioeconomic and biological factors. Thus the infection risk model is not appropriate to predict human-to-human transmission risk.

It is also important to note that the limits of zoonotic transmission, within which the disease map predicts infection risk, were defined using the reservoir and vector species distribution maps and these distributions reflect the fact that species ranges are not fixed. Specifically, these distributions included introduced populations of the two macaque species, for example, pet *M. fascicularis* and *M. nemestrina* macaques are commonly found on Sulawesi where the environment is suitable for the establishment of feral populations. The predictions for infection risk therefore include locations on this island.

Human *P. knowlesi* infections have been identified beyond the ranges of both *M. nemestrina* and *M. fascicularis*. *Macaca leonina*, whose range extends farther north into Myanmar where these human cases were reported [76], has thus been implicated as a putative host species. Predictions were allowed within the range of *M. leonina* but since this species is not found in MBS, it was not used as an explanatory covariate for model fitting. This may have impacted model predictions in the most northern parts of

the study area where the environmental suitability for *M. nemestrina* and *M. fascicularis* is low, but high for *M. leonina*. Again, sampling in these areas, particularly northern Myanmar, would improve the predictions.

Furthermore, two distinct *P. knowlesi* parasite populations, linked to *M. nemestrina* and *M. fascicularis* respectively, have been identified in human patients from Malaysia [115]. It is reasonable to assume that only the strain associated with *M. fascicularis* is circulating and infecting humans in areas of continental Asia, where *M. nemestrina* is absent, and it may have a distinct relationship with environmental and socioeconomic variables compared to the mixture of parasite infections in patients from Malaysia. The presence of Leucosphyrus Complex vectors in Malaysia and Dirus Complex vectors in continental Asia further adds to the possibility of different relationships between disease risk and the environment in these two regions.

Comparing the predicted *P. knowlesi* risk map (Figure 2.5A) with the map of current sampling efforts (Figure 2.2), and the map of malaria eliminating countries (Figure 2.5B), allows one to identify relative surveillance priorities for *P. knowlesi*. These include a number of regions in Thailand (Phisanulok, Phetchuban, Chaiyaphum, Prachan Buri, and southern Nakhon Ratchasima) and Vietnam (Lam Dong, Phu Yen, Gia Lai, and Kon Tum). Further surveillance in previously sampled areas of Thailand, Vietnam and Cambodia is also required to fully understand the distribution of *P. knowlesi* in countries close to eliminating the human malarias.

Among the countries next expected to eliminate the human malarias, my results highlight a need for surveillance in un-sampled, high-risk areas in Myanmar, Laos, and Sumatra and Kalimantan in Indonesia. Initial studies have reported cases in Aceh on Sumatra [89], and South and Central Kalimantan [116, 117] but no published reports are available from the other parts of these regions. Further surveillance is also needed in previously sampled areas, including Palawan in the Philippines.

Importantly, the disease map predicts the environmental suitability for infection, not the prevalence of infection or incidence of cases in these places. The higher numbers of reported cases in Malaysia is not proof that the disease risk is higher here because most locations outside Malaysia simply have not been surveyed and *P. knowlesi* cases could be

misdiagnosed as one of the human malarias. Studies that have investigated numbers of cases or infections have sampled a wide array of communities including malaria patients [72], patients diagnosed as *P. malariae* by microscopy [118], and whole communities [119], meaning the disease prevalence indicators generated are not directly comparable. Until more locations are surveyed using a consistent measurement (ideally infections in a cross section of the community) and diagnostics that distinguish all human malarias, no firm conclusions about relative disease prevalence can be drawn [107]. It will be important to update the predictions presented here when new data become available, and systems are available to generate updated predictive maps [24]. Importantly the map presented here provides key information about the locations where new surveys for *P. knowlesi* infections would be most informative.

2.5 Conclusions

As the volume and quality of geographical data on *P. knowlesi* infections increases across southeast Asia, the maps presented in this chapter will iteratively improve. For now, they provide the best evidence-base currently available for prioritising *P. knowlesi* surveillance to better understand its spatial distribution and its wider contribution to malaria cases.

Using a standard niche modelling approach, with limited occurrence data, one can estimate the spatial variation in suitability for disease transmission, which in this case is a useful metric for prioritising locations for future *P. knowlesi* surveys. However, suitability for infection is not necessarily proportional to a more useful epidemiological metric, such as case incidence or infection prevalence, and can be difficult to interpret if such metrics are desired. In *Chapter 4*, this issue is explored further and I describe the development of a novel niche modelling approach that is able to produce more epidemiologically interpretable outputs.

3

Estimating global yellow fever vaccination coverage from 1970 to 2016

Contents

3.1 Background to yellow fever 38
3.2 Introduction 40
3.3 Methods 41
3.4 Results 54
3.5 Discussion 56
3.6 Conclusions 62

This chapter describes a retrospective analysis tracking vaccination coverage for YF in all age cohorts from 1970 to 2016. In addition to estimating the additional vaccination coverage needed to prevent further YF outbreaks, one of the aims of this chapter is to produce coverage data surfaces to inform the YF risk model developed in *Chapter 4*. The work in this chapter is published in *The Lancet Infectious Diseases* [120]. The content of *Chapters 3* and *4* are preceded by general background information on YF.

3.1 Background to yellow fever

Substantial outbreaks of YF in Angola, Democratic Republic of the Congo (DRC), and Brazil in the past two years [121], coupled with the global distribution of its main urban vector,

Aedes aegypti, suggest that YF has the propensity to spread further internationally and increase its financial burden on health systems as well as its toll on population health.

Yellow fever is an acute viral haemorrhagic disease that is vaccine preventable, yet incomplete vaccination coverage means that the disease is still widely distributed in the tropics and sub-tropics of Latin America and Africa [122], where it caused an estimated 5,800 deaths and 374,015 DALYs in 2016 [123, 124]. Transmission of YF has never been reported in Asia, despite multiple opportunities for introduction and seemingly suitable ecological and climatic conditions [125].

Yellow fever virus belongs to the family *Flaviviridae*, an extended phylogenetic family also containing dengue, Zika, and West Nile viruses [126]. Multiple transmission cycles of YFV co-exist with different mosquito genera serving as vectors in each cycle. The virus is principally maintained by a sylvatic transmission cycle, where non-human primates are hosts and forest-dwelling mosquitoes of *Aedes spp* (Africa) or *Haemogogus spp* (Americas) act as vectors. Humans can sporadically be incidental hosts, usually through occupational or recreational exposure to infected mosquitoes. In the urban cycle, humans are primary hosts and day-feeding *Ae. aegypti* mosquitoes transmit the virus, which can lead to large-scale epidemics. In Africa, there is also an intermediate transmission cycle that occurs in rural areas, typically at the edges of forests, where both humans and non-human primates are hosts and *Aedes spp* act as vectors [5].

The spectrum of human clinical disease caused by YFV is broad, ranging from asymptomatic (inapparent) infections and mild febrile illness, to severe disease and death [127]. Since no specific anti-viral drug exists for YF, there is a heavy emphasis on disease control to prevent infections [128]. Due to diagnostic complexity and limitations of health care systems in many affected countries, YF cases are substantially underreported, with one study estimating case detection probabilities to be of the order of 10^{-5} in parts of Africa [129]. The disease history, epidemiology and public health impact of YF is very different in the two risk zones, Latin America and Africa.

The epidemiology of YF on the African continent, where the vast majority of cases and epidemics are reported, involves a mix of urban, sylvatic and intermediate transmission cycles. In recent decades, almost all cases of YF in Latin America arose from sylvatic

cycles, with inter-human transmission only reported a few times since the early 1940s [130]. While eradication of YF is not considered feasible due to its sylvatic reservoir, outbreak control is achievable due to the availability of a safe, low-cost and highly effective vaccine. If control is insufficient, YFV can cause devastating epidemics, especially in naïve populations or where population immunity is low, with case-fatality rates reported as high as 58% [131, 132].

3.2 Introduction

The YF vaccine is a live, attenuated vaccine that rapidly stimulates immunity (within 30 days for 99% of people vaccinated) and can provide lifelong protection [128]. Since the vaccine became available in 1937, multiple vaccination strategies have been implemented, including the introduction of the vaccine into routine childhood schedules, mass preventive and reactive campaigns across all age groups, as well as the vaccination of travellers to YF risk zones. Combined with vector control programs, this has led to a notable reduction in disease burden at targeted locations and times [129, 131]. Yet in many YF risk areas, vaccine coverage remains too low to prevent outbreaks. The WHO recommends population vaccination coverage of 80% or more to prevent and control outbreaks [122].

The recent outbreak that started in Angola in December 2015, developed into the largest and most widespread outbreak of YF reported in Africa for more than 20 years with 884 and 78 confirmed cases in Angola and neighbouring DRC, respectively [133, 134]. Before 2015, YF had not troubled Angola since 1988 when a much smaller outbreak resulted in 37 cases [135]. Angola introduced the vaccine into routine infant immunisation programs in 1999 [136], but there had been no campaigns to protect adults born before 1999 since an outbreak response campaign in Luanda in 1988 [135].

The Angolan outbreak also exposed vulnerabilities in the preparedness of local and international agencies to respond to such emergencies. Efforts to control the outbreak through mass reactive vaccination campaigns were hampered by dwindling international vaccine supplies and other operational challenges [137]. With the global YF vaccine emergency stockpile depleted by mid-2016, the WHO approved the temporary use of fractional doses to stretch remaining stocks [138].

The course of the outbreak in Angola and DRC reiterates the need for a sustained policy of preventive vaccination for at-risk populations to reduce the risk of epidemics. Planning for vaccination campaigns and other control strategies requires current vaccination coverage rates that have resulted from cumulative campaigns, but this information does not exist for all countries and years and there are a number of barriers to generating a global estimate of contemporary coverage. Spatial estimates of vaccination coverage for YF have been produced previously [129, 139] but are restricted to specific age cohorts or countries, with one study estimating population-wide coverage across all at-risk countries on the African continent, but not however, for Latin America [129]. The aims of this chapter are to generate global YF vaccination coverage maps from 1970 to 2016, and to estimate the additional coverage needed to prevent further outbreaks.

3.3 Methods

3.3.1 Data collation

Yellow fever vaccination is primarily delivered through three programme types: routine childhood immunisation targeting infants around nine months of age in at-risk regions; periodically conducted mass preventive and outbreak response campaigns that target a broader age range; and vaccination of people travelling to high-risk areas. To track vaccination coverage of the entire population over time, I compiled a dataset of the coverage level achieved by each of the first two classes of vaccination activity for each age group at specific locations and time periods from the earliest use of the vaccine in 1939 to 2016. Table 3.1 displays a summary of datasets obtained *via* the WHO, United Nations World Population Prospects (UNWPP) and Centers for Disease Control and Prevention (CDC) websites.

3.3.2 Routine infant vaccination data

Data on routine infant vaccination coverage were obtained from annual national estimates, based on health-service-provider registries of YF coverage reported yearly by each at-risk country to the WHO and United Nations International Children's Emergency Fund (UNICEF) using the joint reporting form (JRF) on immunisation [147, 148]. The data were

Table 3.1: Summary of online vaccination and demographic data sources. Vaccination activity and demographic data accessed via World Health Organization (WHO), United Nations World Population Prospects (UNWPP) and Centers for Disease Control and Prevention (CDC) websites. Mass vaccination = preventive and outbreak response vaccination activities. Routine immunisation = routine infant immunisation. JRF = Joint Reporting Form.

Data description	Timeframe	Date accessed	Data type
WHO Weekly Epidemiological Records [140]	1970-2016	26/08/2016	Mass vaccination
WHO Disease Outbreak News [141]	1996-2016	26/08/2016	Mass vaccination
CDC Morbidity and Mortality Weekly Reports [142]	1982-2016	03/10/2016	Mass vaccination
WHO-UNICEF reviews of national immunisation coverage [143]	1980-2014	26/08/2016	Routine vaccination
WHO-UNICEF JRF national immunisation coverage [144]	1980-2015	03/10/2016	Routine vaccination
UNWPP Annual Population by Five Year Age Groups [145]	1950-2015	13/10/2016	Demographic
UNWPP Infant Mortality Rate [146]	1950-2016	10/10/2016	Demographic
UNWPP Under-five Mortality [146]	1950-2016	10/10/2016	Demographic

available from the year of YF vaccine introduction into the country's routine immunisation schedule up to 2015. Province or state level data, as estimated by the WHO/UNICEF based on data reported in the JRF [144], were also available for certain countries with localised areas of YF risk, namely, Panama, Argentina, and Kenya [143]. The WHO/UNICEF estimates for Brazil indicated that vaccines were only administered to infants in risk zones and it was assumed that these estimates only applied to municipalities within areas classified as 'endemic' by WHO range maps [27]. Since other types of vaccination data reporting up to 2016 were used, to obtain estimates of all routine coverage for 2016, it was assumed that the percent coverage was the same as for 2015.

3.3.3 Preventive and outbreak response vaccination campaign data

Data on mass preventive and outbreak response campaigns were primarily extracted from published online sources including the Weekly Epidemiological Records (WERs) [140], Disease Outbreak News reports (DONs) [141], Morbidity and Mortality Weekly Reports (MMWRs) [142], and two WHO reports: one on mass vaccination in West and

Central Africa through the 1940s to 1960s [149], and another on the 2016 epidemic in the DRC [150]. For the WERs, DONs, and MMWRs, all reports within the specified date range (see Table 3.1) were searched for the keywords 'yellow fever', of which only those containing information on YF vaccination activities were downloaded. The detail in these reports varied greatly from campaign to campaign and was reported at a variety of spatial scales. If vaccination coverage for a campaign was not given, it was estimated using other information, such as the number of vaccine doses administered, combined with population estimates (see Section 3.3.5). The following information was recorded for each vaccination campaign, where available: campaign time period, target location(s), target age range, number of vaccine doses (either administered or supplied), target population size, proportion of target population vaccinated (coverage), and whether coverage was estimated from a post-campaign survey or administrative data (*i.e.* health-service-provider reported estimates).

Where available, coverage values were derived from post-campaign coverage surveys that asked participants whether they could recall having ever received a YF vaccine (and/or their children) and whether they could present an official YF vaccination record. Post-campaign coverage surveys were used in preference of administrative data because these generally provide more reliable coverage estimates [151–153]. Finally, duplicate data were removed for campaign records covering the same location and time period, retaining the most precise records. A description of the data issues encountered and actions taken is included in Table 3.2 and a summary of the information extracted from the WERs, DONs, MMWRs and WHO reports can be found in Table 3.3.

Table 3.2: Processing of preventive and outbreak response vaccination data. Details of extraction issues with preventive and outbreak response vaccination campaign data and actions taken.

Data issue	Action taken
Campaign described as 'planned' or 'in progress' at the time of the report but the number of vaccine doses to be administered was specified	Assumed that one dose was administered per person in targeted locations
Campaign described as 'planned' and only the target population size was specified	Campaign was excluded

Continued on next page

Table 3.2 – *Continued from previous page*

Data issue	Action taken
Target age group was described as 'children'	Assumed that this referred to children less than five years of age
Target age groups not specified	Assumed that individuals in all age groups had an equal chance of receiving a vaccine dose
Coverage was reported to be greater than 100%	Coverage resolved to 100%
Coverage range was stated (*e.g.* 90-95%)	Recorded lower limit (*e.g.* 90%)
The report quoted the number of people vaccinated rather than number of doses or coverage	Recorded as the number of vaccine doses
Coverage was not directly reported (*i.e.* only the number of vaccine doses administered, number of people vaccinated or target population size *etc.* was provided)	Estimated coverage by calculating the size of target population (denominator) using WorldPop and UNWPP data
Coverage reported without specifying the size of the target population	Assumed that target population size equalled population of target locations
Target population size reported without specifying the coverage value	Coverage was assumed to be 100%
The number of vaccine doses, the number of people vaccinated or the size of the target population was reported to be greater than the total estimated number of people in the specified target locations	Target population size was replaced with estimated population size of target locations, meaning that coverage was 100%
The number of vaccine doses was provided without reporting the achieved coverage or the target population size	Assumed that number of people vaccinated equalled number of vaccine doses
Target population size not reported	Assumed target population size equalled population of target locations
Coverage or number of vaccine doses was aggregated over a number of target locations	Assumed that individuals in each location had an equal chance of receiving a vaccine dose
Unclear whether the campaign was targeting locations at the third or fourth order administrative unit (*i.e.* locations have the same name and it was not clear which was targeted in the report)	Assumed campaign targeted the third order administrative unit

Continued on next page

Table 3.2 – *Continued from previous page*

Data issue	Action taken
Target locations provided at the third or fourth order administrative level	Coverage at the second administrative level was calculated
Target locations not specified	Divided number of people vaccinated by estimated population size within broad target locations or within the entire country
Target locations listed as 'endemic' or 'at risk' regions	Divided number of people vaccinated by estimated population size within districts classified as endemic or at-risk by the WHO [27]
Number of vaccine doses aggregated over multiple years	Assumed that an equal number of doses were administered each year
Different coverage values reported from a post-campaign survey based on 1) combination of respondent recall and the presence of official vaccination records and 2) vaccination records only	Used coverage value based on combination of respondent recall and presence of official vaccination records

Table 3.3: Summary of preventive and outbreak response vaccination data. Preventive and outbreak response campaigns reported in each Weekly Epidemiological Report (WER) [140], Disease Outbreak News report (DON) [141], Morbidity and Mortality Weekly Report (MMWR) [142], and two World Health Organization reports: one on mass immunisation in West and Central Africa in the 1940s and 1950s [150] and another on the 2016 YF epidemic in the Democratic Republic of the Congo (DRC) [149]. The DONs are identified by the date of issue in the format ddmmyy and the WERs by four digits representing the volume and issue numbers. CAR = Central African Republic.

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
DON011102	Senegal	Mbacké, Bambey (2)	2002
DON031105	Mali	Kadiolo, Kolondieba (2)	2005
DON031213	Sudan	Western Darfur, Southern Darfur, Nothern Darfur (1)	2012
DON070601	Peru	Puinahua, San Pablo , Iquitos (3)	2001
DON080110	Côte d'Ivoire	Kaniso, Sandégéu, Minignan, Samantiguila (3)	2009
DON080808	Côte d'Ivoire	Abidjan (2)	2001
DON081013	Cameroon	Dibombari, Edéa, Loum, Manjo, Manoka, Mbanga, Melong, Ndom, Ngambé, Nkondjock, Nkongsamba, Pouma, Yabass (3)	2013

Continued on next page

Table 3.3 – *Continued from previous page*

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
DON110304	Liberia	Margibi, Grand Bassa, Grand Gedeh (1)	2004
DON281105	Mali	Kangaba, Kati, Kayes, Kéniébe, Kita, Bafoulabé (2)	2005
DON300903	Sierra Leone	Tonkolili (2)	2003
DON260816	Angola	Luanda, Benguela, Cuanza Sul, Huambo, Huíla, Uíge (1)	2016
Durieux, 1956	Mauritania	NA	1939-53
	Senegal	NA	1939-53
	Mali	NA	1939-53
	Guinea	NA	1939-53
	Côte d'Ivoire	NA	1939-53
	Burkina Faso	NA	1939-53
	Benin	NA	1939-53
	Niger	NA	1939-53
	Togo	NA	1940-53
	Chad	NA	1944-53
	CAR	NA	1944-53
	Republic of Congo	NA	1944-53
	Gabon	NA	1944-53
WER4543	Peru	Oxapampa, Tarma, Satipo (2)	1969-70
WER4631	Ghana	Upper East, Upper West (1)	1969
	Ghana	Brong-Ahafo Region (1)	1970
	Nigeria	Plateau, Bauchi (1)	1969
WER4724	DRC	Gemena, Dungu (3)	1958
	DRC	Gemena (3)	1972
WER5034	Nigeria	Abia, Anambra, Ebonyi, Enugu, Imo, Delta, Rivers (1)	1974
WER5139	Bolivia	Chuquisaca, La Paz, Santa Cruz, Cochabamba, Tarija (1)	1975
	Ecuador	Napo Province (1)	1975
WER5224	Colombia	Casanare (1)	1976
WER5423	Gambia	NA	1978
WER5440	Colombia	Norte de Santander (1)	1979
WER5440	Venezuela	Zulia, Trujillo, Mérida (1)	1941-78
WER5511	Ghana	Volta, Eastern, Brong-Ahafo, Upper East, Upper West (1)	1979

Continued on next page

Table 3.3 – *Continued from previous page*

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
WER5545	Senegal	Fatick, Kaolack, Kaffrine, Tambacounda, Kédougou, Kolda, Sédhiou, Ziguinchor (1)	1979
WER5545	Trinidad and Tobago	NA	1979
WER5633	Cameroon	Diamaré (2)	1980
WER5739	Bolivia	NA	1981-82
	Brazil	Maranhão, Acre, Amazonas, Pará, Mato Grosso, Mato Grosso del Sul, Goiás, Rondonia, Federal District, Roraima, Amapá (1)	1937-83
	Senegal	Thiès (1)	1981
WER5841	Ghana	NA	1980-82
WER5943	Burkina Faso	Centre (1)	1983
	Ghana	Northern, Upper East (1)	1983
WER6056	Brazil	Presidente Prudente (1)	1985
WER6406	Mali	Kita, Kati, Kangaba, Kolokani, Bafoulabé, Diéma, Kéniéba, Nioro du Sahel, Yélimané, Banamba, Dioïla, Koulikoro, Nara, Mopti, Ségou, Sikasso (2)	1987
WER6423	Bolivia	Santa Cruz, Montero, Carrasco, Ichilo (3)	1985-89
WER6528	Angola	Luanda (1)	1988
WER6733	Nigeria	Abuja, Akwa Ibam, Anamba, Bauchi, Delta, Edo, Benue, Borno, Cross River, Adamawa, Taraba, Imo, Kaduna, Kano, Katsina, Kwara, Lagos, Niger, Oyo, Ogun, Ondo, Plateau, Rivers, Sokoto (1)	1986-89
WER7010	Kenya	Baringo, Elgeyo Marakwet (2)	1993
WER7015	Nigeria	Orsu (2)	1994
WER7050	Liberia	Grand Bassa (1)	1995
WER7051/52	Sierra Leone	Kenema (2)	1995
WER7124	Gambia	NA	1978-79
WER7142	Liberia	Buchanan, Grand Bassa (1)	1995
WER7145	Benin	Atakora, Borgou (1)	1996
WER7211	Bolivia	Cochabamba, La Paz, Beni, Santa Cruz (1)	1997
WER7335	Brazil	Afuá (3)	1998
WER7345	Venezuela	Amazonas (1)	1998
WER7346	Benin	Atakora, Borgou (1)	1998

Continued on next page

Table 3.3 – *Continued from previous page*

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
	Ghana	Upper East Region (1)	1996-97
	Liberia	Voinjama, Kolahun, Foya (2)	1997
WER7401	Burkina Faso	Batié (3)	1998
WER7540	Bolivia	Cordillera (2)	1999
	Brazil	NA	1990-2001
	Liberia	Nimba (1)	1998
WER7743	Senegal	NA	2002
WER8029	Burkina Faso	Bobo-Dioulasso, Nouna, Solenzo, Sindou, Gaoua (3)	2004
	Colombia	Cesar, Magdalena, La Guajira (1)	2003-04
	Liberia	Bong, Nimba (1)	2004
	Mali	Kita (2)	2004
	Venezuela	Unspecified high risk areas	2006
	Venezuela	Unspecified enzootic areas	2002-04
WER8051/52	Côte d'Ivoire	Bondoukou, Bouna (3)	2005
WER8133	Burkina Faso	Batié, Gaoua, Banfora (3)	2005
	Senegal	Bakel (2)	2005
	Sudan	South Kordofan (1)	2005
WER8307	Brazil	Federal District, Goiás, Mato Grosso do Sul (1)	2008
WER8308	Cameroon	NA	2006
	CAR	Yaloke-Bossembele (2)	2006
	Côte d'Ivoire	Ouragahio, Korhogo (4)	2006
	Ghana	Sene (2)	2006
	Mali	Yanfolila (2)	2005-06
WER8340	Guinea	NA	2005
WER8350	Mali	NA	2005
	Senegal	NA	2002-06
WER8413	Bolivia	NA	2007
	Cameroon	Akonolinga, Zoétélé (3)	2007
	Senegal	Dagana (and Richard Toll), Podor (and Pete), Linguère, Kédougou, Salémata, Saraya, Thiés (and Pout), Kaolack (and Ndofane), Foundiougne (and Sokone), Guinguinéo, Mbour (and Poponguine, Joal and Dioffior), Tivaouana (and Mekhe), Louga, Saint Louis, Saraya (2)	2007
	Togo	Savanes, Kara , Centrale, Plateaux, Maritime (1)	2007

Continued on next page

Table 3.3 – *Continued from previous page*

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
WER8604	Argentina	Bella Vista, Berón de Astrada, Capital, Concepción, Curuzú Cuatiá, Empedrado, Esquina, Ganeral Paz, General Alvear, Goya, Itatí, Ituzaingó, Lavalle, Mburucuy, Mercedes, Monte Caseros, Paso de los Libres, Saladas, San Cosme, San Luis del Palmar, San Martin, San Miguel, San Roque, Santo Tomé, Sauce, 25 de Mayo, Apóstoles, Cainguás, Candelaria, Eldorado, General Manuel Belgrano, Guarani, Iguazu, Leandro N. Alem, Libertador General San Ma, Montecarlo, Oberá, San Ignacio, San Javier, San Pedro (2)	2008
	Burkina Faso	Ouahigouya (3)	2008
	Cameroon	Abong Mbang, Nguelemendouka (3)	2008
	Cameroon	Buea, Mbongue (3)	2009
	CAR	Bozoum, Lobaye, Sangha Mbaéré, Boali, Damara, Bria, Ouadda, Yalinga (2)	2008
	CAR	Yaloke-Bossembelle (2)	2009
	Côte d'Ivoire	Abidjan (2)	2008
	Côte d'Ivoire	Madinani, Minignan (3)	2009
	Guinea	Bounouma, Urbain, Kouankan (3)	2008
	Guinea	Mandiana (2)	2009
	Guinea	Faranah (2)	2009
	Liberia	Nimba, Zorzor, Voinjama (1,2)	2009
	Paraguay	NA	2008
	Peru	Unspecified endemic areas	2004-08
	Republic of Congo	Cuvette-Ouest (1)	2009
	Sierra Leone	Bo (2)	2008
WER8634	Cameroon	Bandjoun, Foumbot, Kumba, Muyunka (3)	2010
	DRC	Titule, Buta (3)	2010
	Guinea	NA	2010
WER8828	Cameroon	NA	2012
	Chad	Biltine, Assongha, Sila (1)	2012
	Côte d'Ivoire	Béoumi, Katiola, Mankon, Séguéla (3)	2011
	Ghana	Builsa, Kassena-Nankana-West, Kintampo-South (2)	2012
	Senegal	Kédougou, Salémata, Saraya (2)	2012
	Sudan	Algeniema, Beida, Habila, Krenik, Azoum, Nertati, Wadi-Salith, Zalingi, Kass, Sharg-Aljebel, Alsiraif, Saraj-Omgra, Greater Bandasi, Forbranga, Mkjar, Nyala, Shataia, Edd Elfirsan, Elwihda, Kabkabyaia, Shaeriya (2)	2012

Continued on next page

Table 3.3 – *Continued from previous page*

Data source	Target country	Target subnational location/s (administrative level/s of target locations/s)	Timeframe
	Uganda	Abim, Agago, Ktigum, Lamwo, Pader (1)	2011
WER8927	Cameroon	Dschang, Ngoumou (3)	2014
	DRC	Lubao, Kamana, Ludimbi-Lukula (3)	2013
	Ethiopia	South Ari, Bena Tsmay, Dasenech, Gnangatom, Hammer, Jinka City, Malle, North Ari, Selamago (3)	2013
	Nigeria	Akwa Ibbom, Cross Rivers, Nasarawa (1)	2013
	Sudan	Northern Kordofan, Southern Kordofan (1)	2013
WER9026	Cameroon	West Region, Southwest Region (1)	2014
	DRC	Bondo, Buta, Kinkondja (3)	2014
WER9132	Cameroon	Adamaoua, Centre, East, North, North-West, West, South (1)	2015
	Sudan	Gedarif, Kassala, North Kordofan, Red Sea (1)	2015
WMMR020482	Ghana	NA	1981
WHO, 2016	DRC	Kwango, Cataractes Lukaya, Lulua, Lualaba, Kinshasa, Bas-Fleuve, Kasai (2)	2016

3.3.4 Processing supplementary data for Brazil

The Brazilian national immunisation program information system provided more detailed data for Brazil that contained the number of vaccine doses administered in each second administrative level (municipality) in 1-year age bands (from 0 to 4 years), 5-year age bands (5 to 14 years) and broader age groups, 15 to 59 years and 60 plus years, annually from 2006 to 2015. It was assumed that this data captured all doses administered to the Brazilian population between 2006 and 2015 and thus routine and mass preventive and outbreak campaign data from other sources during this period were excluded from coverage calculations. The WHO/UNICEF estimates for Brazil reported routine coverage of infants in YF risk areas from 1998 onwards [144], so these estimates were included up to 2005.

3.3.5 Processing demographic data

Annual age-structured population counts for each district were required to calculate population vaccination coverage and to estimate denominators for preventive and

outbreak response vaccination campaigns where coverage values could not be directly extracted from the data source. Estimates of these population counts were derived from a combination of UNWPP [145], WorldPop [21] and GPW [100] data. The latter two data sets were mosaicked together, with WorldPop data used in preference where they overlapped, to produce a fine scale human population density data surface for 2015, for each 1×1 km grid square.

The UNWPP estimates of the population size by country in five-year age bands for each year between 1950 and 2015 were disaggregated into annual birth cohorts using the methods described by Garske and colleagues [129]. This method required UNWPP data on infant and child mortality rates [146]. To achieve higher spatial resolution of population distribution, these estimates were combined with the fine scale human population density data for 2015. By allocating each 1×1 km grid square (or portion thereof) to a district, the proportion of each country's population living in any particular district was estimated. It was assumed that age distributions and population growth were homogeneous within each country, and that the population proportions for each district obtained from WorldPop and GPW for 2015 were applicable for all other years *e.g.* trends in urbanisation were not captured. To estimate the population in each district for 2016, the same growth rate was assumed as for 2015.

3.3.6 Applying correction terms to non-survey vaccination data

Where available, vaccination data from post-campaign coverage surveys were used in preference to administratively reported coverage data (*i.e.* health service provider reported estimates) because these generally provide more reliable coverage estimates [151–153]. Additionally, all administratively reported coverage data were multiplied by bias correction terms to adjust for variability in health system reporting reliability. These data included all routine vaccination data and any preventive or outbreak response data, for which the percentage of the target population vaccinated was not estimated from a post-campaign coverage survey (only six campaigns reported results of a post-campaign survey). Bias correction terms for each country and year were calculated as ratios of survey-derived Global Burden of Diseases, Risk Factors and Injuries Study

(GBD) [1] estimates of mean coverage for third dose of diphtheria-tetanus-pertussis-containing vaccine (DTP3) and WHO/UNICEF JRF administratively reported estimates [147, 148]. Coverage for DTP3 is often used as the main indicator for performance of routine vaccination services [152]. Data were available to calculate bias correction terms for each country from 1980 to 2015. These terms varied geographically and temporally, sometimes increasing estimates and other times decreasing them. For outbreak response campaigns conducted prior to 1980, bias corrections terms for 1980 were applied, and for 2016, correction terms for 2015 were used. The WHO also corrects reported administrative data [143], but the DTP3 terms were more comprehensive in time and space.

3.3.7 Estimating vaccination coverage

Using the age, time and location specific vaccination coverage and population datasets described above, each age cohort (from ages 0 to 99) in every district was tracked through time – from their birth year through to 2016 (*i.e.* the earliest cohort was born in 1871 in order to track coverage of individuals aged 99 in 1970) – with the coverage level updated whenever a routine, preventive or outbreak response campaign was conducted. Whilst the spatial distribution of populations was accounted for, it was assumed that the mortality rate for vaccinated and unvaccinated individuals was the same and that there was no mixing of populations between districts.

Because the strategies for within-population targeting of vaccinations were unknown, three alternative vaccination coverage estimates were generated, each corresponding to one of three targeting scenarios. *Untargeted*, *unbiased*: assuming vaccination history was not taken into account and all individuals had an equal chance of receiving a vaccine regardless of their previous vaccination status. *Targeted*: assuming that vaccination history was taken into account and only non-vaccinated individuals were targeted by immunisation campaigns. *Untargeted*, *biased*: assuming that vaccination history was not taken into account and that previously vaccinated individuals were more likely to be targeted inadvertently *i.e.* due to demographic biases in vaccination uptake. In the *untargeted*, *biased* scenario, for each vaccination campaign, it was assumed that all previously vaccinated individuals received vaccines before any unvaccinated individuals.

This scenario produces maximally conservative estimates of vaccination coverage, while estimates from the *targeted* scenario are maximally optimistic. To estimate population-wide vaccination under each scenario, the weighted mean proportion vaccinated across all age cohorts was calculated, weighted by population size.

Equations 3.1 to 3.3 below demonstrate how the proportion vaccinated, P, for each age cohort, was calculated under each of the three vaccination-targeting scenarios, given the proportion of the cohort previously vaccinated, V_1, and the proportion of the cohort targeted by any new vaccination activity, V_2.

Untargeted, *unbiased*:

$$P = V_1 + V_2 - V_1 V_2 \tag{3.1}$$

Untargeted, *biased* (conservative):

$$P = \max(V_1, V_2) \tag{3.2}$$

Targeted (optimistic):

$$P = \min(V_1 + V_2, 1) \tag{3.3}$$

3.3.8 Estimating the number of people requiring vaccination in 2016 to meet targets

Using each of the three sets of values of vaccination coverage for 2016 based on optimistic through to conservative scenarios of historical targeting, three sets of estimates were calculated for the number of individuals across all ages who still require vaccination to reach the 80% population coverage threshold recommended by the WHO to prevent outbreaks [122] for each at-risk district (districts classified by the WHO range maps for YF as either completely or partially 'endemic' or as having 'low potential for exposure' [27]). That is, the number of people who would require vaccination if one assumes that in the future unvaccinated individuals will be targeted and that the 80% coverage threshold applies across all populations. The resulting figures were then summed to obtain national, regional and global estimates. A global estimate of the number of people who would have required vaccination in 1970 was also calculated for comparison.

3.4 Results

The estimates of YF vaccination coverage, based on the *untargeted*, *unbiased* campaigns scenario, highlight substantial international and subnational variation, both spatially and temporally (Figure 3.1). Population vaccination coverage rates in countries within risk zones ranged from a maximum of 100% in parts of Amazonas State, Brazil, to zero coverage in parts of Central and East Africa (routine infant immunisation programs have not yet been introduced in these areas). Throughout the decades, vaccination coverage rates were higher overall in Latin America compared to Africa. Coverage rates were particularly high in Brazil during the 1970s and 1980s, declined slightly in the 1990s, and were again very high in most parts of the country by 2016. A similar pattern of waxing and waning was observed for many countries in West and Central Africa. Vaccination coverage rates were moderately high throughout these two regions in the 1970s, but coverage declined between 1970 and 2000, before reaching high rates again by 2016. However, a number of localised areas did achieve high coverage rates at earlier time points, including parts of Senegal, The Gambia, Ghana and Mali.

When contemporary coverage is considered *i.e.* the 2016 coverage map (Figure 3.1F), important gaps are apparent within the risk zones of Africa including large areas of Central and East Africa and parts of Nigeria, Niger, Sierra Leone, Liberia and Guinea-Bissau. In Latin America, low coverage was estimated for Guyana, Suriname, French Guiana and Colombia.

There was substantial geographical variation in the sensitivity of estimated vaccination coverage rates to alternative assumptions for vaccination-targeting (Figure 3.2), which demonstrates the potential impact of an important area of uncertainty in vaccination data. The percentage difference (Figure 3.2D) in coverage between the *targeted* and *untargeted*, *biased* alternative vaccination-targeting scenarios range from zero percentage difference in most of Central and East Africa and most countries in Latin America, to up to 73% in a small number of municipalities within Brazil.

Country-level estimates of YF vaccination coverage by age group in 2016 highlight the progress of routine infant immunisation programs in protecting children and young adults

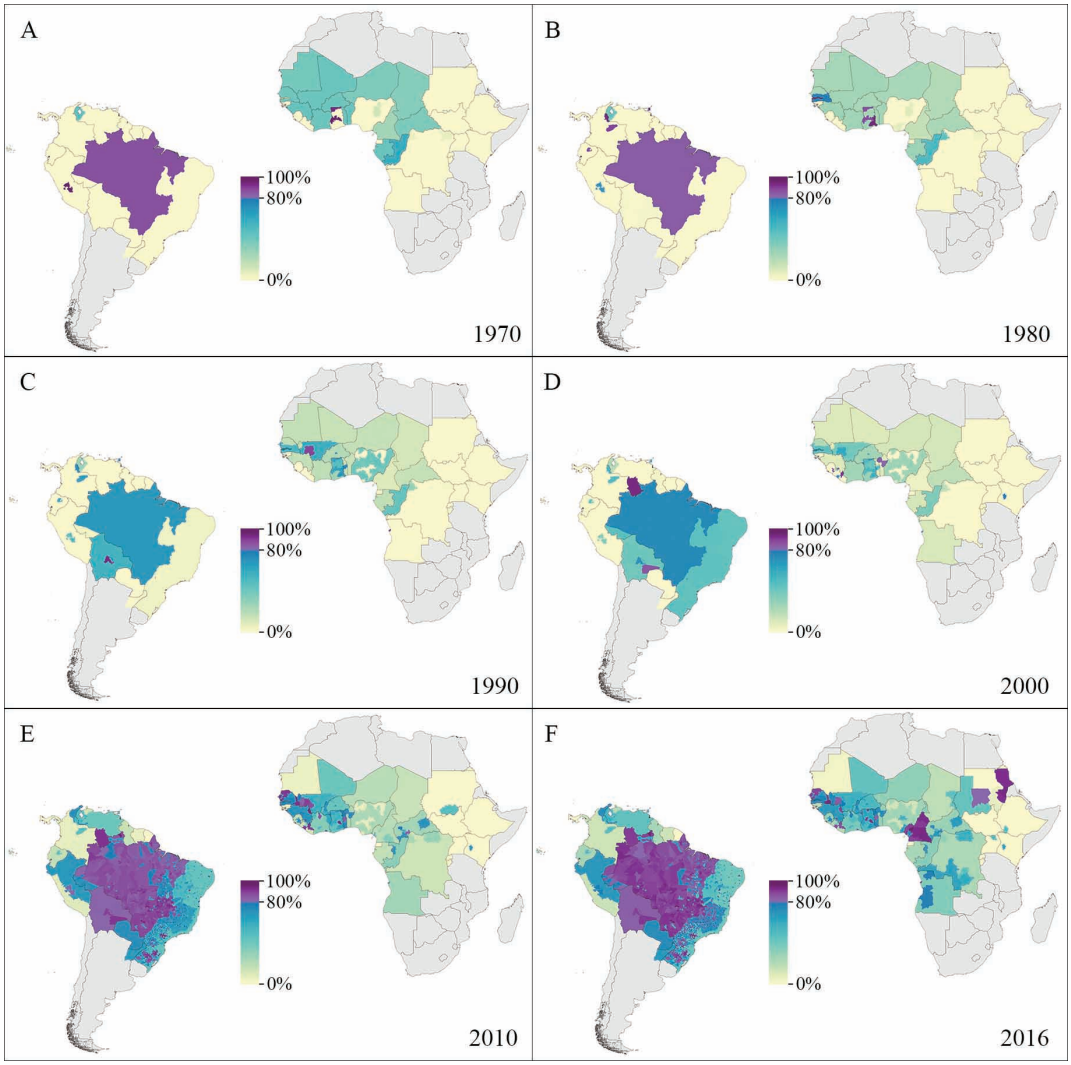


Figure 3.1: Yellow fever (YF) vaccination coverage from 1970 to 2016. Estimated proportion of the population across all age cohorts who have ever received a YF vaccine at the beginning of each decade (and in 2016) at the second administrative level for countries at risk of YF virus transmission, based on the *untargeted*, *unbiased* vaccination-targeting scenario.

on both continents (Figure 3.3), but they also revealed coverage gaps in adult populations for most countries. Angola, Guinea, Senegal, Togo, Cameroon, Paraguay, Bolivia and Brazil were exceptions to this, with moderate to high coverage estimated across all age groups. Estimates by age group from 1970 to 2010 are provided in Appendix B.

From the 2016 outputs, it was estimated that between 393.7 and 472.9 million individuals still require vaccination within at-risk districts globally, in order to achieve the 80% population coverage threshold recommended by the WHO to prevent outbreaks (Table 3.4). Between 361.4 and 396.0 million of these people live in countries on the

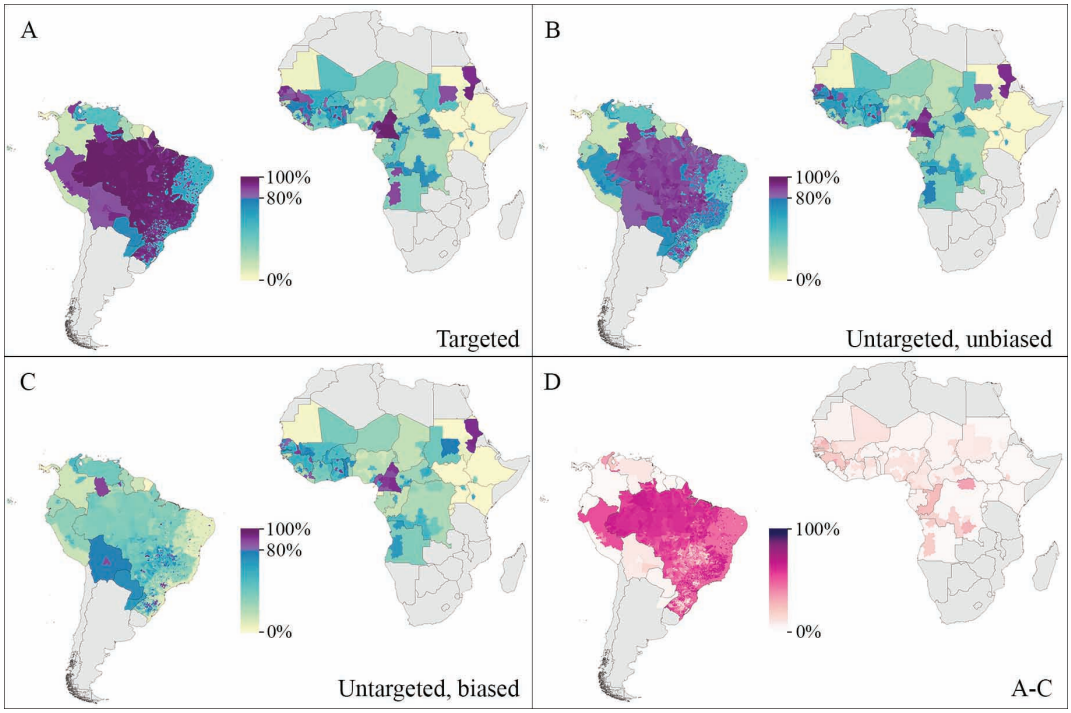


Figure 3.2: Contemporary yellow fever (YF) vaccination coverage scenarios. Estimated proportion of the population in 2016 across all age cohorts who have ever received a YF vaccine at the second administrative level in countries at risk of YF virus transmission. Vaccination coverage was calculated using three alternative vaccination-targeting scenarios; *targeted* (**A**), *untargeted*, *unbiased* (**B**), and *untargeted*, *biased* (**C**). Panel **D** displays the percentage difference in coverage between targeting scenarios **A** and **C**.

African continent and between 32.2 and 76.9 million live in Latin America. The national estimates of the total number of individuals requiring vaccination within each country are provided in Table 3.4. All countries except Bolivia, Peru, and Senegal contained at-risk districts with less than 80% population coverage under all three historical vaccination-targeting scenarios. The global estimates for 2016 represent between 43% and 52% of the population, compared to between 66% and 76% of the population who would have required vaccination in at-risk districts in 1970.

3.5 Discussion

This chapter describes the mapping of population-wide vaccination coverage for YF across all countries within risk zones from 1970 to 2016. Contemporary vaccination coverage is the result of cumulative campaigns, therefore it is important to incorporate information on past vaccination strategies to understand the present position. The coverage map for

Table 3.4: Number of individuals still requiring yellow fever (YF) vaccination in 2016. Estimated number of individuals (in millions) that still need to be vaccinated against YF in 2016 in order to achieve the population coverage threshold of 80% recommended by the WHO to prevent outbreaks, in every district or municipality in each country of Africa and Latin America classified as either completely or partially at-risk for YF [27]. Separate estimates are given based on the three alternative targeting scenarios for historical campaigns (conservative = *untargeted*, *biased*; untargeted = *untargeted*, *unbiased*; optimistic = *targeted*). Countries are ordered from greatest number needed to vaccinate to smallest across at-risk districts, for each continent. *Districts within areas classified by WHO range map as endemic or transitional. **Districts in countries partially or completely classified as endemic or transitional by the WHO range map in addition to districts classified as having low potential for exposure to YF virus.

	Number of individuals requiring vaccination, in millions					
	At-risk districts only*			All districts**		
Country	**Conservative**	**Untargeted**	**Optimistic**	**Conservative**	**Untargeted**	**Optimistic**
Nigeria	112.0	101.2	95.9	112.0	101.2	95.9
Ethiopia	74.4	74.3	74.2	80.9	80.8	80.8
Kenya	32.4	32.4	32.4	37.6	37.6	37.6
Uganda	32.2	32.2	32.2	32.4	32.4	32.4
DRC	30.4	28.6	26.9	34.9	32.5	30.3
Sudan	12.9	12.5	12.1	19.5	19.0	18.7
Rwanda	10.2	10.2	10.2	10.2	10.2	10.2
South Sudan	9.8	9.8	9.8	9.8	9.8	9.8
Niger	9.8	9.6	9.5	9.8	9.6	9.5
Burundi	8.9	8.9	8.9	8.9	8.9	8.9
Chad	8.4	8.2	8.0	8.5	8.3	8.1
Ghana	6.7	5.9	5.4	6.7	5.9	5.4
Côte d'Ivoire	6.4	5.9	5.5	6.4	5.9	5.5
Benin	4.8	4.6	4.4	4.8	4.6	4.4
Burkina Faso	4.3	3.9	3.6	4.3	3.9	3.6
Mali	4.7	3.7	3.2	4.7	3.7	3.2
Angola	4.6	3.5	3.3	4.6	3.5	3.3
Cameroon	3.3	3.0	2.8	3.3	3.0	2.8
Sierra Leone	2.5	2.5	2.5	2.5	2.5	2.5
Congo	3.0	2.2	1.7	3.0	2.2	1.7
CAR	2.3	2.2	2.1	2.3	2.2	2.1
Mauritania	2.1	2.0	2.0	3.3	3.2	3.2
Guinea	2.9	1.6	0.2	2.9	1.6	0.2
Liberia	1.6	1.6	1.6	1.6	1.6	1.6
Guinea-Bissau	1.1	1.1	1.1	1.1	1.1	1.1
Gabon	1.0	1.0	1.0	1.0	1.0	1.0
Senegal	2.2	0.7	0.0	2.2	0.7	0.0
Equatorial Guinea	0.7	0.7	0.7	0.7	0.7	0.7
Togo	0.4	0.3	0.2	0.4	0.3	0.2
Gambia	0.1	0.1	0.1	0.1	0.1	0.1
United Republic of Tanzania	-	-	-	41.7	41.7	41.7
Eritrea	-	-	-	5.3	5.3	5.3
Somalia	-	-	-	4.7	4.7	4.7
Zambia	-	-	-	1.5	1.5	1.5
São Tomé and Príncipe	-	-	-	0.1	0.1	0.1
Totals for Africa	396.0	373.9	361.4	473.4	450.8	437.8
Colombia	23.5	23.0	22.7	29.9	29.3	29.0
Brazil	40.3	6.8	2.1	121.8	49.7	27.1
Venezuela	6.2	4.7	3.9	12.6	10.3	9.1
Panama	1.0	1.0	1.0	1.0	1.0	1.0
Peru	3.1	0.8	0.0	18.7	15.3	14.1
Ecuador	0.6	0.6	0.6	8.7	8.7	8.7
Paraguay	0.5	0.5	0.4	0.5	0.5	0.4
Guyana	0.4	0.4	0.4	0.4	0.4	0.4
Suriname	0.4	0.4	0.4	0.4	0.4	0.4
Argentina	0.2	0.2	0.2	1.3	1.3	1.3
Trinidad and Tobago	0.2	0.2	0.2	0.2	0.2	0.2
French Guiana	0.2	0.2	0.2	0.2	0.2	0.2
Bolivia	0.1	0.0	0.0	0.2	0.0	0.0
Totals for Latin America	76.9	38.8	32.2	196.1	117.4	92.1
Global totals	472.9	412.8	393.7	669.5	568.2	529.9

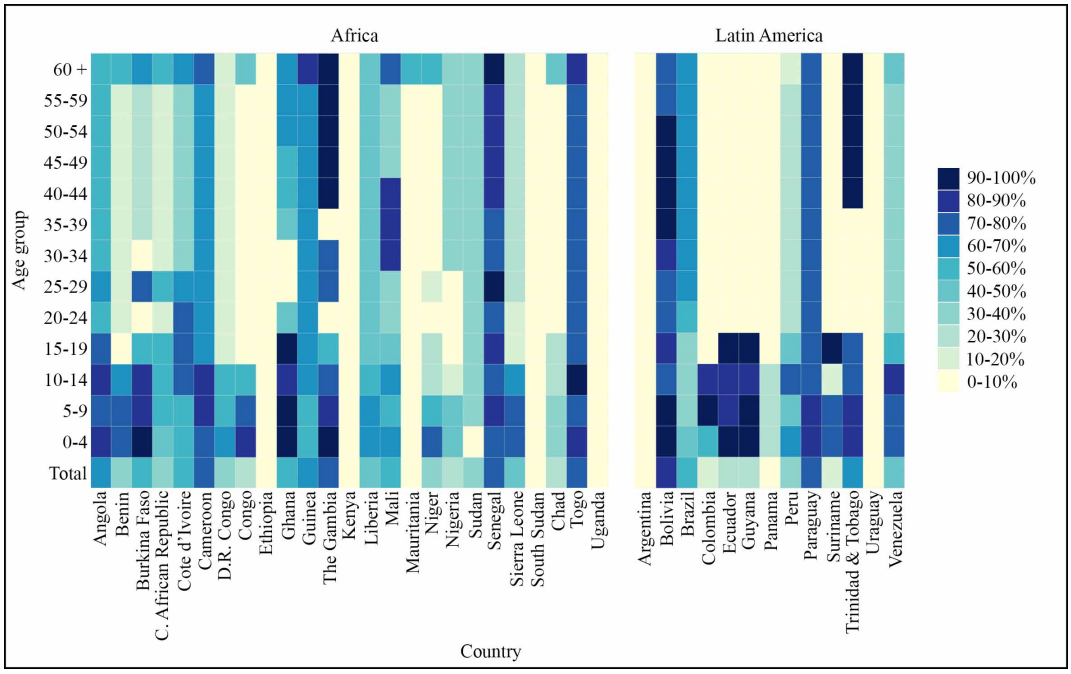


Figure 3.3: National yellow fever (YF) vaccination coverage by age group. Estimated proportion of individuals within each age group in 2016 that have ever received a YF vaccine for each country at risk of YF virus transmission, based on the *untargeted*, *unbiased* vaccination-targeting scenario. Plots for the beginning of each decade from 1970 to 2010 are provided in Appendix B. Congo = Republic of the Congo. D.R. Congo = Democratic Republic of the Congo. C. African Republic = Central African Republic.

2016 highlights key gaps in present levels of vaccination coverage within YF risk zones and enables the identification of areas that currently do not meet the WHO recommendation of 80% population coverage, and are therefore at risk of an uncontrolled outbreak.

Conservative estimates of the number of people requiring vaccination in 2016 to meet the 80% threshold for Africa and Latin America are approximately 10% smaller and 3% greater, respectively (Table 3.4), than the WHO's estimates of doses required for mass campaigns as part of the Global Strategy to Eliminate Yellow fever Epidemics, 2017-2026 [122]. The district level estimates can assist in the planning of vaccine delivery, and if combined with estimates of the number of doses required for future routine infant immunisation programs, they can be used in the planning of emergency stock-piling and manufacturing surge capacity. Notably, the 80% threshold is an average; in reality, the critical vaccination coverage needed to prevent outbreaks is expected to vary substantially across different settings depending tightly on the basic reproduction number, R0 [154]. Using estimates presented in this thesis of the proportion vaccinated

for each district, others can re-calculate the number of individuals requiring vaccination to reach any new threshold.

The estimates of vaccination coverage over time (Figure 3.1) show the cumulative impact of multiple different vaccination strategies combined with the different demographic profiles of the two at-risk regions. The results for Africa from 1970 to 2010 corroborate the analysis by Garske and colleagues [129]. I employed a similar approach but, importantly for this study, I assessed the range of estimates produced when alternative vaccination-targeting scenarios are considered. I also included more recent vaccination coverage data and extended the map to Latin America. Overall this work highlights the stark differences between Latin America and Africa, which may, in part, explain the difference in the observed number of outbreaks between these two regions [131].

Combined with vector control efforts, high levels of vaccination coverage in Brazil has led to a substantial reduction in the number of YF cases since the 1930s, especially those arising from urban transmission cycles. Coverage rates were particularly high in Brazil during the 1970s and 1980s, but then declined slightly, resulting in a resurgence of sylvatic transmission in the 1990s and early 2000s [131]. This resurgence prompted further mass vaccination campaigns and the resulting 2016 vaccination coverage rates were estimated to be very high for most parts of the country. Lower levels of vaccination coverage were estimated at the eastern edge of the risk zone in Brazil for 2016, including within the State of Minas Gerais, where a YF outbreak arose in December 2016 [121].

Moderately high vaccination coverage rates throughout much of West and Central Africa in the 1970s (Figure 3.1) was the result of mass preventive campaigns through the 1940s to 1960s [150]. Coverage waned between 1960 and 2000 in most areas, due to limited vaccination activity, the birth of new unvaccinated cohorts and the gradual reduction in the proportion of older protected cohorts through mortality. Vaccine stock shortages, which have been frequently reported in the African region [122], would have also contributed to lower than recommended coverage levels. A number of localised areas, including parts of Senegal, The Gambia, Ghana and Mali, were exceptions to this pattern as they conducted outbreak response campaigns that reached high levels of coverage. In 2010, vaccination coverage in Angola was estimated to be low across all

districts, except Luanda with 64% coverage. Following outbreak response campaigns in 2016, coverage had reached nearly 90% for Luanda, and just under the 80% threshold across a number of other districts in the north and west of the country. Additionally, the most recent DON report included in this study (published on June 14, 2016) indicated that vaccination campaigns were ongoing in a number of these districts [155].

Many countries in West and Central Africa introduced the YF vaccine into their routine infant immunisation schedules between the late 1980s and early 2000s [136], but the impact of these efforts will take time to influence population level coverage in the absence of campaigns targeting adults born before the late 1980s (Figure 3.3). For adequate protection against outbreaks, the vaccine should be introduced into routine infant immunisation schedules, followed by catch-up campaigns for adult populations. The success of this approach is exemplified by The Gambia, which reported no locally-acquired cases for more than two decades since 1979, when they conducted a post-outbreak mass vaccination campaign targeting all ages and in tandem introduced the vaccine into their routine infant immunisation schedule. Increased coverage rates were achieved overall in West and Central Africa in 2010 and 2016 since large-scale preventive campaigns supported by the Yellow Fever Initiative and the GAVI Alliance were implemented in 2006 [156].

My dataset of vaccination coverage, compiled from a variety of sources, highlights concerns about the completeness and accuracy of the reports on vaccination activity. To assess the impact of uncertainty regarding whether vaccination history was taken into account when targeting individuals in mass campaigns, vaccination coverage was calculated using three different vaccination coverage scenarios ranging from conservative to optimistic. Brazil was most sensitive to different assumptions about the vaccination-targeting scenarios (Figure 3.2), because they have conducted sustained vaccination activities since the 1940s; when vaccination coverage in each Brazilian age cohort was tracked through time, there was ample opportunity for alternative estimates to diverge from one another. Additionally, the vaccination activity data for Brazil used in my analysis were more complete, and spatially and temporally disaggregated, than for other countries. Lessler and colleagues [157] describe another approach for estimating uncertainties

associated with vaccination targeting, which uses administrative data and cross-sectional survey data, across a range of ages, on measles vaccination to calculate the size of the population systematically missed by vaccination activities for specific countries. However, this method requires cross-sectional survey data on YF, and these are insufficiently available, especially for adult populations.

Estimates of vaccination coverage at the country or province level, including most of the routine infant vaccination data and much of the preventive and reactive mass vaccination campaigns used in this study, often smooth out important sources of spatial heterogeneity. Characterisation of such heterogeneities is important for planning spatially targeted interventions. High-resolution vaccination coverage data, such as that used here for Brazil, would improve certainty of coverage estimates for other countries. Data from household surveys would afford increased certainty and resolution of vaccination coverage estimates, but are only available for limited times, locations and age cohorts, and thus require geospatial modelling to interpolate across time and space.

The maps of vaccination coverage do not necessarily translate into an absolute measure of vaccine-derived immunity, since vaccine efficacy is not 100% (although evidence suggests that YF vaccine efficacy is close to 100% and the WHO concluded recently that a single dose gives lifelong protection) [128]. Likewise, the proportion of individuals with immunity to YFV is not necessarily equivalent to the proportion of individuals with vaccine-derived immunity; this is due to the possible acquisition of immunity *via* natural infection, especially in locations where outbreaks have occurred. The coverage estimates presented here could be combined with estimates of vaccine efficacy, and of natural infection rates to estimate the number of individuals susceptible to symptomatic YF infection and to quantify the impact of vaccination on YF incidence and outbreak potential.

Vaccination coverage was tracked across several decades, over which I assumed that mortality rates were the same for vaccinated and unvaccinated individuals, and that there was no population mixing between districts. The prevalence of YF across the study period would not have been high enough for this first assumption to have a large effect on the proportion of the population vaccinated over time. Data were not

available to account for the latter, but it is important to note that this may have impacted results. If migration rates between districts were not balanced between vaccinated and unvaccinated individuals, this would have resulted in overestimates of coverage for some districts, and underestimates for others.

3.6 Conclusions

Yellow fever cannot be eliminated due to the presence of non-human, wildlife reservoirs, which maintain an enzootic transmission cycle of the virus in non-urban settings. However, elimination of YF epidemics can be achieved if effective vector control, vaccination and disease surveillance are enforced and maintained. The results of this chapter highlight both important progress and gaps in YF vaccination coverage within risk zones and provide credible estimates of the doses required for supplementary vaccination campaigns. In the following chapter, this information is coupled with fine-scale estimates of geographical disease risk, in order to identify populations most vulnerable to YF infection. These outputs can assist policy makers in developing the most effective vaccination strategies that work towards the prevention of outbreaks.

4

Modelling the existing and potential global infection risk zones of yellow fever

Contents

4.1 Introduction **64**
4.2 Methods **66**
4.3 Results **84**
4.4 Discussion **86**
4.5 Conclusions **90**

In order to account for human vaccination coverage, and thus produce the most accurate predictive map of YF risk, the standard disease niche modelling approach applied in *Chapter 2* must be developed further. This chapter describes the development of a novel Poisson point process niche model, which is then used to predict YF infection risk in humans, and demonstrates how vaccination coverage can be efficiently accounted for in disease niche models. The model explicitly incorporates: the locations of reports of YF infection, estimates of the population susceptible to YFV infection (calculated from the vaccination coverage maps produced in *Chapter 3*), spatial bias in disease reporting, and a range of environmental and biological variables thought to influence disease transmission. The work in this chapter has been published in *The Lancet Global Health* [158]. General background on YF can be found at the beginning of *Chapter 3*.

4.1 Introduction

As discussed in *Chapter 3*, substantial outbreaks of YF in Angola and Brazil in the past two years suggest a need to assess current control strategies. Before interventions can be effectively implemented, more information is needed about which populations within risk zones should be targeted. Yellow fever cases are substantially underreported and the global distribution of the disease remains inadequately characterised. Additionally, YF has been conspicuously absent from Asia to-date, despite the presence of competent vector populations, and suitable climatic conditions. Epidemiologists have long been concerned about the potential for YFV introduction into *Ae. aegypti*-inhabited cities, where human populations lack naturally acquired or vaccine-derived immunity. In March 2016, several YFV-infected Chinese workers returned to China from the outbreak zone in Angola, which, coupled with the depletion of global emergency vaccine stockpiles, elevated concerns that YF would gain an uncontrollable foothold in East and Southeast Asia [159]. Fortunately these importation events occurred in Beijing, Shanghai and Fujian, where *Ae. aegypti* is not established, and no secondary cases resulting from local transmission were detected [160]. Figure 4.1 displays global estimated coverage of YF vaccination (from *Chapter 3*) overlaying the predicted range of *Ae. aegypti*, highlighting large areas within the vector's range where human populations are 100% susceptible to YFV infection.

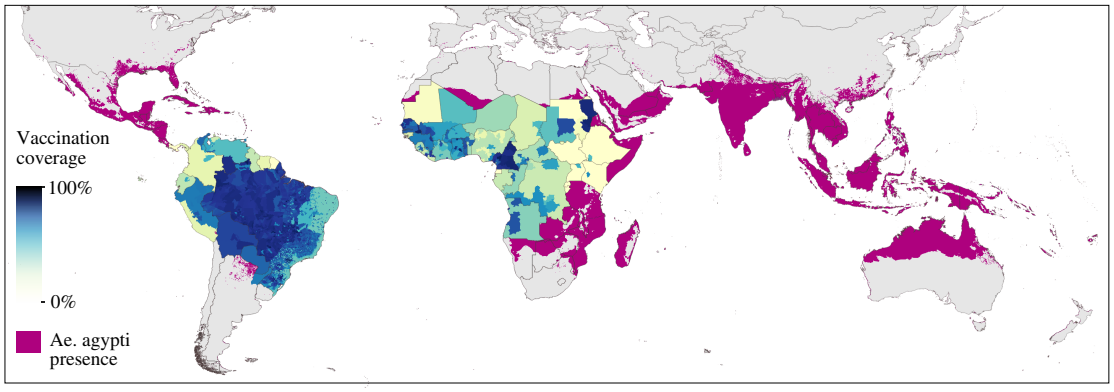


Figure 4.1: Yellow fever (YF) vaccination coverage versus range of *Aedes aegypti*. Estimated YF vaccination coverage (from *Chapter 3*), overlaying the binary distribution map for YF's main urban vector, *Ae. aegypti*, highlighting areas outside the limits of YF vaccination activities where coverage is zero and *Ae. aegypti* is present. *Aedes aegypti* presence was indicated for all grid squares above the threshold value of the continuous species distribution map that incorporated 90% of observed occurrences of *Ae. aegypti* [71].

The course of the Angolan outbreak, and international response, reiterates the need for a sustained policy of preventive vaccination of at-risk populations, as well as travellers to at-risk areas, not only to reduce the risk of epidemics, but also of exportation events into previously unexposed populations. An improved understanding of YF's contemporary distribution and potential for spread into new areas is required to inform optimal control and prevention strategies. Previous studies have quantified YF risk at national levels, as well as at first or second order subnational units, or by ecological zones [27, 122, 129, 161]; however, these results may mask considerable differences in YF risk at more local levels. In 2006, Rogers and colleagues mapped yellow fever risk across both Africa and South America, as well as predicted climatic suitability for transmission outside these regions. This study analysed yellow fever data up to 2005, did not provide any estimate of spatial variation in disease incidence, or account for vaccination. Any effort to understand the distribution of YF within risk zones must account for vaccination coverage, otherwise, for example, risk may be overestimated in areas of high vaccination coverage. An analysis of YF data up to 2011, evaluated the impact of YF vaccination on disease burden in Africa, but the application of these findings may be limited when used for more recent years, and for other regions that experience YF epidemics (*e.g.* Amazon Basin). To date, no study has quantified YF risk or incidence across global risk zones at high geospatial resolution, or evaluated potential for receptivity to YF in areas outside current risk zones. For areas outside the defined risk zones, the most recent efforts to represent populations at risk have been based on the distribution of *Ae. aeygpti* [132, 162]. Vector presence is necessary for transmission, but it is not sufficient, and these estimates ignore many other factors that contribute towards the establishment and maintenance of YFV transmission [6].

In this chapter, subnational variation in YF risk is estimated across its risk zones and receptivity to YFV transmission is predicted outside the defined risk areas, using a model based on YF infection data, vaccination coverage rates, and a range of environmental and biological covariates. Using previously published estimates of annual national case numbers, this model is then used to map subnational variation in YF incidence across at-risk countries, as well as estimate the number of cases averted by vaccination globally.

4.2 Methods

4.2.1 Overview

Geo-referenced reports of symptomatic YFV infection in humans were collated for the 47-year period from 1970 to 2016. Records prior to 1970 were not included since the aim of this chapter was to map the contemporary spatial distribution of YF, and this time point coincides with the re-infestation of Latin America with *Ae. aegypti* following 50 years of eradication efforts [163], and the cessation of organised large-scale YF vaccination activities in West and Central Africa [129]. A Poisson point process boosted regression tree model was fitted to the database of YF disease reports along with a range of environmental and socioeconomic variables hypothesised to influence YF transmission. By explicitly accounting for population density, spatial variation in disease reporting rates, and vaccination coverage, the relative risk of apparent infection was estimated for each 5×5 km grid square across risk zones. To map within-country variation in annual incidence, previously published national case estimates from GBD were used [164]. The incidence map was averaged over 47 years to visualise spatial variation without the large temporal fluctuations seen for YF. Finally, the fitted model for relative infection risk was used to predict areas of receptivity to YFV transmission beyond contemporary risk zones. A schematic overview of the process followed is provided in Figure 4.2.

4.2.2 Assembly of the geo-referenced YFV human infection occurrence dataset

A database was assembled of locations where at least one symptomatic human infection of YFV has been reported in any given year. This information was extracted from a variety of sources including online sources, peer-reviewed literature and WHO reports. The data from online sources were collated by the automated HealthMap system, as described elsewhere [22, 24].

HealthMap alerts for YF were obtained for the years from 2006 to 2015, and then checked manually for validity. To collate the data from published literature, ISI Web of Science and PubMed were searched using the term ‘yellow fever’ for all articles up until April 1, 2016. This returned 5,387 records. The article abstracts were reviewed

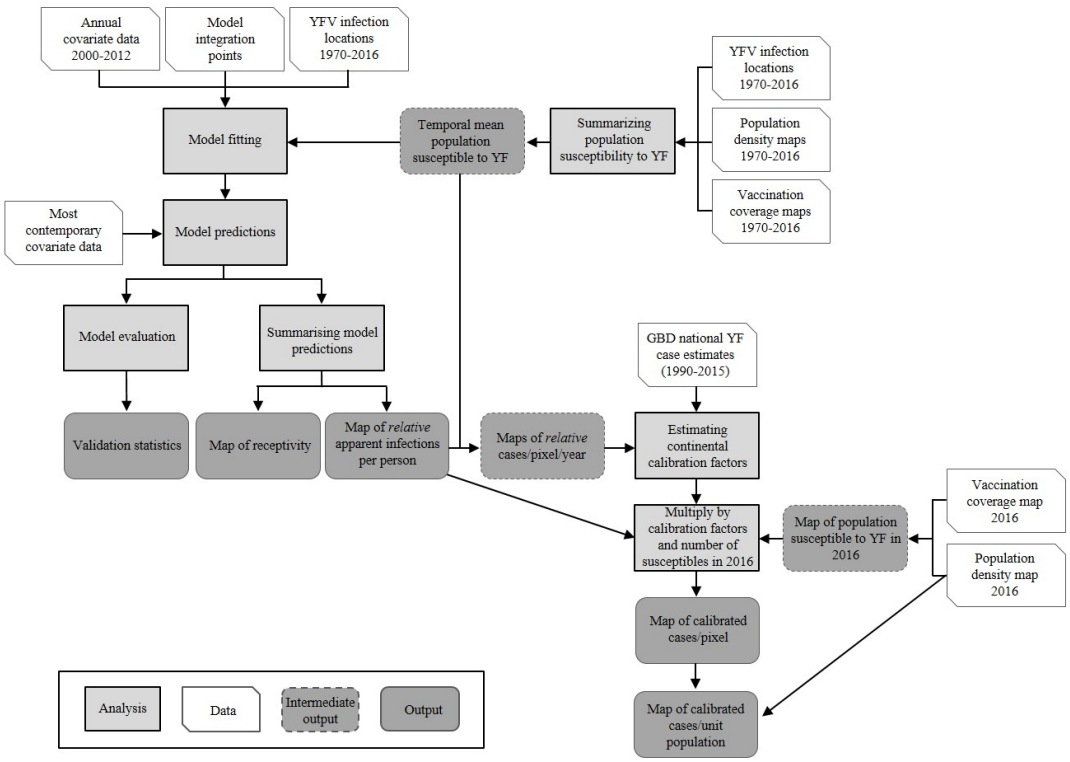


Figure 4.2: Yellow fever (YF) disease mapping methods. Schematic overview of the method used to map YF risk, representing input data, analyses, intermediate outputs and final outputs. YFV = yellow fever virus. GBD = Global Burden of Diseases, Risk Factors and Injuries Study.

and for those containing possible geographical information on YFV infection, the full text was obtained (725 articles).

Location information regarding laboratory confirmed YFV infections in humans was extracted from each peer-reviewed article or report (602 articles/reports) according to previously established protocols [36, 37]. Serological evidence of infection in healthy individuals was not included in the database due to uncertainty regarding time and site of infection and the patient's YF vaccination history. If an article reported multiple cases in the same location (same 5×5 km grid square) within the same calendar year, this was recorded as a single YF occurrence record. Spatial coordinates were assigned to each reported site of infection using a combination of information from the article or report and online gazetteers such as 'Google Maps', 'OpenStreetMap' or 'Geonames'. For locations smaller than 5×5 km in area (termed points), only the latitude and longitude in decimal degrees were recorded. The remaining sites, referred to as polygons, were

assigned an administrative area code (*e.g.* for a province or district) which were linked to shapefiles from the FAO [165], for first and second administrative subnational units, or Global Administrative Areas [166], for third and fourth administrative subnational units.

The final dataset included 1,155 records, comprising 751 point locations and 404 polygon locations (Figure 4.3). Of all records in the final dataset: 402 were diagnosed *via* PCR-based or other genetic detection techniques; 444 were diagnosed *via* serological evidence; and the remaining 308 were reported as confirmed cases without specifying the diagnostic test used.

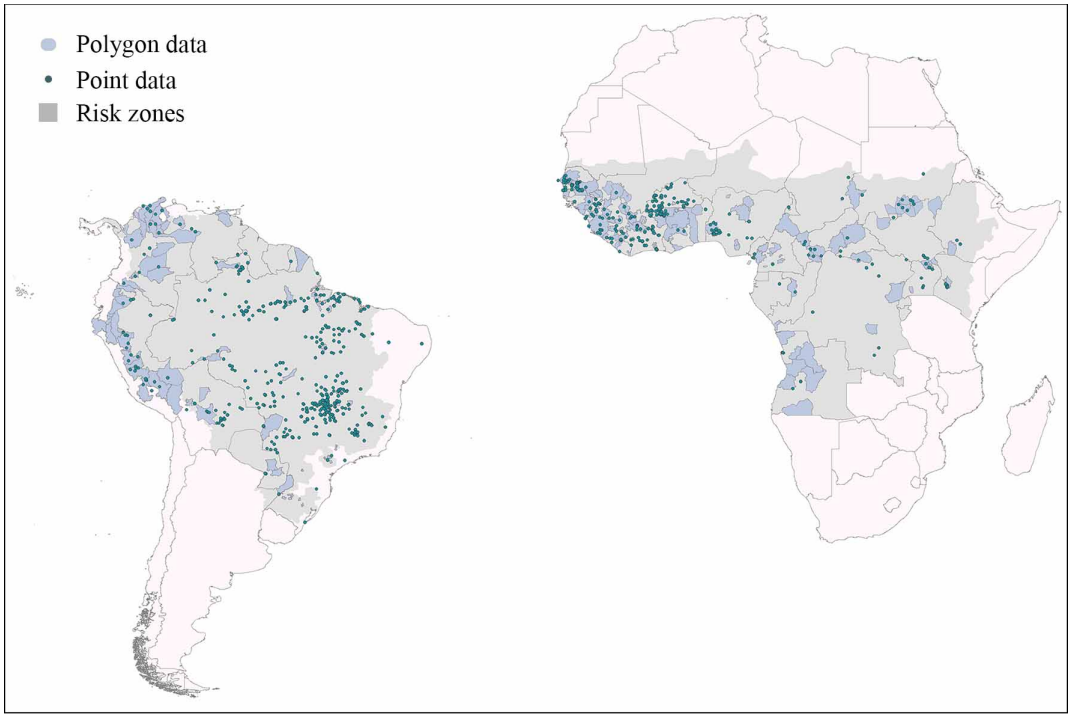


Figure 4.3: Yellow fever (YF) occurrence data. Polygon and point locations of reports of symptomatic human YF virus (YFV) infection from 1970 to 2016. Points = YFV infections reported in locations smaller than 5×5 km in area, where only the latitude and longitude of the site were recorded. Polygons = YFV infections reported in locations larger than 5×5 km in area, which were assigned an administrative area code (*e.g.* province or district). Grey areas = contemporary YF risk zones as defined by Jentes and colleagues [27].

4.2.3 Summarising YF vaccination coverage and population density data

Accounting for the effects of YF vaccination in a temporally static model of YF infection risk, required an estimate of the average susceptible population size across the study time period. To calculate this temporal average, I used estimates of vaccination coverage

for every five years from 1970 to 2016 (developed in *Chapter 3*), and a combination of WorldPop [21], GPW [100], and UNWPP human population data [145].

Since rates of YF reporting varied over the study time period, a weighted average for vaccination coverage and population size were calculated, with weights representing the proportion of YF occurrences reported in each five-year time period. The vaccination coverage estimates were based on the *untargeted*, *unbiased* vaccination-targeting scenario described in *Chapter 3*. A plot of average vaccination coverage is provided in Appendix C.5, alongside the vaccination coverage map for 2016.

To estimate average population size over the study time period, human population density surfaces at 5×5 km resolution for 2000, 2005, 2010 and 2015 were generated using a combination of WorldPop [21] and GPW [100] gridded population data. These two datasets were mosaicked together, with WorldPop data used in preference where they overlapped. Since WorldPop data were only available from 2000 to 2015, the UNWPP national population estimates [145] were used to calculate population growth rates from 1970 to 2000, and for 2016. These rates were applied to the WorldPop/GPW surfaces to generate population density surfaces for every five years from 1970 to 1995, and for 2016.

4.2.4 Model explanatory covariates

Ten 5×5 km gridded data surfaces of a range of environmental and biological factors thought to influence YFV transmission were supplied to the model as potential explanatory covariates (see Table 1). Plots of each covariate data surface are included in Appendices C.1 and C.2.

The environmental covariates included measures of the mean values for each pixel for TCW and EVI (both related to surface moisture), elevation, and three land cover classes: urban and built-up, evergreen broadleaf forests, and cropland/natural vegetation mosaic. Further details on the construction of these surfaces is provided in *Chapter 2*. A data surface of temperature suitability for persistent established *Ae. aegypti* populations was also included [26], since it has been shown that an index of temperature suitability that incorporates the relationships between adult mosquito longevity and temperature can be a better predictor of mosquito-borne disease risk than temperature alone [30].

Table 4.1: Yellow fever (YF) model explanatory covariates. List of covariates supplied to the YF model and their data source. Plots of each covariate data surface are included in Appendices C.1 and C.2

Covariates	Date source
Evergreen broadleaf forest, urban and built up, and cropland/natural vegetation mosaic land cover classes (proportional)	MODIS land cover product [99]
Elevation	Shuttle Radar Topography Mission [101]
tasselled cap wetness, a measure of surface moisture (mean)	Gap-filled MODIS satellite data [97, 98]
Enhanced vegetation index (mean)	Gap-filled MODIS satellite data [97, 98]
Aedes aegypti temperature suitability index	Brady and colleagues, 2014 [26]
Ae. aegypti habitat suitability	Kraemer and colleagues, 2015 [50]
Habitat suitability for suspected YFV reservoir species	*Chapter 4*

Finally, fine-scale predictive species distribution maps for YF's main urban vector species, *Ae. aegypti*, and candidate non-human primate reservoir species were also included.

Details on the construction of this covariate are provided below. Since the original data for these covariates came from a variety of sources, all covariate grids were standardised to ensure identical spatial resolution, extent and boundaries.

4.2.5 Mapping the distribution of suspected non-human primate reservoirs of yellow fever

Predictive distribution maps of suspected non-human primate reservoir species of YFV were produced using a standard Bernoulli niche modelling approach, as described in *Chapter 2*. The Poisson point process niche modelling approach described later in this chapter allows for the incorporation of other factors (such as vaccination coverage), and clearer interpretation of the model output in certain circumstances, but at the cost of greater computational intensity, especially in model validation. Since there were no such factors to account for here, and the output was only being used as a covariate (thus the absolute value was not important), the Bernoulli approach was preferable.

Identifying candidate non-human primate reservoir species

In Africa, and Central and South America, various species of non-human primate have been implicated as natural hosts and reservoirs of YFV [5]. The role of a particular host species in the epidemiology of YF is dependent on their ability to circulate YFV at sufficiently high titers and to provide subsequent transmission of virus to vectors above

minimum threshold levels [167]. When infected with YFV, certain non-human primate species on the African continent develop viremias sufficient to infect mosquitoes, but they are relatively resistant to the pathogenic effects of the virus. Conversely, certain non-human primate species in Central and South America are highly susceptible to YFV infection and develop severe and often fatal disease. Epizootics involving the deaths of these non-human primate species can serve as a warning to nearby human populations of YF risk. The virus moves slowly from monkey group to neighbouring monkey group in large epizootic waves that leave behind a few surviving, immune individuals [5, 167].

To determine which non-human primate species should be included in the analysis, PubMed was searched on February 11th 2017, using the search terms (yellow[All Fields] AND (‘fever’ [MeSH Terms] OR ‘fever’[All Fields])) AND (reservoir[All Fields] OR (‘primates’[MeSH Terms] OR ‘primates’[All Fields] OR ‘primate’[All Fields]) OR host[All Fields]) without any language or date restrictions. This search returned 3,830 records. Article abstracts were reviewed, and the full text was obtained for those articles containing possible evidence implicating a non-human primate species as a possible reservoir host for YFV (99 articles). Information was extracted from 61 of 99 full text articles.

Due to the short duration of viremia (2-5 days) and lack of clinical disease, most studies of African sylvatic YF transmission involved either serological prevalence studies of wild monkeys [168–175], or laboratory experiments testing for sufficient levels of viremia following inoculation with the virus [176–178]. Virus isolation (or positive antibody tests) from sera or tissue samples of diseased or deceased wild monkeys during active searches for epizootics were frequently reported in studies conducted in Central and South America [179–188]. In some cases, transmission from monkey-to-monkey *via* a mosquito vector was demonstrated, under laboratory conditions [189, 190].

All non-human primate species of Africa or Central and South America, where naturally acquired infection with YFV has been confirmed by serological or PCR-based techniques, or sufficient levels of viremia post-inoculation with YFV under laboratory conditions has been demonstrated, were included in the analysis. Non-human primate species meeting these criteria are listed in Table 4.2. Occurrence data were not available

for *Marikina geoffroyi*, *Cercopithecus ihoesti*, or *Papio doguera*, so these species were subsequently not included in the analysis.

Table 4.2: Non-human primate species included in candidate reservoir habitat suitability model. All non-human primate species where naturally acquired infection with YFV has been confirmed by serological or PCR-based techniques, or sufficient levels of viremia post-inoculation with YFV under laboratory conditions has been demonstrated, were included in the analysis.

Region	Common name	Scientific name	References
Africa	Ornate Titi Monkey	*Callicebus ornatus*	[190]
	Grey-cheeked Mangabey	*Cercocebus albigena*	[173, 174, 191]
	Grivet Monkey	*Chlorocebus aethiops*	[190]
	Guereza	*Colobus guereza/abyssinicus*	[174, 178, 191–193]
	King Colobus	*C. polykomos*	[168, 173]
	Blue Monkey	*Cercopithecus mitis*	[173, 174, 194]
	Mona Monkey	*C. mona*	[173, 175, 191]
	Putty-nosed Monkey	*C. nictitans*	[171, 175]
	De Brazza's Monkey	*C. neglectus*	[175]
	L'hoest's Monkey	*C. lhoesti*	[195]
	Red-tailed Monkey	*C. ascanius*	[170, 172, 174, 175, 191]
	Diana Monkey	*C. diana*	[168]
	Patas Monkey	*Erythrocebus patas*	[169, 175, 196, 197]
	Senegal Bushbaby	*Galago senegalensis*	[192, 198, 199]
	Greater Bushbaby	*G. crassicaudatus* (*Otolemur crassicaudatus*)	[194, 199]
	Yellow Baboon	*Papio hamadryas/cynocephalus*	[175, 193]
	Guinea Baboon	*P. papio*	[169]
	Dog Face Babbon	*P. doguera*	[173, 174]
	Bay Colobus	*Piliocolobus badius*	[168, 174, 200]
Central and South America	Brown Howler Monkey	*Alouatta guariba*	[180, 181, 201]
	Mantled Howler Monkey	*A. palliata*	[202–205]
	Black Howler Monkey	*A. caraya*	[180, 181, 184, 201, 206, 207]

Continued on next page

Table 4.2 – *Continued from previous page*

Region	Common name	Scientific name	References
	Venezuelan Red Howler Monkey	*A. seniculus*	[187, 188, 208, 209]
	Red-handed Howler Monkey	*A. belzebul*	[182]
	Black Spider Monkey	*Ateles paniscus*	[210]
	Black-headed Spider Monkey	*A. fusciceps*	[211]
	Geoffrey's Spider Monkey	*A. geoffroyi*	[203, 211]
	Night Monkey	*Aotus trivirgatus*	[190]
	Panamanian Night Monkey	*A. zonalis*	[211]
	Tufted Capuchin	*Cebus apella*	[212–214]
	White-headed Capuchin	*C. capucinus*	[203, 213]
	Marmoset	*Marikina geoffroyi*	[204, 211]
	White-faced Saki	*Pithecia pithecia*	[208, 209]
	Golden-handed Tamarin	*Saguinus midas*	[208, 209]
	Cotton-top Tamarin	*S. oedipus*	[190]
	Common Squirrel Monkey	*Saimiri scieureus*	[190, 215]
Asia	Long-tailed Macaque	*Macaca fascicularis*	[216]
	Toque Macaque	*M. sinica*	[217]
	Rhesus Macaque	*M. mulatta*	[215, 218]

In order to predict receptivity to YFV transmission in regions outside its current risk zones, I also mapped the distribution of non-human primate species of Asia that have demonstrated susceptibility to YFV infection under laboratory conditions. These included *Macaca fascicularis*, *M. sinica* and *M. mulatta* [215–218] (Table 4.2). In particular, *Macaca mulatta* (rhesus macaques) infected with YFV exhibit severe disease that generally models the course of the disease in humans and serve as a highly suitable animal model for YFV research [219].

Reservoir model occurrence data

Separate species distribution models were carried out to predict the distribution of potential reservoir species on each continent. Occurrence data for each species were retrieved from GBIF [93] totalling 931, 1,193, and 820 records for African, Latin American, and Asian species, respectively. Records falling outside the IUCN range maps [113] for each species were excluded from the analysis. Following standard practice for species distribution mapping, occurrence records were supplemented with a large number of background records to represent areas where the species has not been reported as present. To improve model performance and account for spatial bias in the occurrence data, background records were selected to reflect the spatial distribution of survey effort, as per Phillips and colleagues [55]. Specifically, the locations of all other mammal surveys held by GBIF were used as background data. This resulted in 5,661, 14,274, and 743, background points for the African, Latin American and Asian reservoir models, respectively.

Reservoir model covariates

Ten explanatory covariates were included in each model, including elevation, EVI (mean), LST (daytime mean), and TCW (mean) as well as six proportional land cover classes: cropland/natural vegetation mosaic, woody savannas, savannas, urban and built-up, evergreen broadleaf forest and deciduous broadleaf forest. Further details on the source and construction of each covariate data surface are provided in *Chapter 2*. Plots of each covariate data surface are included in Appendix C.3. Since the collection date was not available for the vast majority of occurrence records, land cover data surfaces for 2012 were linked to all occurrence records.

Reservoir model fitting

An ensemble of 100 Bernoulli BRT models (sub-models) was fitted to each continent dataset using the R package 'gbm3' [220]. Each sub-model was trained to a separate bootstrap dataset randomly sampled with replacement from the complete occurrence/background dataset for that continent. The model algorithm hyper-parameters were set to the following values: cross-validation folds=10, tree complexity=3, learning

rate=0.0005, and bag fraction=0.5. Otherwise, I followed the procedure used in *Chapter 2* for mapping the distribution of the macaque reservoir species of *P. knowlesi* malaria, and further details on model fitting can be found there.

To generate the final prediction maps, a mean predicted value of environmental suitability was calculated across the model ensembles. Prior to inclusion in the YF disease model, each prediction was clipped to the joint species IUCN range maps for that continent. Finally, the predictions for each continent were joined to produce a single YF reservoir covariate data surface (Figure 4.4).

Reservoir model evaluation

To evaluate the predictive performance of each continent model, the AUC was calculated, which reflects the ability of the model to discriminate between occurrence and background records [108]. The overall statistic for each model was calculated as the mean of the AUCs for each sub-model in the model ensemble. The models demonstrated moderate to high predictive performance, with overall AUCs of 0.754 (SE ± 0.008), 0.806 (SE ± 0.006), and 0.771 (SE ± 0.012) for the African, Latin American and Asian species distribution models, respectively.

4.2.6 Modelling YF occurrences as an inhomogeneous Poisson point process

Species distribution modelling has been widely applied in epidemiology to produce predictive maps of reservoir, vector and disease distributions from reports of disease occurrence [41]. Previous applications of SDM for disease mapping have typically used background or pseudo-absence points to represent areas where the disease has not been reported, and then model the presence/background label for each point as a draw from a Bernoulli likelihood, even though the data do not meet the assumptions of this model [221]. Recently, point process models have been proposed in the ecological SDM literature as a more appropriate tool for SDM of occurrence-only data. These models directly consider the data as a set of point locations where a species has been recorded as occurring [221–223]. The point process framework therefore resolves a number of

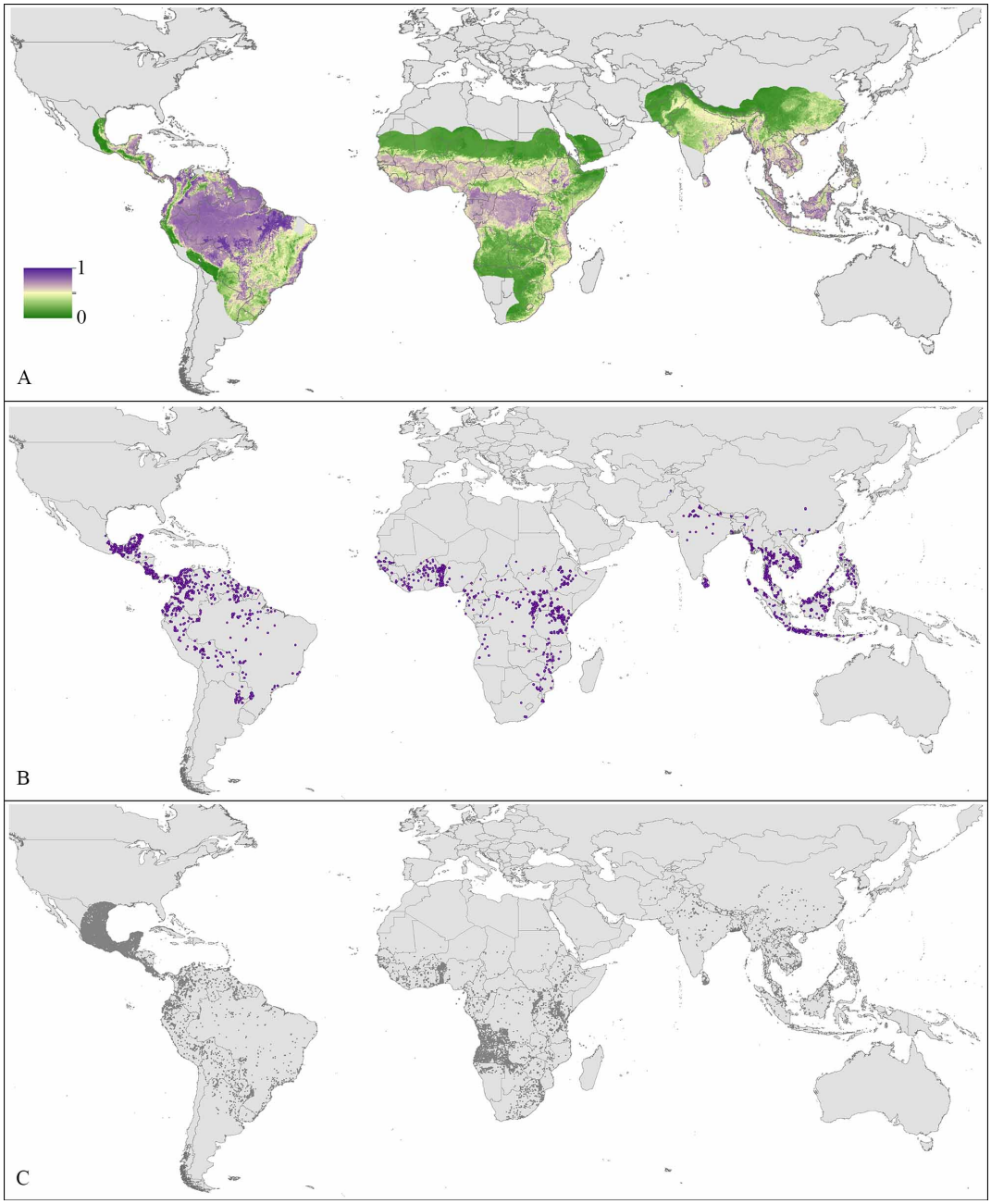


Figure 4.4: Distribution map of candidate non-human primate reservoir species for yellow fever virus (YFV). Composite habitat suitability maps for candidate non-human primate reservoir species of YFV for Latin America, Africa and Asia (**A**), the locations of occurrence data (**B**) and background data (**C**) used to generate each continent map.

important issues for occurrence-only SDM including interpretation, selection of pseudo-absence points, accounting for biases, and checking of model assumptions [221, 223].

Inhomogeneous Poisson point process models

The inhomogeneous Poisson point process is among the simplest and most well studied point processes that can be used to analyse occurrence-only data. Given a set of locations at which a species (or disease) has been observed in geographic region A, one can characterise the point process from which the locations are drawn via an intensity function over space $\lambda(s)$ – the expected number of observations per unit area. Conditional on this intensity function, the locations of the points are assumed to be independent of each other, and the total number of points in A is thus a Poisson random variable with mean:

$$\Lambda = \int_A \lambda(s) ds \tag{4.1}$$

One may also assume that the intensity surface of the point process $\lambda(s)$ varies spatially and *via* some relationship with spatially varying environmental conditions $x(s)$. The expected number of points per unit area can then be modelled as a function of covariates measured throughout the study region, A, to form a Poisson point process model (hereafter PPM) [224].

The most common approach to fitting a PPM is to maximise its likelihood function. Evaluating this likelihood requires calculation of the integral of the intensity function over A —the expected number of occurrence points. In general, the likelihood cannot be calculated directly, but can be approximated efficiently by evaluating a weighted Poisson likelihood at a set of quadrature points in space. These quadrature points are superficially similar to background points in a presence-background SDM analysis, but have clear mathematical interpretation under the PPM framework. Hence prior to model fitting, one needs to decide how quadrature points will be chosen, and how to calculate the quadrature weights at each quadrature and presence point. Frameworks for selecting quadrature points, along with examples, are provided in a number of recent publications [170, 223, 225].

A useful implication of the PPM approach is that if it is assumed that the observed locations represent all disease events in the area, then the intensity $\lambda(s)$ represents

the expected abundance of events. Most of the time this is not the case as not all occurrences of disease are reported, and the fraction reported is rarely known. However in these cases the intensity can instead be interpreted as being directly proportional to abundance, provided any spatial variation in sampling effort has been taken into account (*e.g.*, *via* offsets or covariates included in the model). This explicit link with abundance enables integration of distribution modelling with epidemiological models that consider immunity and metrics of disease dynamics.

Model description

A PPM was fitted to the dataset of YF disease reports using a BRT model to characterise the intensity function. By explicitly accounting for human population density, vaccination rate, and variable reporting effort, this model estimated the number of apparent YFV infections per person to within a constant of proportionality at each 5×5 km grid square in the study region.

It was assumed that the number of YF disease reports in any given area a within the study region follows a Poisson distribution:

$$N_A^+ \sim Poisson(\Lambda_A) \tag{4.2}$$

It was further assumed that the set of locations (points) arose from an intensity function $\lambda(s)$ comprising the product of three independent processes: the number of apparent infections per person, the number of individuals susceptible to infection and the proportion of infections reported:

$$\Lambda_A = \int_A \lambda(s)ds = \int_A I(s)P(s)R(s)ds \tag{4.3}$$

Where $I(s)$ gives the expected number per person of reported apparent infections over the study period at location s, $P(s)$ gives the susceptible population density at point location s, and $R(s)$ gives the number of reports per infection (reporting rate) at location s. The aim was to infer the spatial process $I(s)$ governing incidence, and so I sought to specify $P(s)$ and $R(s)$ *a priori*.

The population density of susceptible individuals $P(s)$ was directly estimated as the product of the overall population density and the fraction of the total population who had not received a vaccine against YFV averaged over the study period. The model does not take into account the proportion of the population with immunity from natural infections. It was also assumed that the YF vaccine provides lifelong 100% protection, which is supported by recent WHO recommendations [128].

Whilst the absolute reporting rate process $R(s)$ could not be estimated reliably, spatial variation in reporting rates of disease occurrence can be inferred from the distribution of reports generally [55, 224]. Therefore, $R(s)$ was factorised to $kr(s)$, where $r(s)$ is a probability distribution giving the relative reporting rate (and integrating to 1 over the study area), and k is a constant of proportionality which cannot be estimated *a priori*. Rather than inferring $r(s)$ directly, I followed Philips and colleagues [55] in collating a set z of n coordinates giving the locations of reports of human infectious diseases and treated these as random samples from $r(s)$. Specifically z were drawn from a database of locations within the study area where at least one case of human infectious disease has been reported. This approach assumes that YF and all the diseases in this database are subject to a common spatial variation in reporting rate. Further details on the assembly of the database of disease reports are provided in Section 4.2.7 below.

This sample from $r(s)$ enables Eq. 4.3 to be represented *via* a Monte Carlo approximation during model fitting. Writing the Monte Carlo approximation to a function f over a probability distribution π as:

$$\int_A f(s)\pi(s)ds \approx \frac{1}{n}\sum_i^n f(z_i); z_i \sim \pi(s) \tag{4.4}$$

and substituting $f(s) = kI(s)P(s)$ and $\pi(s) = r(s)$, gives:

$$\int_A I(s)P(s)R(s)ds \approx \frac{1}{n}\sum_i^n kI(z_i)P(z_i) \tag{4.5}$$

A PPM was fitted to the point pattern of YF reports, with a BRT model for the intensity function $\lambda(s)$. The PPM likelihood was approximated *via* the Monte Carlo sum above, using the Poisson likelihood and the set of locations z as integration points, following

the approach previously proposed for modelling PPMs in generalised linear modelling software using numerical quadrature [170, 223, 225]. The value of the susceptible population density process $P(z_i)$ was provided for each quadrature point as an offset value, leaving the BRT to estimate only the function $kI(s)$. Equation 4.3 assumes a linear relationship between the number of YFV infections and the number of susceptible individuals, thus the values of $I(s)$ predicted by the BRT model represent the incidence of apparent infections in susceptible individuals over the study time period, to within an unknown constant of proportionality k.

Model fitting

An ensemble of PPM BRT models were fitted using the R package 'gbm3' [220]. Rather than exclusively using synoptic (averaged across time) covariate values for each of the occurrence locations irrespective of the occurrence date, annual land cover data surfaces from 2001 to 2012 were used (data were not available for years outside this time period). Occurrence data collected between 2001 and 2012 were matched with covariate values for the relevant year; 66% of data points fell within this time period. Covariate values for 2001 were used for occurrence data prior to this date and covariate values for 2012 were used for post-2012 data. Final predictions were made to the most contemporary covariate values available. Land cover data were not available to match with occurrence data reported prior to 2001, and I was therefore unable to account for changes to the landscape (such as from urbanisation, agriculture and deforestation) that occurred from 1970 to 2000.

To increase the robustness of model predictions and quantify model uncertainty, I fitted an ensemble of 100 BRT models (sub-models), each trained to a separate bootstrap dataset. A spatial block bootstrapping procedure was applied where the study space was stratified into 160 regular grid squares (spatial blocks), and each bootstrap dataset consisted of all occurrence data and model integration points falling within a set of blocks, randomly sampled with replacement from the complete set. Each cross-validation fold consisted of data and integration points falling within a non-contiguous sample of spatial blocks, generated for the bootstrapping procedure. A Poisson likelihood was used

to approximate the PPM likelihood (specifying the Poisson density as a loss function) and integration weights were applied to each model integration point, equalling the estimated number of individuals susceptible to YF infection at that location, which were incorporated in the model as a logarithmic offset (passed *via* the offset algorithm). The algorithm hyper-parameters were set to the following values: cross-validation folds=10, tree complexity=3, learning rate=0.005, and bag fraction=0.75. The model was restricted so that YF transmission had a monotonic positive relationship with both habitat suitability and temperature suitability for *Ae. aegypti*. To incorporate uncertainty in the precise location of infection for polygon occurrence records, each bootstrap dataset randomly selected a 5×5 km pixel within each polygon, biased towards human population density, since more populous areas have a greater probability of reporting at least one case.

To generate the final prediction surface, a mean predicted value was calculated across the 100 sub-models for each 5×5 km pixel across the risk zones. Standard deviation values for each pixel were also calculated across the model ensemble to provide an estimate of model uncertainty (Appendix C.7).

4.2.7 Selection of model integration points

Eight thousand model integration points were drawn from a database of reports of human infectious diseases. A plot of these points is included in Appendix C.4. The aim of this dataset was to reflect the spatial bias in reporting rates of apparent YFV infections. A number of approaches for estimating spatial bias in occurrence data are described in the ecological literature, where datasets are similarly subject to spatial bias in survey effort [55]. As discussed and applied throughout this thesis, one approach is to use occurrence records for a broad biological group (such as birds or mosquitoes) to represent overall survey effort. Multiple factors are likely to impact disease reporting, including the strength of surveillance systems, treatment seeking behaviour, availability of accurate diagnostics, accessibility to healthcare facilities, and disease prevalence [226]. The most efficient way of encompassing these multiple issues in the integration dataset was to include reports from a broad range of human infectious diseases.

One drawback of this approach is that certain locations will experience higher rates of disease in general than others. In other words, the distribution of the integration points will actually represent a combination of spatial variation in reporting rates (what I want to incorporate in the model) and spatial variation in the number of diseases that could be reported (what I do not want to incorporate in the model). However, since no global-scale, spatially-resolved maps of disease reporting rates are available, I believe this approach is the best practical solution to accounting for these biases.

Reports of human infectious diseases were collated from four publicly available datasets: 1) for falciparum and vivax malaria [227], 2) the leishmaniases [37], 3) dengue [36], and 4) for numerous diseases (full list below) *via* the web-based system HealthMap, which automatically collates infectious disease reports from informal online data sources [22, 24]. The HealthMap records included disease reports from 2006 to 2014. This yielded 13,387 geo-referenced reports from within the study region, after removing any reports falling within the same 5×5 km pixel. To avoid generating a set of model integration points biased towards any one disease, sampling was biased towards diseases with lower counts and diseases with less than 20 reports were grouped before sampling.

Reports of the following diseases were collated *via* HealthMap: aeromonas and marine vibrio infection, anthrax, ascariasis, aspergillosis, avian influenza virus subtype H5N1, Bolivian haemorrhagic fever, botulism, brucellosis, campylobacteriosis, chikungunya, chlamydia infections, cholera, clostridium difficile colitis, common cold, Crimean-Congo haemorrhagic fever, cryptosporidiosis, cysticercosis, dengue, diphtheria, dracunculiasis (Guinea worm), Eastern equine encephalitis, Ebola virus disease, enterovirus infection, Escherichia coli diarrhoea, filariasis (bancroftian), giardiasis, glanders, hepatitis A, hepatitis B, hepatitis C, hepatitis E, histoplasmosis, HIV/AIDS, hookworm, infectious mononucleosis or EBV infection, influenza, Lassa fever, legionellosis, leishmaniasis (cutaneous), leishmaniasis (visceral), leprosy, leptospirosis, listeriosis, Marburg virus disease, measles, melioidosis, meningitis (viral), monkey pox, mumps, onchocerciasis, ornithosis, *P. falciparum* malaria, *P. malariae* malaria, *P. ovale* malaria, *P. vivax* malaria, parvovirus B19 infection, pediculosis, pertussis, plague, poliomyelitis, Q fever, rabies, respiratory syncytial virus infection, Rift Valley fever, Rocky Mountain spotted fever, rotavirus infection, rubella,

salmonellosis, severe acute respiratory syndrome, scabies, scarlet fever, schistosomiasis (haematobium), schistosomiasis (mansoni), shigellosis, St Louis encephalitis, syphilis, taeniasis, toxocariasis, toxoplasmosis, trachoma, trichinosis, trichuriasis, trypanosomiasis (African), trypanosomiasis (American), tuberculosis, typhoid and enteric fever, varicella, Venezuelan equine encephalitis, West Nile fever, and Zika.

4.2.8 Model validation

The model's predictive performance was evaluated using spatially stratified ten-fold cross-validation and calculating the predictive deviance of the model; the mismatch between the predicted number of occurrence records and the number observed. Specifically, McFadden's pseudo-R squared was calculated, which is the ratio of the log likelihood of the full model (with covariates) and the log likelihood of the null model (without covariates). The ratio of the likelihoods indicates the level of improvement offered by the full model over the null (or intercept) model [228]. This ratio was calculated for each cross-validation fold model of each bootstrap sub-model and averaged across folds and sub-models, resulting in a bootstrapped estimate of the model's predictive capacity.

4.2.9 Model calibration

The model estimate of the relative number of YF cases was calibrated using the average annual number of cases estimated by GBD [164] to obtain an estimate of the annual number of YF cases per grid cell, *i.e.* a continental-scale estimate of incidence was used to estimate the average case detection rate, which is not identifiable from presence-only data alone. The GBD mean estimates (as well as lower and upper bounds) of national incidence were available for every five years from 1990 to 2015. The average annual number of YF cases predicted by GBD for each country over this time period was calculated. Using the map of relative annual YF cases (based on the average number of individuals susceptible to YFV infection over the study time period), the total number of cases estimated by the model was summed for each country. Three separate generalised linear models were then fitted to each of the three GBD estimated case numbers (mean, lower, and upper), with the model estimated case numbers included as a logarithmic offset, and continent included as a fixed effect. Using the fitted model, calibration

factors were predicted for each continent, 155.15 for Africa (lower bound=44.27, upper bound=405.99), and 38.95 for South America (lower bound=10.32, upper bound=108.20). The mean calibration factors were then applied to the model output.

This method assumes a constant case detection rate within continent and over time. However, one might expect that reporting rates have improved over time, and were therefore lower in the 1970s and 1980s. Since the calibration factors were calculated using data from 1990 to 2015, and the average rate of apparent YF infections was calculated using data from 1970 to 2016, incidence may be underestimated due to inadequate correction.

4.2.10 Cases averted by vaccination

The average annual number of YF cases per pixel was calculated using estimated vaccination coverage rates in 2016, based on the three vaccination targeting scenarios for historical campaigns described in *Chapter 3*: 1) *untargeted*, *biased* (conservative), 2) *untargeted*, *unbiased* and, 3) *targeted* (optimistic). To estimate the average annual number of YF cases averted within risk zones by 2016 vaccination levels, the outputs of conservative and optimistic cases per pixel were compared to a fourth scenario assuming 100% of the population in 2016 were susceptible to YFV infection (*i.e.* vaccination coverage rates assumed to be zero). Since the model estimates the relative risk of YF infection from 1970 to 2016, and therefore assumes a constant rate of YF infection through time, these values represent the cases averted by 2016 vaccination levels given the average annual infection rate, not the specific infection rate in 2016, and assuming constant exposure to a source of YFV infection.

4.3 Results

The model predicted substantial international and subnational spatial variation in individual infection risk for YF (Figure 4.5A). This risk can be interpreted as the rate at which susceptible individuals are expected to acquire a symptomatic YFV infection in a given location. The underlying risk for YF can therefore be high in locations with negligible population density or high vaccination coverage. Areas with highest predicted individual

infection risk within Latin America includes the Amazonian regions of Brazil (particularly Pará, Acre and Roraima States), Bolivia, Colombia, Ecuador, French Guiana, Guyana, Peru, Suriname, and Venezuela. Within the African risk zone, the highest areas of predicted apparent infection risk includes a number of locations within Burkina Faso (Cascades and Sud-Ouest), Cameroon (Centre, Est, Littoral, Sud-Ouest, and Nord), Côte d'Ivoire (Savanes, Zanzan, and Montagnes), DRC (Sankuru, Maï-Ndombe, Équateur, Maniema, Mongala, Tshuapa, Kwilu, Nord-Ubangi, Bas-Uele, Haut-Uele, Ituri, and Tshopo), Ghana (Northern, and Upper West), Liberia (Bong, Gbarpolu, Grand Gedeh, Lofa, and Nimba), Republic of Congo (Sangha), Sierra Leone (Eastern), Mali (Sikasso) and South Sudan (Jonglei). Model uncertainty in spatial predictions of individual infection risk for YF is shown in Appendix C.7; there is low uncertainty in predictions in most areas.

Validation statistics for the model ensemble indicated that model performance was high, with 92% of out-of-sample deviance explained by the model. The main predictors for YFV infection risk were the following: proportional cover of land classified as urban and built-up, temperature suitability for *Ae. aegypti*, habitat suitability for suspected non-human primate reservoir species, and enhanced vegetation index. Marginal effect plots for each of these covariates are included in Appendix C.6.

Considered together, the maps of individual apparent infection risk (Figure 4.5A), of average annual incidence (Figure 4.5B) and of average annual cases (Figure 4.5C), highlight the impact of vaccination on YF incidence. For example, populations of Brazil, Bolivia, Venezuela, Guinea, Cameroon, and Togo have high predicted rates of underlying risk of YF (Figure 4.5A), but low predicted annual incidence (Figure 4.5B), based on vaccination coverage levels achieved by 2016. Conversely, they also indicate that many populous regions, with patchy or limited vaccination coverage, have a high exposure risk to YFV infection and are poorly protected against transmission. For example, areas with the highest predicted average annual cases included large parts of the DRC (Sankuru, Maï-Ndombe, Équateur, Mongala, Tshuapa. Kwilu, Nord-Ubangi, Sud-Ubangi, Bas-Uele, Haut-Uele, and Tshopo), Nigeria (South-West, South-East, South-South, and North-Central Zones), and South Sudan (Jonglei, Eastern Nile, Western Nile, Terekeka, Yei River, and Jubek), as well as smaller areas within Côte d'Ivoire (Montagnes), Sierra Leone (Eastern),

Ghana (Ashanti and Northern), Mali (Sikasso), Ethiopia (Amhara and S.N.N.P.) and Uganda (Adjumani). For all these locations, vaccination coverage in 2016 was estimated to be less than the 80% recommend by the WHO to prevent outbreaks [122], and for the majority of them, less than 40%.

Using the total average annual cases estimated by the model, it was calculated that globally between 94,336 and 118,500 cases of YF are averted each year by vaccination coverage rates achieved in 2016 (based on conservative and optimistic vaccination coverage scenarios, respectively). This represents 33% to 42% of total predicted cases if vaccination coverage was zero across risk zones. The global totals represent between 84,385 and 99,840 cases averted on the African continent (33% to 39% of total predicted cases) and between 9,951 and 18,660 cases within Latin America (36% to 68% of total predicted cases).

Outside contemporary risk zones, the model predicted high receptivity to YFV transmission across much of Southeast Asia, where YF has never been reported, including large areas of Cambodia, India, Indonesia, Laos, Malaysia, Myanmar, Papua New Guinea, the Philippines, Thailand, and Vietnam (Figure 4.6A). High-receptivity is also predicted in several countries within Central and South America, contiguous with risk zones. Historically, YF has been reported in many of these countries, such as Costa Rica, Guatemala, Honduras, Nicaragua, and Mexico (Figure 4.6B) [229, 230]. Interestingly, low receptivity to YFV transmission was predicted in several locations where *Ae. aegypti* is known to be present, for example southern China (Figure 4.6C), and southern United States.

4.4 Discussion

The high-resolution maps of YF risk and incidence produced in this chapter highlight the success of vaccination in reducing incidence in a number of high-risk regions, including Brazil, Cameroon and Togo (Figure 4.5). The maps also identify areas with high-predicted average annual case numbers, where vaccination coverage in 2016 was less than the recommended threshold to prevent outbreaks, such as large parts of Nigeria, DRC, and South Sudan. These maps provide an evidence base to prioritise areas for vaccination and vector control programmes in current risk zones. Receptivity to YFV transmission

in areas outside risk zones was also mapped (Figure 4.6), including areas where YF has been controlled, such as Central America, or has never persisted, such as southern Asia. My findings highlight areas where public health authorities should be most vigilant for potential spread or importation events. Furthermore, the areas of high receptivity near the edge of the current risk zone on the southeast coast of Brazil (States of Bahia, Minas Gerais, São Paulo, Esparíto Santa, and Rio de Janeiro), reflect a need to geographically expand the existing risk limits because locally acquired cases have recently been reported in this area [231].

Rogers and colleagues [125] used discriminant analysis techniques to predict areas of climatic suitability for YF within Asia, including eastern Thailand and parts of Malaysia and Indonesia. Building on their analysis, my model predicts more extensive receptivity to YFV transmission within these countries, as well as other parts of the continent. I have extended the work of Rogers and colleagues by fitting, to the most recent occurrence data, a model that has been demonstrated to perform well, accounts for bias in disease reporting rates, and incorporates vaccination coverage rates. Furthermore, my model includes mechanistic knowledge of habitat suitability for vectors, a range of covariate data on land cover, and distribution maps for reservoir and vector species. This work is the first example of a disease niche modelling approach that incorporates vaccination coverage rates. My results for Africa (33–39% of cases averted from 1970 to 2016) add to a previous analysis that estimated that vaccination campaigns averted the number of cases by 22% to 31% from 1987 to 2011 [129]. The work in this chapter includes more recent (and historical) occurrence data, and overall highlights the differences between Latin America and Africa.

Predictive distributions of *Ae. aegypti* and candidate non-human primate hosts were included in the model, because countries where both exist are thought to be most vulnerable to the introduction and establishment of YFV [6]. Importation of the virus by an infected traveller could initiate an urban epidemic and, if suitable vector and reservoir species are present, subsequently trigger an enzootic transmission cycle, leading to a long-term infection risk for the local population [232]. This occurred when YFV was introduced into the Americas during the slave trade era, with the first epidemic reported in 1648 in Yucatán, Mexico [232]. The present ease and frequency of international travel,

compounded by poor enforcement of travel vaccination requirements, low levels of vaccination coverage in many at-risk regions, and existing vaccine shortages, raises serious concerns of YFV introduction into naive human populations, particularly in southern Asia [160].

Yellow fever has both urban and sylvatic transmission cycles, the spatial and temporal distribution of cases fluctuates considerably. The periodicity of YF epidemic activity is driven in part by the rapid depletion and slow replenishment of susceptible hosts, as illustrated by upsurges at irregular intervals in parts of West and East Africa. My model did not attempt to capture temporal and spatial spikes in YF cases; the focus was on predicting spatial variation in the underlying risk of apparent infection. Model outputs represent the time period from 1970 through to 2016, averaged over the large fluctuations that occur. Thus my results may smooth over important secular trends across time, and the estimated incidence for a particular location will be different from the 2016 incidence, or a five-year average. Indeed, the five most recently reported outbreaks of YF began in areas of variable predicted risk. Additionally, since all reports of symptomatic human infection with YFV were included in the model, regardless of whether infection was the result of a sylvatic, intermediate, or urban transmission cycle, the model predicts apparent infection risk from any transmission cycle. The BRT model is capable of capturing different relationships in the data arising from these distinct transmission cycles, with the given covariates.

My model predicts receptivity to YFV transmission based on the relation between locations where YF has been reported from 1970 to 2016, and the values of environmental and biological covariates at those locations. For this reason, predictions of high receptivity in southeast Asia should be interpreted with caution. Potential variables that distinguish current risk zones in Africa and the Americas from *Ae. aegypti*-inhabited areas of southeast Asia might be missing from my analysis. Indeed, most theories that have been used to explain the absence of YF in Asia involve biological factors rather than climatic or environmental ones. These factors include cross-protection from other flaviviruses, lower competence of local populations of *Ae. aegypti*, and competition between other flaviviruses within mosquito cells [5, 233], none of which are included

in my model. The strength of the BRT approach is in its predictive power and ability to fit complex non-linear functions, with the given covariates. It is not possible to use this type of analysis to identify causal associations between the covariates and suitability for disease transmission. It would not be appropriate to make further inferences as to why the pattern of high receptivity predicted by my model is more contracted than the distribution of YF's main urban vector, *Ae. aegypti*, or to further speculate on reasons for the disease's absence from Asia to date. This work, however, highlights areas for future improvement in our understanding of YFV ecology.

This chapter involved multiple stages of analysis, each containing potential sources of bias and uncertainty, much of which was difficult to account for or quantify. Each stage of the analytical process involved fitting a model to an independent, fixed dataset. Since each of these models was fitted to fixed data, not the output of previous models, it was not meaningful to propagate uncertainty through these steps. However, there were some steps where it was important to account for or report uncertainties. In fitting the model of risk for apparent YFV infection, uncertainty in the spatial locations of reported cases was propagated through the model *via* Monte Carlo simulation. Similarly, uncertainty in the final step of the analytical process calibrating spatial predictions of yellow fever risk against GBD 2015 estimates of annual cases was estimated and reflected in reported results.

High quality spatial data on YF are lacking, largely due to diagnostic complexity and limitations of health care systems in many affected countries. The occurrence dataset assembled for this chapter included human YFV infections confirmed *via* both genetic and serological diagnostic methods. As a result of cross-reactivity among flaviviruses, the effectiveness of serological diagnostic techniques is limited, particularly when considering historical records [234]. Additionally, it was assumed that estimates of vaccination coverage, combined with data on population size, were proportional to the number of people susceptible to YFV infection. Translating maps of unvaccinated individuals into maps of susceptible people is complicated by the acquisition of immunity *via* natural infection, which is difficult to measure. Modelling efforts will be improved as the volume and quality of geographical data on YF increases, ideally using diagnostics that distinguish

past YF infection from vaccination, and from other flavivirus infections. For now, the work presented in this chapter provides the best evidence base for identifying populations most vulnerable to YF.

4.5 Conclusions

To develop the most cost-effective vaccination strategies that prevent outbreaks and minimise adverse events, vaccination policy makers require a clear understanding of geographical disease risk. Within risk zones, my model improves understanding of geographical variation in YF risk and comes at a time when current policies may require re-evaluation given the demonstrated inadequacy of emergency stocks and surge capacity of vaccine manufacture to meet the needs of the recent outbreak in Angola [159]. Since the vector of urban YFV transmission also transmits dengue, chikungunya, and Zika viruses, as well as other human pathogens, the work presented in this chapter can also inform integrated disease control strategies. Beyond risk zones, the map of receptivity identifies areas where governments should consider the need for requiring that travellers to and from YF endemic countries are vaccinated, under the framework of the International Health Regulations, as well as assessing their capabilities for early detection and intervention. If local YFV transmission was to occur in these locations, the map additionally provides an evidence base for predicting and limiting its spread.

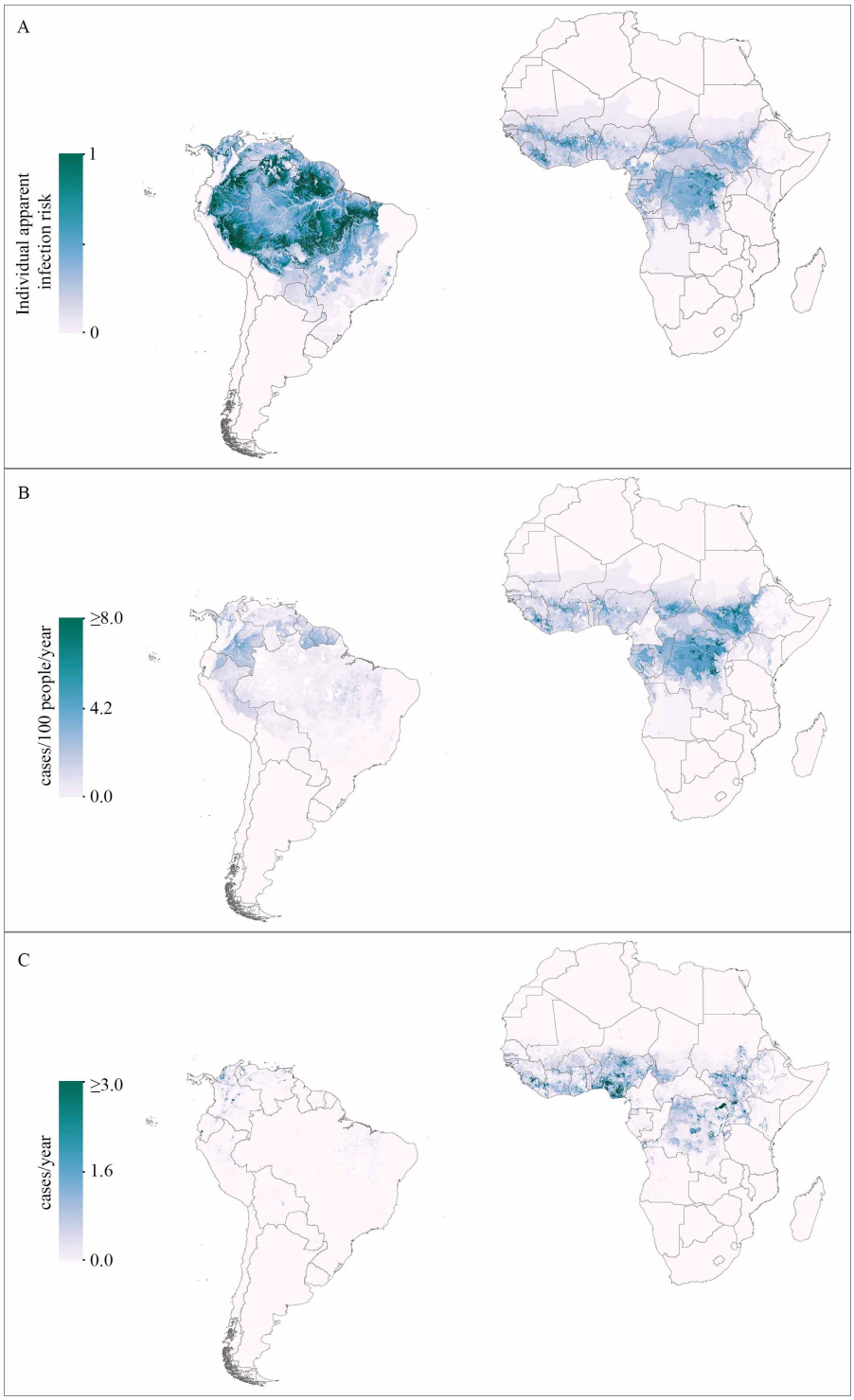


Figure 4.5: Predicted distribution of yellow fever (YF) within contemporary risk zones. A. Predicted spatial variation in individual risk of YF virus infection. **B.** Predicted average annual YF incidence (99.8% of grid squares within the disease's range were predicted to have fewer than 8 cases per 100 people per year, the highest predicted value was 20 cases per 100 people per year). **C.** Average annual YF case numbers (99.5% of grid squares within the disease's range were predicted to have fewer than 3 cases per year, the highest predicted value was 109 cases per year). Continental calibration factors have been applied to the latter two outputs, calculated from GBD estimates of national incidence averaged from 1990 to 2015. They also use vaccination coverage rates achieved in 2016. All predictions were restricted to areas within the contemporary risk zones defined by Jentes and colleagues [27] and were averaged over the 47-year study period from 1970 to 2016.

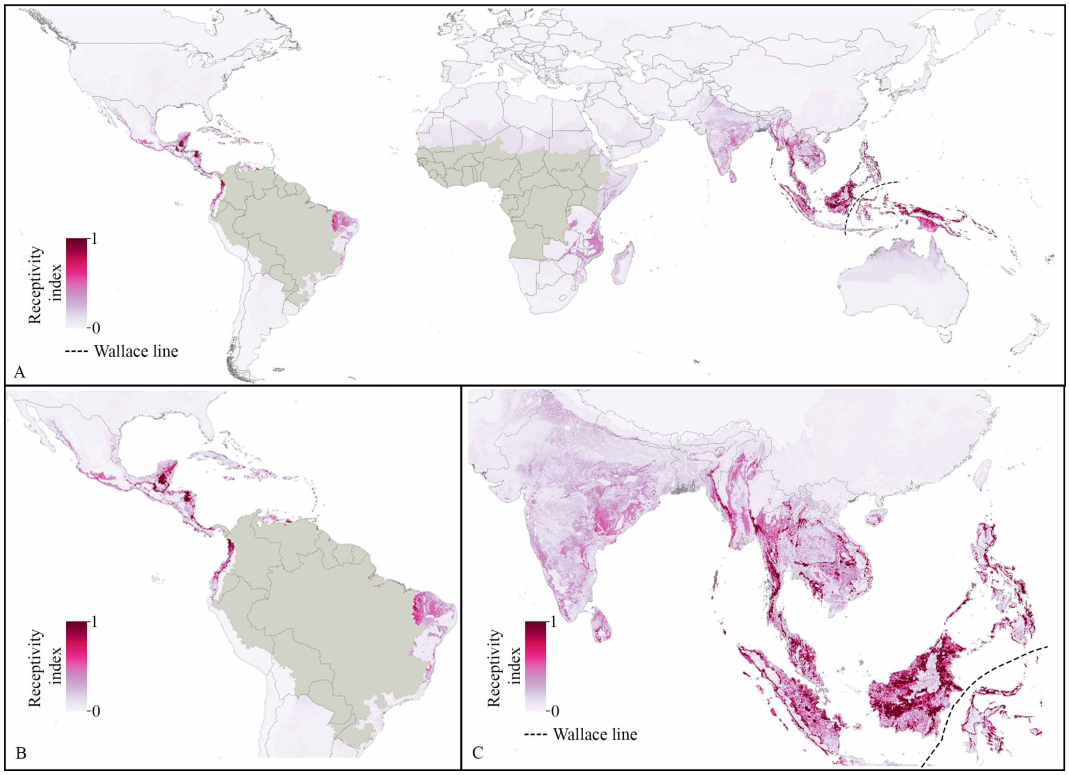


Figure 4.6: Predicted receptivity to yellow fever transmission outside existing risk zones. Receptivity to YF transmission in areas outside the contemporary risk zones, as defined by Jentes and colleagues [27], and shown in light brown. No species known to be potential hosts of YFV persist in areas east of the Wallace line (faunal boundary).

5
General discussion

The aims of this thesis were to investigate the geographical distribution of two vector-borne, zoonotic diseases of public health importance: *P. knowlesi* malaria and YF. To achieve these aims, there was a focus on the development of novel approaches for: mapping diseases whose distributions are impacted by multiple host and reservoir species, ameliorating spatial bias in disease reporting rates, and accounting for human vaccination coverage.

The following general discussion recapitulates the content and principal findings of this thesis, discusses some relevant methodological and policy issues, and draws overall conclusions.

5.1 Recapitulation

Chapter 2 investigated the spatial variation in risk of human *P. knowlesi* malaria across Southeast Asia. A standard niche modelling approach was used to map reservoir and vector species distributions, and a novel joint distribution model was used to predict disease risk, so that infection data could be leveraged from all three host species: humans, macaques and mosquitoes. Other modifications were made to previously described niche modelling approaches to allow for temporal changes in land cover to be incorporated, and improve handling of sampling bias and wide-area data. The final *P. knowlesi* malaria

map represents an initial evidence-base for identifying areas of human disease risk that should be prioritised for surveillance and future epidemiological investigations, particularly in the context of malaria elimination in the region.

Substantial outbreaks of YF in Angola and Brazil in the past two years, coupled with the global expansion of the range of *Ae. aegypti*, highlight a pressing need to assess current control strategies, and to improve understanding of YF's contemporary distribution and potential for spread into new areas. In *Chapter 3*, prior to mapping the distribution of YF risk (*Chapter 4*), population coverage of YF vaccination was estimated across all countries within YF risk zones, from 1970 to 2016. To account for uncertainty in how vaccination campaigns were targeted, three population coverage values were calculated to encompass alternative scenarios.

In order to account for vaccination coverage in the model of YF risk, the standard niche modelling approach applied in *Chapter 2* had to be developed further. Thus *Chapter 4* describes the development of a novel Poisson point process niche model, which is then used to predict YF infection risk in humans and demonstrates how vaccination coverage can be efficiently accounted for in disease niche models. Using previously published estimates of annual national case numbers, this model was used to map subnational variation in YF incidence across at-risk countries, as well as estimate the number of cases averted by vaccination globally. The fitted model of YF risk was also used to predict receptivity to YFV transmission outside current risk zones.

5.2 Methodological discussion

The mapping of YF and *P. knowlesi* malaria was complicated by multiple host and reservoir species, reporting rates biases and intervention coverage. In this section of the discussion, I will outline how each of these issues was approached in this thesis, the limitations of these approaches, and opportunities for further development.

5.2.1 Modelling diseases with multiple host and reservoir species

For spatial epidemiology to capture the full complexity of vector-borne, zoonotic disease transmission systems, especially those with multiple vector or reservoir species, it is

important to consider the existing and potential distributions of all component species. One approach is to add the distributions maps of one or more component species, alongside environmental and socioeconomic variables, as predictors in disease niche models. This has resulted in substantial improvements in model performance [235].

For the mapping of *P. knowlesi* malaria and YF performed in this thesis, reservoir and vector distributions were mapped first, and then included as potential predictors in disease models. I aimed to be comprehensive in the mapping of reservoir and vector species, which led to the inclusion of both confirmed and putative host species. Presently, insufficient evidence exists to confirm the relative roles of candidate YF reservoir species in YFV transmission. In addition, low volumes of occurrence data were available for many of these species. Thus for each continent within the study region, a multi-species SDM was fitted to occurrence data from a range of non-human primates with either serological or genetic evidence of YF infection. Similarly, as very low volumes of data were available for a number of species within the *Anopheles leucosphyrus* group, a multi-species model was used to generate a vector distribution surface, to supply to the *P. knowlesi* malaria risk model as a covariate. One advantage of the BRT algorithm is its ability to fit a single overall model to multiple distinct patterns arising in the data [51] (*e.g.* simultaneously fitting different relationships with the environment, for each species).

Habitat suitability for *M. nemestrina* (reservoir host of *P. knowlesi*) and suspected YF reservoir species, were among the most influential predictors for the *P. knowlesi* malaria and YF risk models, respectively. This suggests that reservoir/vector distributions were important disease predictors. These inferences are complicated, however, because the distributions of reservoir/vector predictor species are governed by similar environmental variables to the pathogen, which can induce multicollinearity within the disease model [236]. Thus it is difficult to disentangle the relative contributions of reservoir/vector suitability and environmental predictors of reservoir/vector suitability in explaining spatial variation in disease risk. One might expect that the relationship between environmental variables and reservoir/vector occurrence is a better predictor of disease risk than environmental variables alone, but a covariate selection exercise would be required to explore this further.

Since all reports of symptomatic YFV human infection were included in the YF model, regardless of whether infection was the result of a sylvatic, intermediate, or urban transmission cycle, the model predicted apparent infection risk from any transmission cycle. As forementioned, the BRT model is capable of capturing different relationships in the data arising from these distinct transmission cycles, with the given covariates. Data permitting, it would be interesting to map the distribution of YF risk from sylvatic, intermediate and urban transmission separately, for African risk zones. Human-vector-human transmission of YFV is likely to have a different niche to zoonotic transmission *i.e.* a different relationship with environmental, socioeconomic and biological factors. These maps would be of interest to policy-makers because the risk of human infection from each transmission cycle would elicit different control and preparedness strategies.

Purse and Golding note that for the most complete understanding of the relative roles of vector and reservoir species in transmission from distribution models, it is necessary to include data on the occurrence of all potential interacting species in the study region [41]. Indeed, the distributions of species that are not directly involved in the transmission of pathogens to humans can still be used to inform disease models. For example, pig density relative to cattle density is an important factor in Japanese encephalitis virus transmission ecology. The primary disease vector, *Culex tritaeniorhynchus*, blood feeds on cattle in preference to pigs and humans, and whilst pigs are amplifying hosts of the virus, cattle are unable to transmit the virus to other susceptible hosts [237]. In future models of *P. knowlesi* malaria risk, it is worth considering whether the distributions of *M. mulatta* (rhesus monkeys) and *Presbytis (Semnopithecus) entellus* should be included as potential predictors. It is thought that neither species can co-exist with *P. knowlesi* because they exhibit such high fatality rates when infected with the parasite [238, 239], and could therefore be used as negative indicators for a disease reservoir.

One challenge of incorporating reservoir/vector species distributions in models of disease risk, is distinguishing actual distributions of reservoir/vector species from their potential distributions. Niche models assume that a species is in equilibrium with its environment (*i.e.* that the species has had sufficient time to sample all potentially suitable habitat in the model training region). This assumption is clearly violated when

the species' range is expanding (*e.g.* newly invading pathogen or vector), or parts of its range are unoccupied (*e.g.* due to human intervention or a geographical barrier). In these cases, the species' potential distribution, rather than its actual distribution will be mapped [41, 110, 240].

For species that are not in equilibrium with their environment, outputs representing their actual distribution can be achieved by restricting predictions to areas within expert opinion or evidence consensus range maps. Both the existing and potential distributions of reservoir/vector species can provide information for public health decision-makers (for example, potential distributions are useful for forecasting disease distributions), but it is important to recognise what has been mapped. Otherwise, if potential reservoir/vector distributions are naively used to predict disease risk, the current limits of disease risk may be overestimated. This is especially true for vector-borne, zoonotic diseases without human-vector-human transmission cycles, since their geographic distributions are so intimately tied to reservoir/vector distributions. In this thesis, I used carefully constructed range maps for *P. knowlesi* reservoir and vector species and two versions of the disease map were produced: (1) the disease risk map, restricted to areas within the existing geographic range of the species required for zoonotic transmission, and (2) an unmasked version of the model output which represents disease risk for a hypothetical scenario where known reservoir and vectors species have expanded their ranges to reach all suitable environments.

One limitation of the reservoir/vector SDMs used as disease predictors in this thesis is that they quantify the relationship between reservoir/vector species' and their environments without considering other important factors that shape species distributions. However, ecologists and epidemiologists are increasingly developing new methods to incorporate other factors, such as dispersal constraints and biotic interactions [41].

When the goal is to predict the potential distribution of pathogens, vectors or reservoirs into news areas, a number of approaches have been proposed to deal with species dispersal constraints. These include parameterising models with occurrence data from early in an invasion or developing mechanistic models based on physiological responses of organisms to their environments. Furthermore, Golding and Purse have

recently published a niche modelling method that they found performed better than BRTs, and can be combined with mechanistic models, which may hold promise for incorporating these dispersal methods into SDMs [241]. These methods could be used to build on existing global maps of the invasive urban vector for YFV, *Ae. aegypti* [50]. Updated maps of the existing and potential distributions of *Ae. aegypti* would improve modelling efforts for a range of diseases of public health importance, since the vector also transmits dengue, chikungunya, and Zika viruses.

Vector-borne, zoonotic disease transmission systems are essentially sets of interacting species, and thus biological interactions also play an important role in shaping the distribution of pathogens. Methods have been developed to incorporate biotic interactions into SDMs, including joint species distribution models [242] and species interaction distribution models [236]. Extensions to SDM that consider biotic interactions and dispersal constraints are highly relevant to epidemiological applications, and will generate further insight into vector-borne, zoonotic disease patterns.

5.2.2 Spatial bias in disease reporting rates

The data available for mapping *P. knowlesi* malaria and YF were reports of disease occurrence, which are subject to spatial bias in reporting rates. Whilst this is an issue with all occurrence-only disease mapping, it is likely to be more pronounced for diseases such as *P. knowlesi* malaria and YF, where significant resources are required to accurately diagnose infection.

Plasmodium knowlesi infection is not routinely considered a possible cause of malaria within the ranges of its reservoir/vector species, with the exception of MBS. A disease model was therefore fitted using only data from MBS (the model training region), where reporting bias could be accounted for through the selection of background data. Additionally, a MESS surface was generated to identify regions where the model was required to extrapolate to novel environments not found within MBS, and therefore evaluate the appropriateness of inferring risk in those regions. To account for spatial bias in reporting of *P. knowlesi* infections within MBS, I followed the approach of Phillips and colleagues by selecting background records that reflected the same bias as the occurrence data.

Since the infection data contained records from reservoirs, vectors and human hosts, this process was tailored to account for specific biases within each host infection dataset.

For the YF risk model, disease reporting rate bias was accounted for through the selection of model integration points. These points were randomly drawn from a large global database of reports of human infectious diseases, which aimed to reflect the spatial bias in reporting rates generally. This approach assumes that YF and all other infectious diseases in the database are subject to a common spatial variation in reporting rate. In reality, reporting rates will likely vary between individual diseases or groups of similar diseases. Indeed, Meyer and colleagues recently found in their study of bias in global biodiversity datasets that the relative importance of bias-causing factors varied strongly across taxonomic groups [243].

Further opportunities for improvements in modelling of disease reporting rate bias may lie in better use of available information. For example, the database of infectious disease reports assembled for this thesis could be used to generate global-scale, spatially resolved maps of disease reporting rates. This could be achieved by fitting a joint PPM BRT to the disease reports, with disease features (such as pathogen taxonomy, symptoms, or diagnostic method) included as a large set of binary covariates, to produce a combined distribution of reporting effort for diseases with those features. For example, a predictive surface of reporting rate bias for PCR-diagnosed, viral infections presenting with fever, may improve geospatial models of YF risk.

This joint PPM BRT model could be further improved by incorporating variables thought to influence variation in reporting rates. For global biodiversity data, drivers of species survey effort have recently been explored in-depth, with region or species appeal, spatial distance to data-contributing institutions and locally available research funding, identified as key drivers. Additionally, drivers of bias have been explicitly incorporated into biodiversity methods, through the joint analysis of occurrence data and comparatively limited presence-absence data [224, 244]. Multiple factors are likely to influence disease reporting, including the strength of surveillance systems, treatment seeking behaviour, availability of accurate diagnostics, accessibility to healthcare facilities, and disease prevalence; but the relative importance of these factors has not yet been investigated

on a global scale, or across a range of diseases, or for different types of disease reports [226]. Nonetheless, variables such as urban accessibility, which has recently been comprehensively mapped [245], or other possible proxies to bias-influencing processes, could be supplied to the bias model as potential predictors.

For each occurrence dataset used in this thesis, potential biases were assessed, and methods were employed to ameliorate the impact of these biases on predictions. In most cases, this required sourcing a separate dataset to represent the distribution of survey effort/reporting. This process was time-consuming, and could be streamlined if explicit models of surveying/reporting bias existed, such as the one described above.

5.2.3 Accounting for public health interventions in niche models

Standard niche modelling frameworks do not provide an efficient method for dealing with the potential impact of public health interventions (such as vaccination coverage) on disease risk. This thesis describes the development and application of a novel PPM BRT model where population immunity was explicitly incorporated into the modelling procedure. In addition, this approach resolves a number of other issues inherent in occurrence-only SDM, including interpretation, selection of background points, accounting for biases, and checking of model assumptions [221, 223].

Standard applications of SDM for disease mapping typically supplement occurrence records with background (or pseudo-absence) points to represent areas where the disease has not been reported, and then use a Bernoulli likelihood, in spite of the data not meeting the assumptions of this model. Because of the arbitrary nature of these background points, a variety of approaches have been proposed for generating them, without a consensus amongst modellers on how they should be chosen [44], and these ad-hoc modelling decisions lack a solid mathematical justification [221]. Since PPMs directly consider the data as a set of point locations where a disease (or species) has been recorded as occurring [221, 222]. The point process framework therefore provides a solid underpinning for occurrence-only SDM.

For example, regarding the interpretation of model outputs, the index of suitability for zoonotic *P. knowlesi* malaria transmission, generated by a standard niche modelling

approach in *Chapter 2*, is a useful metric for prioritising locations for future *P. knowlesi* surveys. However, this index is not necessarily proportional to more useful epidemiological metrics, such as case incidence or infection prevalence, and can be difficult to interpret if such metrics are desired. As described in *Chapter 4*, a useful implication of the PPM approach is that if it is assumed that the observed locations represent all disease events in the area, then the intensity function represents the expected abundance of events. Most of the time this is not the case, as not all occurrences of disease are reported, and the fraction reported is rarely known. However in these cases the intensity can instead be interpreted as being directly proportional to abundance, provided any spatial variation in sampling effort has been taken into account (*e.g. via* offsets). This explicit link with abundance enables integration of distribution modelling with epidemiological models that consider immunity and metrics of disease dynamics. For the YF risk model, this meant that the raw model output could be calibrated to independent estimates of national YF incidence to map subnational variation in disease incidence, which increases the value of these maps to policy-makers. Point process distribution models could also be better integrated with other epidemiological models, for example, incorporating the solution to a susceptible-infectious-recovered-type model in place of the linear combination $I(s)P(s)$ used in *Chapter 4*.

Despite these advantages, and being a more statistically appropriate tool for niche modelling of occurrence-only data, PPMs have not yet been widely adopted by ecological or epidemiological niche modellers. This is most likely due to a lack of published practical applications (including methods for model evaluation) and no user-friendly software for constructing and evaluating models. The code used in this thesis will form part of an R software package, to allow ecologists and epidemiologist to easily fit PPM SDMs to their datasets, and encourage wider-use and further development of this methodology.

5.3 Translating maps into policy

Since the ultimate goal of infectious disease mapping is to reduce the impact of disease on human health, it is worthwhile considering how to maximise the design and interpretation of models for policy development.

The accuracy of disease models, and thus their utility, depends on the reliability of the input data, as well as the underlying assumptions used in model fitting. For many diseases of public health importance, high quality and spatially representative epidemiological data are not available. The paucity of data available for the two diseases studied in this thesis, constrained the type of analyses that could be conducted, and the maps and models presented here are by no means final. They do, however, represent the best currently available estimates of the distributions of *P. knowlesi* malaria and YF risk. These maps will iteratively improve as the volume and quality of data increases, and modelling techniques are developed further, especially for quantifying uncertainty and understanding of species' interactions and dispersal constraints.

Given the data constraints, it was important to clearly communicate to readers (and potentially public health decision-makers): how outputs should be interpreted, the model limitations, and highlight locations where improved data would be most informative. In the case of *P. knowlesi* malaria, the lack of geographical data on this disease was acknowledged throughout *Chapter 2*, and strongly shaped the design of my analysis, as well as any recommendations for researchers and decision-makers. The results were not represented as a definitive map of *P. knowlesi* malaria infection risk, but as an initial evidence-base to help prioritise locations for surveillance and future epidemiological investigations.

Despite advances in the predictive power and relevance of infectious disease models over recent decades, their integration into the public health decision-making process remains limited, with many risk maps remaining unused in scientific literature or reports. To help improve linkages between disease modellers and policy-makers, scientists have called for established standards for assessing and communicating model skill, more effective communication of model predictions, especially uncertainties [41, 246, 247], and the development of systems for operationalising models for repeated use. For example, the recently developed zoon R package [248] is an opportunity for niche modellers to generate reproducible, comparable SDM outputs. The code developed for fitting PPM SDMs in this thesis could be added as a zoon module.

Even outcomes of the most well designed, executed and communicated disease mapping exercises may not be adopted to support public health activities, particularly if modelling goals do not match specific public health needs [41], align with current political narratives [249], or are not delivered in a timely manner [250] (since the policy process tends to be very fast, especially in the context of outbreak response).

During this thesis, most stakeholder input was sought towards the end of the analysis phase, when perhaps they might have felt it was too late to suggest substantial changes. Ensuring that modelling goals match specific public health needs, would perhaps require improved consultation with policy-makers and other stakeholders, prior to model development [246]. Interaction with stakeholders and policy makers helps modellers to understand trade-offs in health priorities, political interests at stake in policy choices, and the optimal format for communicating results. To further increase the value of maps to decision-makers, modellers should also consider providing an interactive interface that allows decision-makers to vary assumptions or key variables [250]. For the YF vaccination coverage mapping, I uploaded the information necessary for stakeholders to re-calculate the number of people requiring YF vaccination based on any new vaccination coverage threshold, to an online data repository, but more user-friendly formats could be explored.

Several scientists have also called for collaboration with the social sciences in order to generate models more relevant to the social and political contexts in which policy is implemented [249, 251, 252]. For example, social and anthropological understanding of human-wildlife interactions and risk perceptions would provide deeper insight into the patterns of vector-borne, zoonotic disease transmission [251, 252]. Indeed, activities that increase human contact with non-human primates and sylvatic mosquitoes, are considered risk factors for YF and *P. knowlesi* malaria zoonotic transmission [187, 253, 254]. An anthropologist was involved in the *P. knowlesi* malaria mapping work conducted in this thesis, and I considered using cultural information about interactions between humans and macaques, but suitable data were not forthcoming.

Leach and Scones [249] further argue that models should not be understood as an objective, neutral science informing policy in a linear manner, but as having social and political norms and values that should be made explicit and aim at triangulation rather

than integration into policy decision making processes. They believe that a multiple model approach is essential to inform policy, so that diverse perspectives, assumptions, and sources of information can be considered.

Exploring and addressing these issues is an opportunity to advance both the science of infectious disease mapping and its utility in public health decision-making processes.

5.4 General conclusions

Using a combination of standard and PPM niche modelling approaches, this thesis provides estimates of the geographical distribution of two vector-borne, zoonotic diseases, *P. knowlesi* malaria and YF. These diseases are currently of substantial importance to global public health and, prior to the work presented in this thesis, knowledge of their existing and potential geographic distributions was lacking. Global estimates of YF vaccination coverage were also provided for all countries across YF risk zones, and were incorporated into the disease risk model. The disease maps produced in this thesis have contributed substantially to our understanding of the magnitude of these public health problems, as well as providing guidance to those working in disease research, surveillance and control.

The risk map of *P. knowlesi* malaria represents an initial evidence-base for identifying areas outside MBS that governments and researchers should prioritise for surveillance (with appropriate diagnostics) and for further epidemiological investigation. This is particularly relevant for identifying areas of potential residual malaria transmission in countries currently eliminating the human malarias.

My results for YF provide the evidence-base for targeting vaccination campaigns within risk zones, as well as emphasising their high effectiveness. The receptivity map highlights areas where public health authorities should be most vigilant for potential spread or importation events. This includes several countries in Asia, where the disease has never persisted, and its emergence could have devastating consequences for local populations. As highlighted in *Chapter 4*, governments in these regions should consider the need for travellers to or from countries endemic for yellow fever to be vaccinated, under the framework of the International Health Regulations, as well as assessing their capabilities for early detection and intervention. If local transmission was to occur in

these locations, the map additionally provides an evidence base for predicting and limiting its spread.

The YF risk model extends the work of others in identifying high environmental, climatic, and to a certain extent biological, suitability for YF transmission in Asia. These results, coupled with recent shortages in the global YF vaccine stockpile, and the global expansion of the distribution of *Ae. aegypti*, emphasises the vulnerability of southern Asian populations to potentially devastating YF outbreaks. Some epidemiologists believe that it is only a matter of time before local transmission of YF occurs on the Asian continent. Whilst it is not possible to use the BRT method to identify causal relationships between the covariates and disease risk, considering the set of environmental variables included in my model, and the generally accepted view that YFV has previously been imported to Asia, I would speculate that YF has never persisted in this region because of an unknown biotic factor, *i.e.* a key interaction between YFV and its host or vector that enables or prevents onward transmission. My work perhaps highlights that the key to understanding YF's absence from Asia is in improving our knowledge of YFV ecology, which should be a research priority.

In addition to the public health policy implications, this thesis provides a number of advances on existing approaches to geospatial mapping of risk for vector-borne, zoonotic pathogens. Using the PPM approach introduced in *Chapter 4*, one can explicitly account for vaccination coverage in disease niche models, incorporate variation in reporting rates, and generate predictions proportional to abundance or incidence. Extensions to this method that account for biological interactions, dispersal constraints and human behaviour could further increase the accuracy and interpretability of these model outputs. Due to data constraints, my predicted disease distributions are by no means final, but they represent the best currently available evidence base for identifying populations most vulnerable to *P. knowlesi* malaria and YF. Furthermore, the models will iteratively improve as the volume and quality of epidemiological data increases, and modelling techniques are developed further. For example, recent methods have emerged for combining machine learning (including BRT) with Gaussian process geostatistical approaches, which have been shown to outperform any individual method in terms of predictive power [255].

The maps resulting from this thesis will act as resources to improve the targeting and evaluation of disease surveillance and control strategies. The methods developed to cope with complex transmission systems, account for intervention coverage and ameliorate reporting rates biases, will be applicable to risk mapping for a range of diseases of public health importance. In addition to improving the accuracy of risk maps, modelling the impact of public health interventions, such as vaccination coverage, enables inference about the past and future impact of vaccination, maximising the utility of these maps for public health policy.

References

[1] Kassebaum, N. J. et al. “Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015”. *Lancet* 388.10053 (2016), pp. 1603–58.

[2] Anderson, R. and May, R. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.

[3] Ghosh, D. and Basu, A. “Japanese encephalitis — a pathological and clinical perspective”. *PLOS Negl Trop Dis* 3.9 (2009), e437.

[4] Solomon, T. et al. “West Nile encephalitis”. *BMJ* 326.7394 (2003), p. 865.

[5] Hanley, K. et al. “Fever versus fever: the role of host and vector susceptibility and interspecific competition in shaping the current and future distributions of the sylvatic cycles of dengue virus and yellow fever virus”. *Infect Genet Evol* 19 (2013), pp. 292–311.

[6] Reisen, W. K. “Landscape epidemiology of vector-borne diseases”. *Annu Rev Entomol* 55.1 (2009), pp. 461–83.

[7] Lambin, E. F. et al. “Pathogenic landscapes: Interactions between land, people, disease vectors, and their animal hosts”. *Int J Health Geogr* 9.1 (2010), p. 54.

[8] Estrada-Peña, A. et al. “Effects of environmental change on zoonotic disease risk: an ecological primer”. *Trends Parasitol* 30.4 (2014), pp. 205–14.

[9] Reperant, L. A. “Applying the theory of island biogeography to emerging pathogens: toward predicting the sources of future emerging zoonotic and vector-borne diseases”. *Vector Borne Zoonotic Dis* 10.2 (2010), pp. 105–10.

[10] Guerra, C. A. et al. “The limits and intensity of *Plasmodium falciparum* transmission: Implications for malaria control and elimination worldwide”. *PLOS Med* 5.2 (2008), e38.

[11] Pullan, R. L. and Brooker, S. J. “The global limits and population at risk of soil-transmitted helminth infections in 2010”. *Parasit Vectors* 5.1 (2012), p. 81.

[12] Bhatt, S. et al. “The global distribution and burden of dengue”. *Nature* 496.7446 (2013), pp. 504–7.

[13] Tatem, A. J. et al. “Ranking of elimination feasibility between malaria-endemic countries”. *Lancet* 376.9752 (2010), pp. 1579–91.

[14] Kilpatrick, A. M. et al. “Predicting pathogen introduction: West Nile virus spread to Galápagos”. *Conserv Biol* 20.4 (2006), pp. 1224–31.

[15] Nunes, M. R. T. et al. “Emergence and potential for spread of chikungunya virus in Brazil”. *BMC Med* 13.1 (2015), p. 102.

[16] Aströöm, C. et al. “Potential distribution of dengue fever under scenarios of climate change and economic development”. *EcoHealth* 9.4 (2012), pp. 448–54.

[17] Hales, S. et al. “Potential effect of population and climate changes on global distribution of dengue fever: an empirical model”. *Lancet* 360.9336 (2002), pp. 830–34.

[18] Williams, C. R. et al. “Bionomic response of *Aedes aegypti* to two future climate change scenarios in far north Queensland, Australia: implications for dengue outbreaks”. *Parasit Vectors* 7.1 (2014), p. 447.

[19] Rogers, D. J. et al. “Satellite imagery in the study and forecast of malaria”. *Nature* 415.6872 (2002), pp. 710–15.

[20] Tusting, L. S. et al. “Socioeconomic development as an intervention against malaria: a systematic review and meta-analysis”. *Lancet* 382.9896 (2013), pp. 963–972.

[21] WorldPop. *Gridded Population Distributions, 2015*. Available from: `http://www.worldpop.org.uk/`. Accessed: January, 2016.

[22] Freidfeld, C. et al. “Healthmap: global infectious disease monitoring through automated classification and visualization of Internet media reports”. *Med Inform Assoc* 15 (2008), pp. 150–7.

[23] Hay, S. I. et al. “Big data opportunities for global infectious disease surveillance”. *PLOS Med* 10.4 (2013), e1001413.

[24] Patching, H. M. M. et al. “A supervised learning process to validate online disease reports for use in predictive models”. *Big Data* 3.4 (2015), pp. 230–7.

[25] Alvar, J. et al. “Leishmaniasis worldwide and global estimates of its incidence”. *PLOS ONE* 7.5 (2012), e35671.

[26] Brady, O. J. et al. “Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission”. *Parasit Vectors* 7.1 (2014), p. 338.

[27] Jentes, E. S. et al. “The revised global yellow fever risk map and recommendations for vaccination, 2010: consensus of the Informal WHO Working Group on Geographic Risk for Yellow Fever”. *Lancet Infect Dis* 11.8 (2011), 622–32.

[28] Reiner, R. C. et al. “A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010”. *J R Soc Interface* 10.81 (2013).

[29] Gething, P. W. et al. “Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*”. *Parasit Vectors* 4.1 (2011), p. 92.

[30] Weiss, D. J. et al. “Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach”. *Malar J* 14.1 (2015), p. 68.

[31] Kearney, M. and Porter, W. “Mechanistic niche modelling: combining physiological and spatial data to predict species’ ranges”. *Ecol Lett* 12.4 (2009), pp. 334–50.

[32] Gething, P. W. et al. “A new world malaria map: *Plasmodium falciparum* endemicity in 2010”. *Malar J* 10.1 (2011), p. 378.

[33] Kabatereine, N. B. et al. “Integrated prevalence mapping of schistosomiasis, soil-transmitted helminthiasis and malaria in lakeside and island communities in Lake Victoria, Uganda”. *Parasit Vectors* 4.1 (2011), p. 232.

[34] Hay, S. I. et al. “Global mapping of infectious disease”. *Philos Trans R Soc B Biol Sci* 368.1614 (2013).

[35] *ProMED Mail*. International Society for Infectious Diseases (2010). Available from: `https://www.promedmail.org/`.

[36] Messina, J. P. et al. “A global compendium of human dengue virus occurrence”. *Sci Data* 1 (2014), p. 140004.

[37] Pigott, D. M. et al. “Global database of leishmaniasis occurrence locations, 1960–2012”. *Sci Data* 1 (2014), p. 140036.

[38] McInerny, G. J. and Etienne, R. S. “Stitch the niche – a practical philosophy and visual schematic for the niche concept”. *J Biogeogr* 39.12 (2012), pp. 2103–11.

[39] Soberón, J. and Nakamura, M. “Niches and distributional areas: concepts, methods, and assumptions”. *Proc Natl Acad Sci* 106.Suppl 2 (2009), pp. 19644–50.

[40] Soberón, J. “Grinnellian and Eltonian niches and geographic distributions of species”. *Ecol Lett* 10.12 (2007), pp. 1115–23.

[41] Purse, B. and Golding, N. “Tracking the distribution and impacts of diseases with biological records and distribution modelling”. *Biol J Linnean Soc* 115 (2015), pp. 664–7.

[42] Soberón, J. and Peterson, A. “Interpretation of models of fundamental ecological niches and species’ distributional areas”. *Biodiversity Informatics* 2 (2005), pp. 1–10.

[43] Elith, J. et al. “Novel methods improve prediction of species’ distributions from occurrence data”. *Ecography* 29.2 (2006), pp. 129–151.

[44] Barbet-Massin, M. et al. “Selecting pseudo-absences for species distribution models: how, where and how many?” *Methods Ecol Evol* 3.2 (2012), pp. 327–38.

[45] Pigott, D. M. et al. “Global distribution maps of the leishmaniases”. *eLife* 3 (2014), e02851.

[46] Pigott, D. M. et al. “Mapping the zoonotic niche of Ebola virus disease in Africa”. *eLife* 3 (2014), e04395.

[47] Sinka, M. E. et al. “The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis”. *Parasit Vectors* 4.1 (2011), p. 89.

[48] Sinka, M. E. et al. “The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis”. *Parasit Vectors* 3.1 (2010), p. 117.

[49] Sinka, M. E. et al. “The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis”. *Parasit Vectors* 3.1 (2010), p. 72.

[50] Kraemer, M. U. G. et al. “The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*”. *eLife* 4 (2015), e08347.

[51] Elith, J., Leathwick, J. R., and Hastie, T. “A working guide to boosted regression trees”. *J Anim Ecol* 77.4 (2008), pp. 802–13.

[52] Han, B. and Vrielink, E. *The Algorithm That’s Hunting Ebola*. Available from: `https://spectrum.ieee.org/biomedical/diagnostics/the-algorithm-thats-hunting-ebola`. Accessed: March, 2018.

[53] Schulman, L., Toivonen, T., and Ruokolainen, K. “Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation”. *J Biogeogr* 34.8 (2007), pp. 1388–99.

[54] Hertzog, L. R., Besnard, A., and Jay-Robert, P. “Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modeling”. *Divers Distrib* 20.12 (2014), pp. 1403–13.

[55] Phillips, S. J. et al. “Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data”. *Ecol Appl* 19.1 (2009), pp. 181–97.

[56] Fourcade, Y. et al. “Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias”. *PLOS ONE* 9.5 (2014), e97122.

[57] Kramer-Schadt, S. et al. “The importance of correcting for sampling bias in MaxEnt species distribution models”. *Divers Distrib* 19.11 (2013), pp. 1366–79.

[58] Varela, S. et al. “Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models”. *Ecography* 37.11 (2014), pp. 1084–91.

[59] Escobar, L. E. et al. “Ecology and geography of transmission of two bat-borne rabies lineages in Chile”. *PLOS Negl Trop Dis* 7.12 (2013), e2577.

[60] Blackburn, J. K. et al. “*Bacillus anthracis* diversity and geographic potential across Nigeria, Cameroon and Chad: further support of a novel West African lineage”. *PLOS Negl Trop Dis* 9.8 (2015), e0003931.

[61] Blackburn, J., McNyset KM Curtis, A., and Hugh-Jones, M. “Modeling the geographic distribution of *Bacillus anthracis*, the causative agent of anthrax disease, for the contiguous United States using predictive ecological [corrected] niche modeling”. *Am J Trop Med Hyg* 77.6 (2007), pp. 1103–10.

[62] Levine, R. S. et al. “Ecological niche and geographic distribution of human monkeypox in Africa”. *PLOS ONE* 2.1 (2007), e176.

[63] Mak, S. et al. “Ecological niche modeling of *Cryptococcus gattii* in British Columbia, Canada”. *Environ Health Perspect* 118.5 (2010), pp. 653–8.

[64] Slater, H. and Michael, E. “Predicting the current and future potential distributions of lymphatic filariasis in Africa using maximum entropy ecological niche modelling”. *PLOS ONE* 7.2 (2012), e32202.

[65] Okorie, P. N. et al. “Lymphatic filariasis in Nigeria; micro-stratification overlap mapping (MOM) as a prerequisite for cost-effective resource utilization in control and surveillance”. *PLOS Negl Trop Dis* 7.9 (2013), e2416.

[66] Stanton, M. C. et al. “Quantifying filariasis and malaria control activities in relation to lymphatic filariasis elimination: a multiple intervention score map (MISM) for Malawi”. *Trop Med Int Health* 19.2 (2014), pp. 224–35.

[67] Bhatt, S. et al. “The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015”. *Nature* 526.7572 (2015), pp. 207–11.

[68] Moyes, C. L. et al. “Predicting the geographical distributions of the macaque hosts and mosquito vectors of *Plasmodium knowlesi* malaria in forested and non-forested areas”. *Parasit Vectors* 9.1 (2016), p. 242.

[69] Shearer, F. M. et al. “Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria”. *PLOS Negl Trop Dis* 10.8 (2016), e0004915.

[70] William, T. et al. “Increasing incidence of *Plasmodium knowlesi* malaria following control of *P. falciparum* and *P. vivax* malaria in Sabah, Malaysia”. *PLOS Negl Trop Dis* 7.1 (2013).

[71] Barber, B. E. et al. “A prospective comparative study of knowlesi, falciparum, and vivax malaria in Sabah, Malaysia: High proportion with severe disease from *Plasmodium knowlesi* and *Plasmodium vivax* but no mortality with early referral and artesunate therapy”. *Clin Infect Dis* 56.3 (2013), pp. 383–97.

[72] Cox-Singh, J. et al. “*Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening”. *Clin Infect Dis* 46.2 (2008), pp. 165–71.

[73] WHO. *World Malaria Report 2014*. Available from: `http://www.who.int/malaria/publications/world_malaria_report_2014/en/`.

[74] Jongwutiwes, S. et al. “*Plasmodium knowlesi* malaria in humans and macaques, Thailand”. *Emerg Infect Diseases* 17.10 (2011), pp. 1799–806.

[75] Lee, K. S. et al. “*Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques”. *PLOS Pathog* 7.4 (2011), e1002015.

[76] Zhu, H., Li, J., and Zheng, H. “Human natural infection of *Plasmodium knowlesi*”. *Chinese Journal of Parasitology and Parasitic Diseases* 24.1 (2006), p. 70.

[77] Marchand, R. P. et al. “Co-infections of *Plasmodium knowlesi*, *P. falciparum*, and *P. vivax* among humans and *Anopheles dirus* mosquitoes, Southern Vietnam”. *Emerg Infect Dis* 17.7 (2011), pp. 1232–9.

[78] Jiram, A. I. et al. “Entomologic investigation of *Plasmodium knowlesi* vectors in Kuala Lipis, Pahang, Malaysia”. *Malar J* 11 (2012), p. 213.

[79] Vythilingam, I. “*Plasmodium knowlesi* and *Wuchereria bancrofti*: their vectors and challenges for the future”. *Front Physiol* 3 (2012), p. 115.

[80] Wong, M. et al. “Seasonal and spatial dynamics of the primary vector of *Plasmodium knowlesi* within a major transmission focus in Sabah, Malaysia”. *PLOS Negl Trop Dis* 9.10 (2015), e0004135.

[81] Vythilingam, I. et al. “*Plasmodium knowlesi* malaria an emerging public health problem in Hulu Selangor, Selangor, Malaysia (2009-2013): epidemiologic and entomologic analysis”. *Parasit Vectors* 7 (2014), p. 436.

[82] Azidah, A. K. et al. “Severe *Plasmodium knowlesi* infection with multiorgan involvement in north east peninsular Malaysia”. *Trop Biomed* 31.1 (2014), pp. 31–5.

[83] Nakaviroj, S. et al. “An autochthonous case of severe *Plasmodium knowlesi* malaria in Thailand”. *Am J Trop Med Hyg* 92.3 (2015), pp. 569–572.

[84] Seilmaier, M. et al. “Severe *Plasmodium knowlesi* infection with multi-organ failure imported to Germany from Thailand/Myanmar”. *Malar J* 13 (2014), p. 422.

[85] Singh, B. and Daneshvar, C. “Human infections and detection of *Plasmodium knowlesi*”. *Clin Microbiol Rev* 26.2 (2013), pp. 165–84.

[86] Barber, B. E. et al. “Limitations of microscopy to differentiate *Plasmodium* species in a region co-endemic for *Plasmodium falciparum*, *Plasmodium vivax* and *Plasmodium knowlesi*”. *Malar J* 12 (2013).

[87] Goh, X. T. et al. “Increased detection of *Plasmodium knowlesi* in Sandakan division, Sabah as revealed by PlasmoNex™”. *Malar J* 12 (2013), p. 28.

[88] Yusof, R. et al. “High proportion of knowlesi malaria in recent malaria cases in Malaysia”. *Malar J* 13 (2014), p. 168.

[89] Coutrier, F. et al. *Serial molecular identification to confirm the presence of Plasmodium knowlesi in Indonesia*. Poster session presented at: 64th Annual Meeting of The American Society for Tropical Medicine and Hygiene, October 25-29th 2015, Philadelphia, PA.

[90] Moyes, C. L. et al. “Defining the geographical range of the *Plasmodium knowlesi* reservoir”. *PLOS Negl Trop Dis* 8.3 (2014), e2780.

[91] Ng, O. T. et al. “Naturally acquired human *Plasmodium knowlesi* infection, Singapore”. *Emerg Infect Dis* 14.5 (2008), p. 814.

[92] Cotter, C. et al. “The changing epidemiology of malaria elimination: new strategies for new challenges”. *Lancet* 382.9895 (2013), pp. 900–11.

[93] *Global Biodiversity Information Facility Home Page*. GBIF.org (2017). Available from: `http://gbif.org`.

[94] *Malaria Atlas Project Data Explorer*. Malaria Atlas Project (2017). Available from: `http://www.map.ox.ac.uk/explorer/`.

[95] WHO. *World Malaria Report 2012*. Available from: `http://www.who.int/malaria/publications/world_malaria_report_2012/en/`.

[96] Huete, A., Justice, C., and Van Leeuwen, W. *MODIS vegetation index (MOD13). Algorithm theoretical basis document*. 1999.

[97] Lobser, S. and Cohen, W. “MODIS tasselled cap: land cover characteristics expressed through transformed MODIS data”. *Int J Remote Sens* 28 (2007), pp. 5079–101.

[98] Weiss, D. J. et al. “An effective approach for gap-filling continental scale remotely sensed time-series”. *ISPRS J Photogramm Remote Sens* 98 (2014), pp. 106–18.

[99] Friedl, M. A. et al. “MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets”. *Remote Sens Environ* 114.1 (2010), pp. 168–82.

[100] SEDAC. *Gridded Population of the World, Version 4 (GPWv4): Population Count Grid*. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY (2005). Available from: `http://www.worldpop.org.uk/`.

[101] TG, F. et al. “The Shuttle Radar Topography Mission”. *Rev Geophys* 45 (2007), RG2004.

[102] Eyles, D. E. “The species of simian malaria: taxonomy, morphology, life cycle, and geographical distribution of the monkey species”. *J Parasitol* 49.6 (1963), pp. 866–887.

[103] Nelson, A. *Travel time to major cities: A global map of accessibility*. Ispra Italy: Global Environmental Monitoring Unit – Joint Research Centre of the European Commission (2008). Available from: `http://forobs.jrc.ec.europa.eu/products/gam/`.

[104] Potapov, P. et al. “Mapping the world’s intact forest landscapes by remote sensing”. *Ecol Soc* 13.2 (2008), p. 51.

[105] Ridgeway, G. *gbm: Generalized Boosted Regression Models, R package version 2.1.3 (2017)*. Available from: `https://cran.r-project.org/web/packages/gbm/index.html`.

[106] Hansen, M. C. et al. “High-resolution global maps of 21st-century forest cover change”. *Science* 342.6160 (2013), pp. 850–3.

[107] Fornace, K. M. et al. “Mapping infectious disease landscapes: unmanned aerial vehicles and epidemiology”. *Trends Parasitol* 30.11 (2014), pp. 514–9.

[108] Fleiss, J., Levin, B., and Paik, M. *Statistical methods for rates and proportions*. Hoboken, New Jersey: John Wiley and Sons, 2003.

[109] Elith, J. and Leathwick, J. R. “Species distribution models: ecological explanation and prediction across space and time”. *Annu Rev Ecol Evol Syst* 40.1 (2009), pp. 677–97.

[110] Elith, J., Kearney, M., and Phillips, S. “The art of modelling range-shifting species”. *Methods Ecol Evol* 1.4 (2010), pp. 330–42.

[111] Hijmans, R. et al. *dismo: Species Distribution Modelling, R package version 1.1-4 (2017)*. Available from: `https://cran.r-project.org/web/packages/dismo/index.html`.

[112] Hijmans, R. et al. *raster: Geographic Data Analysis and Modelling, R package version 2.5-8 (2016)*. Available from: `https://cran.r-project.org/web/packages/raster/index.html`.

[113] IUCN. *Macaca range maps. In: International Union for Conservation of Nature (IUCN) Red List of Threatened Species. Version 2013.2*. Cambridge, United Kingdom. Available from: `http://www.iucnredlist.org/`. Accessed: January, 2015.

[114] Newby, G. et al. “The path to eradication: a progress report on the malaria-eliminating countries”. *Lancet* 387.10029 (2016), pp. 1775–84.

[115] Divis, P. et al. “Admixture in humans of two divergent *Plasmodium knowlesi* populations associated with different macaque host species”. *PLOS Pathog* 11.5 (2015), e1004888.

[116] Hadi, M. S. and Sumedi, D. P. *Malaria Baru Dari Primata [New Primate Malaria]*. Pusat Komunikasi Publik Setjen Kementerian Kesehatan Ri, 2013.

[117] Sulistyaningsih, E. et al. “Diagnostic difficulties with *Plasmodium knowlesi* infection in humans”. *Emerg Infect Dis* 16.6 (2010), p. 1033.

[118] Vythilingam, I. et al. “*Plasmodium knowlesi* in humans, macaques and mosquitoes in peninsular Malaysia”. *Parasit Vectors* 1.1 (2008), p. 26.

[119] Steenkeste, N. et al. “Sub-microscopic malaria cases and mixed malaria infection in a remote area of high malaria endemicity in Rattanakiri province, Cambodia: implication for malaria elimination”. *Malar J* 9 (2010), p. 108.

[120] Shearer, F. M. et al. “Global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis”. *Lancet Infect Dis* (2017), doi:10.1016/S1473-3099(17)30419-X.

[121] Paules, C. I. and Fauci, A. S. “Yellow fever –once again on the radar screen in the Americas”. *N Engl J Med* 376.15 (2017), pp. 1397–99.

[122] WHO. “Eliminate yellow fever epidemics (EYE): a global strategy, 2017-2026”. *Weekly Epidemiological Record* 16.92 (2017), 193–204.

[123] Abajobir, A. A. et al. “Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016”. *Lancet* 390.10100 (2017), pp. 1151–210.

[124] Abajobir, A. A. et al. “Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016”. *Lancet* 390.10100 (2017), pp. 1260–344.

[125] Rogers, D. et al. “The global distribution of yellow fever and dengue”. *Adv Parasitol* 62 (2006), pp. 181–220.

[126] Heinz, F. and Stiasny, K. “Flaviviruses and flavivirus vaccines”. *Vaccine* 30 (2012), pp. 4301–6.

[127] Johansson, M. A., Vasconcelos, P. F. C., and Staples, J. E. “The whole iceberg: estimating the incidence of yellow fever virus infection from the number of severe cases”. *Trans R Soc Trop Med Hyg* 108.8 (2014), pp. 482–7.

[128] WHO. “Vaccines and vaccination against yellow fever. WHO Position Paper –June 2013”. *Weekly Epidemiological Record* 88 (2013), pp. 269–84.

[129] Garske, T. et al. “Yellow fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological data”. *Plos Med* 11.5 (2014), e1001638.

[130] Monath, T. P. and Vasconcelos, P. F. C. “Yellow fever”. *J Clin Virol* 64 (2015), pp. 160–173.

[131] Barrett, A. D. and Higgs, S. “Yellow fever: a disease that has yet to be conquered”. *Annu Rev Entomol* 52 (2007), pp. 209–29.

[132] Wasserman, S., Tambyah, P., and Lim, P. “Yellow fever cases in Asia: primed for an epidemic”. *J Infect Dis* 48 (2016), pp. 98–103.

[133] Kraemer, M. U. G. et al. “Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015-16: a modelling study”. *Lancet Infect Dis* 17.3 (2017), 330–8.

[134] Maguire, H. and Heymann, D. “Yellow fever in Africa”. *BMJ* 354 (2016), p. i3764.

[135] Gobbelaar, A. et al. “Resurgence of yellow fever in Angola, 2015-2016”. *Infect Dis* 22 (2016), 1854–5.

[136] WHO. *Year of introduction of selected vaccines database in excel*. Available from: `http://www.who.int/entity/immunization/monitoring_surveillance/data/year_vaccine_introduction.xls?ua=1`. Accessed: April 23, 2017.

[137] Barrett, A. “Yellow fever in Angola and beyond –the problem of vaccine supply and demand”. *N Engl J Med* 375 (2016), pp. 301–3.

[138] Wu, J. T. et al. “Fractional dosing of yellow fever vaccine to extend supply: a modelling study”. *Lancet* 388.10062 (2016), pp. 2904–11.

[139] Wiysonge, C. S. et al. “Yellow fever control in Cameroon: where are we now and where are we going?” *BMC Med* 6.1 (2008), p. 3.

[140] WHO. *The Weekly Epidemiological Record*. Available from: `http://www.who.int/wer/en/`. Accessed: August 26, 2016.

[141] WHO. *Disease Outbreak News*. Available from: `http://www.who.int/csr/don/en/`. Accessed: August 26, 2016.

[142] CDC. *Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report*. Available from: `http://www.cdc.gov/mmwr/index2016.html`. Accessed: October 3, 2016.

[143] Burton, A. et al. “WHO and UNICEF estimates of national infant immunization coverage: methods and processes”. *Bull World Health Organ* 87 (2009), 535–41.

[144] WHO. *WHO/UNICEF review of national immunization coverage, 1980-2015.* Available from: `http://apps.who.int/immunization_monitoring/globalsummary/wucoveragecountrylist.html`. Accessed: August 26, 2016.

[145] UN. *United Nations World Population Prospects, the 2015 Revision.* Available from: `https://esa.un.org/unpd/wpp/Download/Standard/Population/`. Accessed: October 13, 2016.

[146] UN. *United Nations World Population Prospects, the 2015 Revision.* Available from: `https://esa.un.org/unpd/wpp/Download/Standard/Mortality/`. Accessed: October 10, 2016.

[147] WHO. *Country reported administrative data time series in excel.* Available from: `http://www.who.int/entity/immunization/monitoring_surveillance/data/coverage_series.xls?ua=1`. Accessed: October 13, 2016.

[148] WHO. *WHO/UNICEF Joint Reporting Process.* Available from: `http://www.who.int/immunization/monitoring_surveillance/routine/reporting/reporting/en/`. Accessed: April 24, 2017.

[149] WHO. *Report on the fight against the yellow fever epidemic 2016 in the Democratic Republic of Congo*. World Health Organization, Kinshasa, 2016.

[150] Durieux, K. “Mass yellow fever vaccination in French Africa south of the Sahara”. *Monograph Series World Health Organization* 30 (1956), pp. 115–22.

[151] Bosch-Capblanch, X. et al. “Accuracy and quality of immunization information systems in forty-one low income countries”. *Trop Med Int Health* 14 (2009), 2–10.

[152] Lim, S. et al. “Tracking progress towards universal childhood immunisation and the impact of global initiatives: a systematic analysis of three-dose diphtheria, tetanus, and pertussis immunization coverage”. *Lancet* 372 (2008), pp. 2031–46.

[153] Murray, C. et al. “Validity of reported vaccination coverage in 45 countries”. *Lancet* 362 (2003), pp. 1022–7.

[154] Massad, E. et al. “Yellow fever vaccination: how much is enough?” *Vaccine* 23.30 (2005), pp. 3908–14.

[155] WHO. *Disease Outbreak News 14 June 2016. Yellow fever –Angola.* Available from: `http://www.who.int/entity/csr/don/14-june-2016-yellow-fever-angola/en/index.html`. Accessed: March 16, 2017.

[156] WHO. *Yellow Fever Initiative - Providing an opportunity of a lifetime*. World Health Organization, Geneva, 2010. Available from: `http://www.who.int/csr/disease/yellowfev/YFIbrochure.pdf`.

[157] Lessler, J. et al. “Measuring the performance of vaccination programs using cross-sectional surveys: a likelihood framework and retrospective analysis”. *PLOS Med* 8.10 (2011), e1001110.

[158] Shearer, F. M. et al. “Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis”. *Lancet Glob Health* (2018), doi:10.1016/S2214-109X(18)30024-X.

[159] Vasconcelos, P. and Monath, T. P. “Yellow fever remains a potential threat to public health”. *Vector Borne Zoonot* 16.8 (2016), pp. 566–7.

[160] Chen, J. and Lu Hongzhou, L. “Yellow fever in China is still an imported disease”. *Biosci Trends* 10 (2016), pp. 158–62.

[161] Staples, J. E. et al. “Yellow fever risk assessment in the Central African Republic”. *Trans R Soc Trop Med Hyg* 108.10 (2014), pp. 608–15.

[162] Butler, D. “Fears rise over yellow fever’s next move”. *Nature* 532 (2016), pp. 155–6.

[163] Soper, F. L. “The elimination of urban yellow fever in the Americas through the eradication of *Aedes aegypti*”. *Am J Public Health Nations Health* 53.1 (1963), pp. 7–16.

[164] “Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015”. *Lancet* 388.10053 (2016), pp. 1545–1602.

[165] FAO. *GeoNetwork*. Food and Agriculture Organization of the United Nations (2017). Available from: `http://www.fao.org/geonetwork/`.

[166] Hijmans, R. *GADM database of global administrative areas, version 2.8 (2015)*. Available from: `http://www.gadm.org/`.

[167] Rodhain, F. “The role of monkeys in the biology of dengue and yellow fever”. *Comp Immunol Microbiol Infect Dis* 14.1 (1991), pp. 9–19.

[168] Findlay, G. M. and Maccallum, F. O. “Yellow fever immune bodies in the blood of African primates”. *Trans R Soc Trop Med Hyg* 31.1 (1937), pp. 103–6.

[169] Traoré-Lamizana, M. et al. “Surveillance for yellow fever virus in eastern Senegal during 1993”. *J Med Entomol* 33.5 (1996), pp. 760–5.

[170] Simpson, D. et al. “Going off grid: computationally efficient inference for log-Gaussian Cox processes”. *Biometrika* 103 (2016), pp. 49–70.

[171] Monath, T. P. et al. “Arbovirus studies in Nupeko forest, a possible natural focus of yellow fever virus in Nigeria. I. Description of the area and serological survey of humans and other vertebrate hosts”. *Trans R Soc Trop Med Hyg* 68.1 (1974), pp. 30–8.

[172] Kirya, B. G. and Okia, N. O. “A yellow fever epizootic in Zika Forest, Uganda, during 1972: part 2: monkey serology”. *Trans R Soc Trop Med Hyg* 71.4 (1977), pp. 300–3.

[173] Haddow, A. J. et al. “Monkeys in relation to yellow fever in Bwamba County, Uganda”. *Trans R Soc Trop Med Hyg* 40.5 (1947), pp. 677–700.

[174] Haddow, A. J. et al. “Monkeys in relation to the epidemiology of yellow fever in Uganda”. *Trans R Soc Trop Med Hyg* 45.2 (1951), pp. 189–224.

[175] Mathiot, C. C., Herve, V. M., and Georges, A. J. “Antibodies to haemorrhagic fever viruses and to selected arboviruses in monkeys from the Central African Republic”. *Trans R Soc Trop Med Hyg* 84.5 (1990), pp. 732–3.

[176] Henderson, B. E. et al. “Immunologic studies with yellow fever and selected African group B arboviruses in rhesus and vervet monkeys”. *Am J Trop Med Hyg* 19.1 (1970), pp. 110–8.

[177] Woodall, J. P. “The reaction of a mangabey monkey (*Cercocebus galeritus agilis* Milne-Edwards) to inoculation with yellow fever virus”. *Ann Trop Med Parasitol* 62.4 (1968), pp. 522–7.

[178] Woodall, J. P., Dykes, J. R., and Williams, M. C. “The reaction of a species of colobus monkey to inoculation with yellow fever virus”. *Ann Trop Med Parasitol* 62.4 (1968), pp. 528–35.

[179] Trapido, H. and Galindo, P. “The investigation of a sylvan yellow fever epizootic on the north coast of Honduras, 1954”. *Am J Trop Med Hyg* 4.4 (1955), pp. 665–74.

[180] Holzmann, I. et al. “Impact of yellow fever outbreaks on two howler monkey species (*Alouatta guariba clamitans* and *A. caraya*) in Misiones, Argentina”. *Am J Primatol* 72.6 (2010), pp. 475–80.

[181] Almeida, M. A. de et al. “Yellow fever outbreak affecting *Alouatta* populations in southern Brazil (Rio Grande do Sul State), 2008–2009”. *Am J Primatol* 74.1 (2012), pp. 68–76.

[182] Vasconcelos, P. F. et al. “Yellow fever in Pará State, Amazon region of Brazil, 1998–1999: entomologic and epidemiologic findings”. *Emerg Infect Dis* 7.3 Suppl (2001), pp. 565–9.

[183] Sallis, E. et al. “A case of yellow fever in a brown howler (*Alouatta fusca*) in Southern Brazil”. *J Vet Diagn Invest* 15.6 (2003), pp. 574–6.

[184] Moreno, E. S. et al. “Reemergence of yellow fever: detection of transmission in the State of São Paulo, Brazil, 2008”. *Rev Soc Bras Med Trop* 44.3 (2011), pp. 290–6.

[185] Moreno, E. S. et al. “Yellow fever epizootics in non-human primates, São Paulo state, Brazil, 2008-2009”. *Revista do Instituto de Medicina Tropical de São Paulo* 55 (2013), pp. 45–50.

[186] Leal, S. G. et al. “Frequency of histopathological changes in Howler monkeys (*Alouatta sp.*) naturally infected with yellow fever virus in Brazil”. *Revista da Sociedade Brasileira de Medicina Tropical* 49 (2016), pp. 29–33.

[187] Auguste, A. J. et al. “Enzootic transmission of yellow fever virus, Venezuela”. *Emerg Infect Dis* 21.1 (2015), pp. 99–102.

[188] Anderson, C. R. and Downs, W. G. “The isolation of yellow fever virus from the livers of naturally infected red howler monkeys”. *Am J Trop Med Hyg* 4.4 (1955), pp. 662–4.

[189] Golding, N. et al. “Integrating vector control across diseases”. *BMC Med* 13.1 (2015), p. 249.

[190] Bates, M. and Roca-Garcia, M. “Experiments with various Colombian marsupials and primates in laboratory cycles of yellow fever”. *Am J Trop Med Hyg* 26 (1946), pp. 437–53.

[191] McCrae, A. W. and Kirya, B. G. “Yellow fever and Zika virus epizootics and enzootics in Uganda”. *Trans R Soc Trop Med Hyg* 76.4 (1982), pp. 552–62.

[192] Lumsden, W. H. and Buxton, A. P. “A study of the epidemiology of yellow fever in West Nile District, Uganda”. *Trans R Soc Trop Med Hyg* 45.1 (1951), pp. 53–78.

[193] Sérié, C. et al. “Etudes sur la fièvre jaune en Ethiopie. 6. Etude épidémiologique”. *Bull World Health Organ* 38.6 (1968), pp. 879–84.

[194] Haddow, A. J. “A review of the results of yellow fever protection-tests on the sera of primates from Kenya”. *Ann Trop Med Parasitol* 46.2 (1952), pp. 135–43.

[195] Kading, R. C. et al. “Prevalence of antibodies to alphaviruses and flaviviruses in free-ranging game animals and nonhuman primates in the greater Congo basin”. *J Wildl Dis* 49.3 (2013), pp. 587–99.

[196] Felsenfeld, O., Wolf, R. H., and Dutta, N. K. “Serological responses of patas monkeys (*Erythrocebus patas*) to vaccination against cholera and yellow fever”. *Proc Soc Exp Biol Med* 143.2 (1973), pp. 548–50.

[197] Deubel, V. et al. “Homogeneity among Senegalese strains of yellow fever virus”. *Am J Trop Med Hyg* 34.5 (1985), pp. 976–83.

[198] Taylor, R. M., Haseeb, M. A., and Work, T. H. “A regional reconnaissance on yellow fever in the Sudan; with special reference to primate hosts”. *Bull World Health Organ* 12.5 (1955), pp. 711–25.

[199] Haddow, A. J. and Ellice, J. M. “Studies on bush-babies (*Galago spp.*) with special reference to the epidemiology of yellow fever”. *Trans R Soc Trop Med Hyg* 58.6 (1964), pp. 521–38.

[200] Findlay, G. M. et al. “Yellow fever immune bodies in the blood of African animals”. *Trans R Soc Trop Med Hyg* 29.4 (1936), pp. 419–24.

[201] Laemmert H. W., J. and Kumm, H. W. “The susceptibility of howler monkeys to yellow fever virus”. *Am J Trop Med Hyg* 30.5 (1950), pp. 723–31.

[202] Boshell, J. and Bevier, G. A. “Yellow fever in the lower Motagua Valley, Guatemala”. *Am J Trop Med Hyg* 7.1 (1958), pp. 25–35.

[203] Galindo, P. and De Rodaniche, E. “Surveillance for sylvan yellow fever activity in Panama (1957-1961)”. *Am J Trop Med Hyg* 13 (1964), pp. 844–50.

[204] De Rodaniche, E. “Survey of primates captured in Panama, R. P. during the years 1952-1956 for protective antibodies against yellow fever”. *Am J Trop Med Hyg* 6.5 (1957), pp. 835–9.

[205] Vargas-Mendez, O. and Elton, N. W. “Naturally acquired yellow fever in wild monkeys of Costa Rica”. *Am J Trop Med Hyg* 2.5 (1953), pp. 850–63.

[206] Morales, M. A. et al. “Detection of the mosquito-borne flaviviruses, West Nile, dengue, Saint Louis encephalitis, Ilheus, Bussuquara, and yellow fever in free-ranging black howlers (*Alouatta caraya*) of Northeastern Argentina”. *PLOS Negl Trop Dis* 11.2 (2017), e0005351.

[207] Tranquilin, M. V. et al. “First report of yellow fever virus in non-human primates in the State of Parana, Brazil”. *Rev Soc Bras Med Trop* 46.4 (2013), pp. 522–4.

[208] Thoisy, B. de, Dussart, P., and Kazanji, M. “Wild terrestrial rainforest mammals as potential reservoirs for flaviviruses (yellow fever, dengue 2 and St Louis encephalitis viruses) in French Guiana”. *Trans R Soc Trop Med Hyg* 98.7 (2004), pp. 409–12.

[209] Thoisy, B. de et al. “Health evaluation of translocated free-ranging primates in French Guiana”. *Am J Primatol* 54.1 (2001), pp. 1–16.

[210] Karesh, W. B. et al. “Immobilization and health assessment of free-ranging black spider monkeys (*Ateles paniscus chamek*)”. *Am J Primatol* 44.2 (1998), pp. 107–23.

[211] De Rodaniche, E. “Survey of live forest animals for protective antibodies against yellow fever in panama, R.P.” *Am J Trop Med Hyg* 1.5 (1952), pp. 789–95.

[212] Downs, W. G. “Epidemiological notes in connection with the 1954 outbreak of yellow fever in Trinidad, B.W.I.” *Yellow fever, a symposium in commemoration of Carlos Juan Finlay, 1955. Paper 4.* (1955).

[213] Pinheiro, F. P., Travassos da Rosa, A. P., and Moraes, M. A. “An epidemic of yellow fever in Central Brazil, 1972–1973. II. Ecological studies”. *Am J Trop Med Hyg* 30.1 (1981), pp. 204–11.

[214] Lima, M. A., Romano-Lieber, N. S., and Duarte, A. M. R. d. C. “Circulation of antibodies against yellow fever virus in a simian population in the area of Porto Primavera Hydroelectric Plant, São Paulo, Brazil”. *Rev Inst Med Trop S Paulo* 52 (2010), pp. 11–16.

[215] Price, W. H., Casals, J., and O’Leary, W. “Studies on the sequential immunization against group B arboviruses in squirrel monkeys, cynomolgus monkeys, rhesus monkeys, and chimpanzees”. *Am J Trop Med Hyg* 23.1 (1974), pp. 118–30.

[216] Levenbook, I. S., Pelleu, L. J., and Elisberg, B. L. “The monkey safety test for neurovirulence of yellow fever vaccines: the utility of quantitative clinical evaluation and histological examination”. *J Biol Stand* 15.4 (1987), pp. 305–13.

[217] Stokes, A., Bauer, J., and Hudson, N. “Experimental transmission of yellow fever to laboratory animals.” *Am J Trop Med Hyg* 1 (1928), pp. 103–64.

[218] Stokes, A., Bauer, J. H., and Hudson, N. P. “The transmission of yellow fever to *Macacus rhesus*. 1928”. *Rev Med Virol* 11.3 (2001), pp. 141–8.

[219] Julander, J. G. “Animal models of yellow fever and their application in clinical research”. *Curr Opin Virol* 18 (2016), pp. 64–9.

[220] Ridgeway, G. *gbm3: Generalized Boosted Regression Models, R package version 2.1.3 (2017)*. Available from: `https://github.com/gbm-developers/gbm3`.

[221] Wharton, D. and Shepherd, L. “Poisson point process models solve the ‘pseudo-absence problem’ for presence-only data in ecology”. *Ann Appl Stat* 4 (2010), pp. 1383–402.

[222] Aarts, G., Fieberg, J., and Matthiopoulos, J. “Comparative interpretation of count, presence–absence and point methods for species distribution models”. *Methods Ecol Evol* 3.1 (2012), pp. 177–187.

[223] Renner, I. W. et al. “Point process models for presence-only analysis”. *Methods Ecol Evol* 6.4 (2015), pp. 366–79.

[224] Fithian, W. et al. “Bias correction in species distribution models: pooling survey and collection data for multiple species”. *Methods Ecol Evol* 6.4 (2015), pp. 424–438.

[225] Berman, M. and Turner, T. “Approximating point process likelihoods with GLIM”. *Appl Statist* 41 (1992), pp. 31–8.

[226] Gibbons, C. L. et al. “Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods”. *BMC Public Health* 14 (2014), p. 147.

[227] Moyes, C. L. et al. “Providing open access data online to advance malaria research and control”. *Malar J* 12.1 (2013), p. 161.

[228] Long, J. S. *Regression model for categorical and limited dependent variables*. Thousand Oaks: Sage Publications, 1997.

[229] Elton, N. “Anticipated progress of yellow fever in Gautemala and Mexico, 1955–1959”. *Am J Public Health* 45.7 (1955), pp. 923–7.

[230] “Yellow fever in Yucatan”. *Lancet* 180.4661 (1912), pp. 1830–1831.

[231] WHO. *4th Meeting Report of the Scientific and Technical Advisory Group on Yellow Fever Risk Mapping, January 2017*. Available from: `http://www.who.int/ith/yellow-fever-risk-mapping/4th-GRYF-meeting-report.pdf?ua=1`.

[232] Staples, J. E. and Monath, T. P. “Yellow fever: 100 years of discovery”. *JAMA* 300.8 (2008), pp. 960–962.

[233] Amaku, M., Coutinho, F. A. B., and Massad, E. “Why dengue and yellow fever coexist in some areas of the world and not in others?” *Biosystems* 106.2 (2011), pp. 111–120.

[234] Allwinn, R. et al. “Cross-reactivity in flavivirus serology: new implications of an old finding?” *Med Microbiol Immunol* 190.4 (2002), pp. 199–202.

[235] Purse, B. V. et al. “Impacts of climate, host and landscape factors on Culicoides species in Scotland”. *Med Vet Entomol* 26.2 (2012), pp. 168–77.

[236] D. Kissling, W. et al. “Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents”. *J Biogeogr* 39 (2011), pp. 2163–78.

[237] Lord, J. S., Gurley, E. S., and Pulliam, J. R. “Rethinking Japanese encephalitis virus transmission: a framework for implicating host and vector species”. *PLoS Negl Trop Dis* 9.12 (2015), e0004074.

[238] Wheatley, B. P. “Malaria as a possible selective factor in the speciation of macaques”. *J Mammal* 61.2 (1980), pp. 307–11.

[239] Garnham, P. C. “Distribution of simian malaria parasites in various hosts”. *J Parasitol* 49 (1963), pp. 905–11.

[240] Wardrop, N. A. et al. “Interpreting predictive maps of disease: highlighting the pitfalls of distribution models in epidemiology”. *Geospat Health* 9.1 (2014), pp. 237–46.

[241] Golding, N. and Purse, B. V. “Fast and flexible Bayesian species distribution modelling using Gaussian processes”. *Methods Ecol Evol* 7.5 (2016), pp. 598–608.

[242] Pollock, L. J. et al. “Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)”. *Methods Ecol Evol* 5.5 (2014), pp. 397–406.

[243] Meyer, C. et al. “Global priorities for an effective information basis of biodiversity distributions”. *Nat Commun* 6 (2015), p. 8221.

[244] Dorazio, R. M. “Accounting for imperfect detection and survey bias in statistical analysis of presence-only data”. *Glob Ecol Biogeogr* 23.12 (2014), pp. 1472–84.

[245] Weiss, D. et al. “Mapping inequalities in accessibility: a global assessment of travel time to population centers in 2015” (Submitted).

[246] Boden, L. A. and McKendrick, I. J. “Model-based policymaking: a framework to promote ethical good practice in mathematical modelling for public health policymaking”. *Front Public Health* 5 (2017), p. 68.

[247] Sedda, L. et al. “Risk assessment of vector-borne diseases for public health governance”. *Public Health* 128.12 (2014), pp. 1049–58.

[248] Golding, N. et al. “The zoon R package for reproducible and shareable species distribution modelling”. *Methods Ecol Evol* Epub Jul 30 (2017).

[249] Leach, M. and Scoones, I. “The social and political lives of zoonotic disease models: narratives, science and policy”. *Soc Sci Med* 88 (2013), pp. 10–7.

[250] Whitty, C. J. M. “What makes an academic paper useful for health policy?” *BMC Med* 13.1 (2015), p. 301.

[251] Bardosh, K. “Global aspirations, local realities: the role of social science research in controlling neglected tropical diseases”. *Infect Dis Poverty* 3 (2014), p. 35.

[252] Daszak, P. et al. “Interdisciplinary approaches to understanding disease emergence: the past, present, and future drivers of Nipah virus emergence”. *Proc Natl Acad Sci U S A* 110 Suppl 1 (2013), pp. 3681–8.

[253] Imai, N. et al. “Transmission and control of *Plasmodium knowlesi*: a mathematical modelling study”. *PLoS Negl Trop Dis* 8.7 (2014), e2978.

[254] Cox-Singh, J. and Culleton, R. “*Plasmodium knowlesi*: from severe zoonosis to animal model”. *Trends Parasitol* 31.6 (2015), pp. 232–8.

[255] Bhatt, S. et al. “Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization”. *J R Soc Interface* 14.134 (2017), p. 20170520.

Appendices

A

Plasmodium knowlesi malaria risk mapping

A.1 Plots of model covariates

In this section, Figures A.1, A.2, and A.3 display plots of land cover, environmental and socioeconomic covariates supplied to the niche models for *P. knowlesi* malaria and its reservoir and vector species.

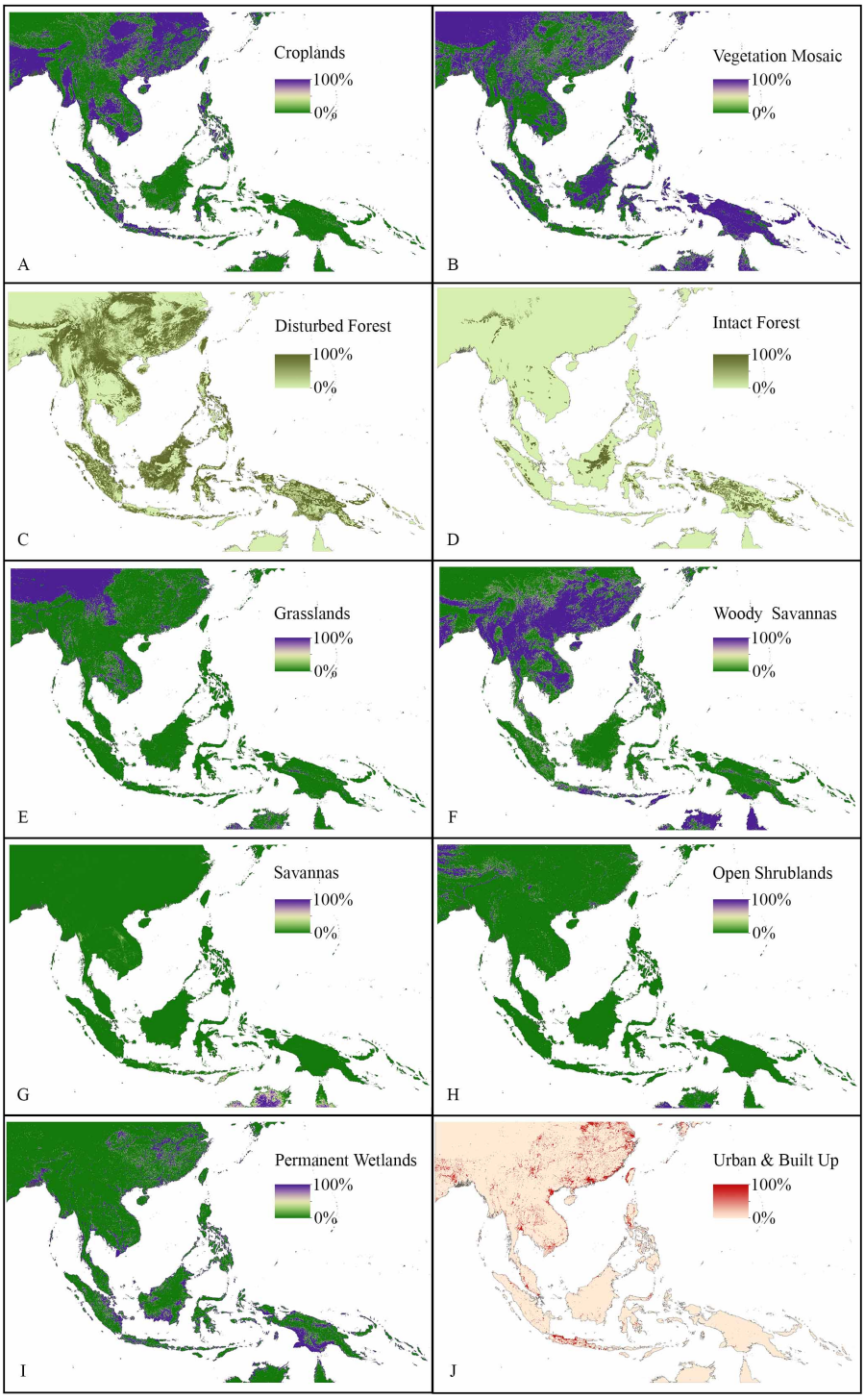


Figure A.1: Land cover covariates supplied to distribution models for *Plasmodium knowlesi* malaria and its reservoir and vector species. A-J. Proportional cover for 2012 of lands with croplands, cropland/natural vegetation mosaic (=Vegetation Mosaic), disturbed forest, intact forest, grasslands, woody savannas, savannas, open shrublands, permanent wetlands and urban and built-up, respectively. For details on how each of these covariates layers were derived, their source, and to which models they were supplied, see Section 2.2.3

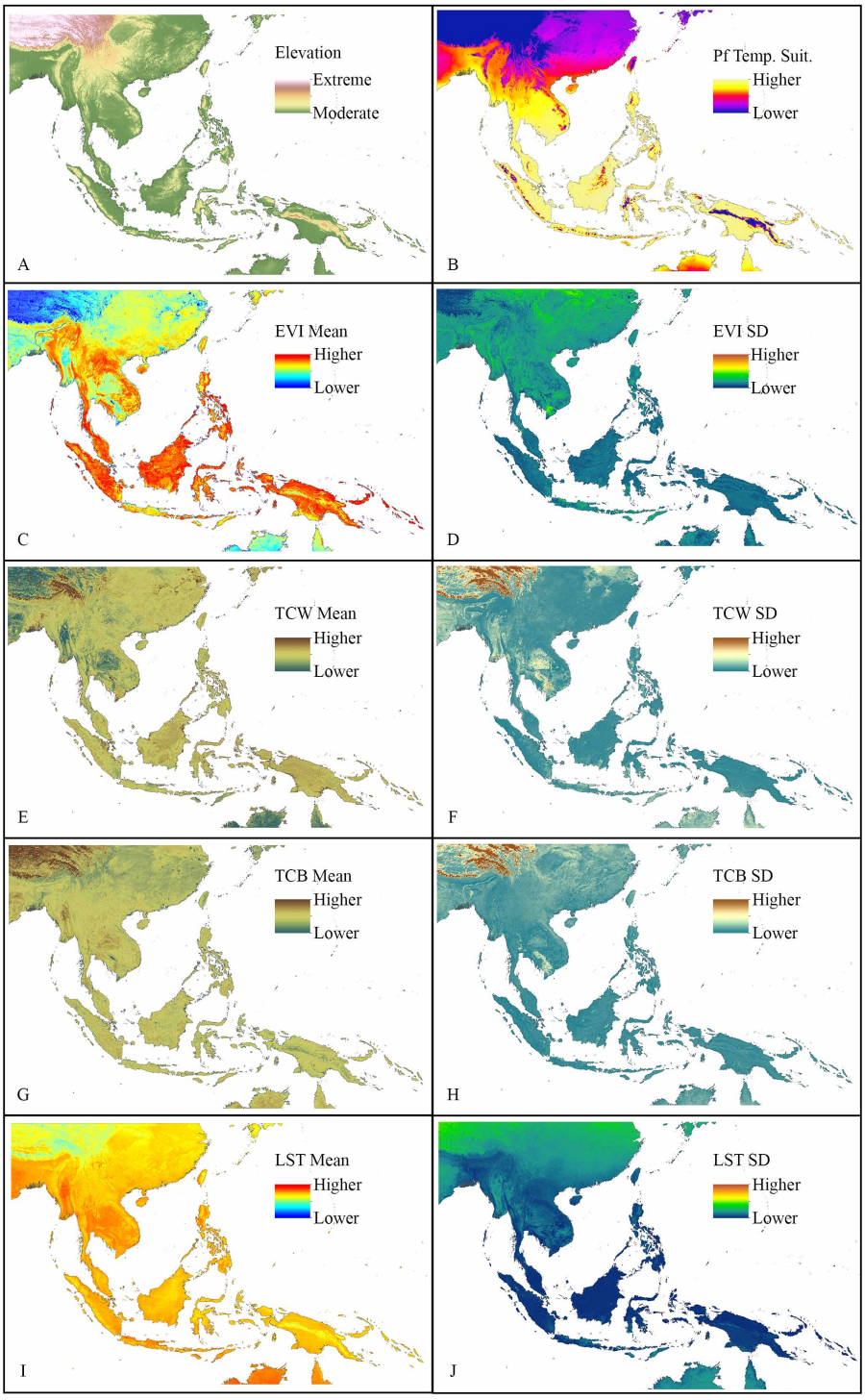


Figure A.2: Environmental covariates supplied to distribution models for *Plasmodium knowlesi* malaria and its reservoir and vector species. A. Elevation. **B.** Index of temperature suitability for *P. falciparum* transmission. **C and D.** Enhanced vegetation index, which is a measure of surface moisture (mean and standard deviation, respectively). **E and F.** tasselled cap wetness, which is a measure of surface moisture (mean and standard deviation, respectively). **G and H.** tasselled cap brightness, which is a measure of moisture on bare surfaces (mean and standard deviation respectively). **I and J.** Daytime land surface temperature (mean and standard deviation, respectively). For details on how each of these covariates layers were derived, their source, and to which models they were supplied, see Section 2.2.3

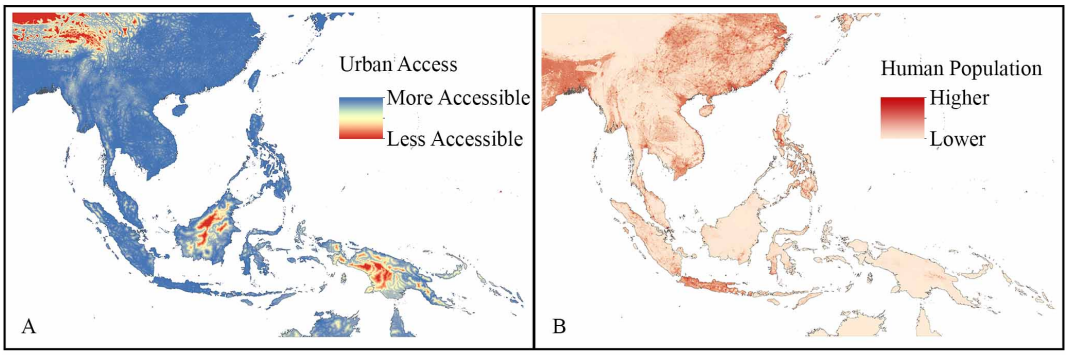


Figure A.3: Socioeconomic covariates supplied to distribution models for *Plasmodium knowlesi* malaria and its reservoir and vector species. A. Time required to travel from each geographic location to a large city *via* land or water-based transport networks. **B.** Human population density. For details on how each of these covariates layers were derived, their source, and to which models they were supplied, see Section 2.2.3

A.2 Species model uncertainty plots and top predictors

In this section, Tables A.1 to A.7 display relative influence values for the most influential covariates for each species niche model. The relative influence of predictor variables can be estimated by calculating the number of times a variable is selected for splitting, weighted by the squared improvement to the model as result of each split and averaged over all trees. These values can be scaled to sum to 100 so that a higher number indicates greater effect of the variables on the response. The most influential predictors were defined as those whose relative influence was greater than 100 divided by the total number of covariates. Figures A.4 to A.10 show the 0.025 (lower) and 0.975 (upper) quantiles for each of the species model ensembles.

Table A.1: Relative influence values of the most influential predictors of suitability for *Macaca nemestrina*.

Predictor	Relative influence (%)
Variation in daytime temperature through time	57.52
Variation in tasselled cap brightness through time (seasonality in moisture)	10.77
Human population density	9.23

Table A.2: Relative influence values of the most influential predictors of suitability for *Macaca leonina*.

Predictor	Relative influence (%)
Variation in the enhanced vegetation index through time (seasonality in greenness/moisture)	15.81
Elevation	12.64
Variation in tasselled cap wetness through time (seasonality in moisture)	10.38
Human population density	9.25
Variation in tasselled cap brightness through time (seasonality in moisture)	8.18
Variation in daytime temperature through time	7.36
Woody savannas (30-60% tree cover) land cover class	6.68
tasselled cap wetness	5.96

Table A.3: Relative influence values of the most influential predictors of suitability for *Macaca fascicularis*.

Predictor	Relative influence (%)
Elevation	18.87
Daytime temperature	13.27
Variation in tasselled cap brightness through time (seasonality in moisture)	9.66
Variation in daytime temperature through time	8.42
Human population density	7.48
Enhanced vegetation index (greenness and moisture)	6.43
Permanent wetlands land cover class	5.66

Table A.4: Relative influence values of the most influential predictors of suitability for *Anopheles dirus*.

Predictor	Relative influence (%)
Variation in the enhanced vegetation index through time (seasonality in greenness/moisture)	35.21
tasselled cap wetness	16.68
Elevation	9.21
Daytime temperature	6.81
Permanent wetlands land cover class	5.43

Table A.5: Relative influence values of the most influential predictors of suitability for members of the Dirus Complex.

Predictor	Relative influence (%)
tasselled cap wetness	21.18
Variation in daytime temperature through time	14.67
Human population density	11.20
Variation in the enhanced vegetation index through time (seasonality in greenness/moisture)	9.55
Elevation	8.24
Cropland/natural vegetation mosaic land cover class	6.40

Table A.6: Relative influence values of the most influential predictors of suitability for members of the Leucosphyrus Complex.

Predictor	Relative influence (%)
Woody savannas (30-60% tree cover) land cover class	38.75
Variation in daytime temperature through time	13.74
Disturbed forest land cover class	10.08
Enhanced vegetation index (greenness and moisture)	7.97
Permanent wetlands land cover class	6.54
Human population density	5.28

Table A.7: Relative influence values of the most influential predictors of suitability for members of the Leucosphyrus Group.

Predictor	Relative influence (%)
tasselled cap wetness	18.94
Variation in the enhanced vegetation index through time (seasonality in greenness/moisture)	12.79
Human population density	11.03
Variation in daytime temperature through time	9.87
Elevation	8.85
Variation in tasselled cap brightness through time (seasonality in moisture)	6.03

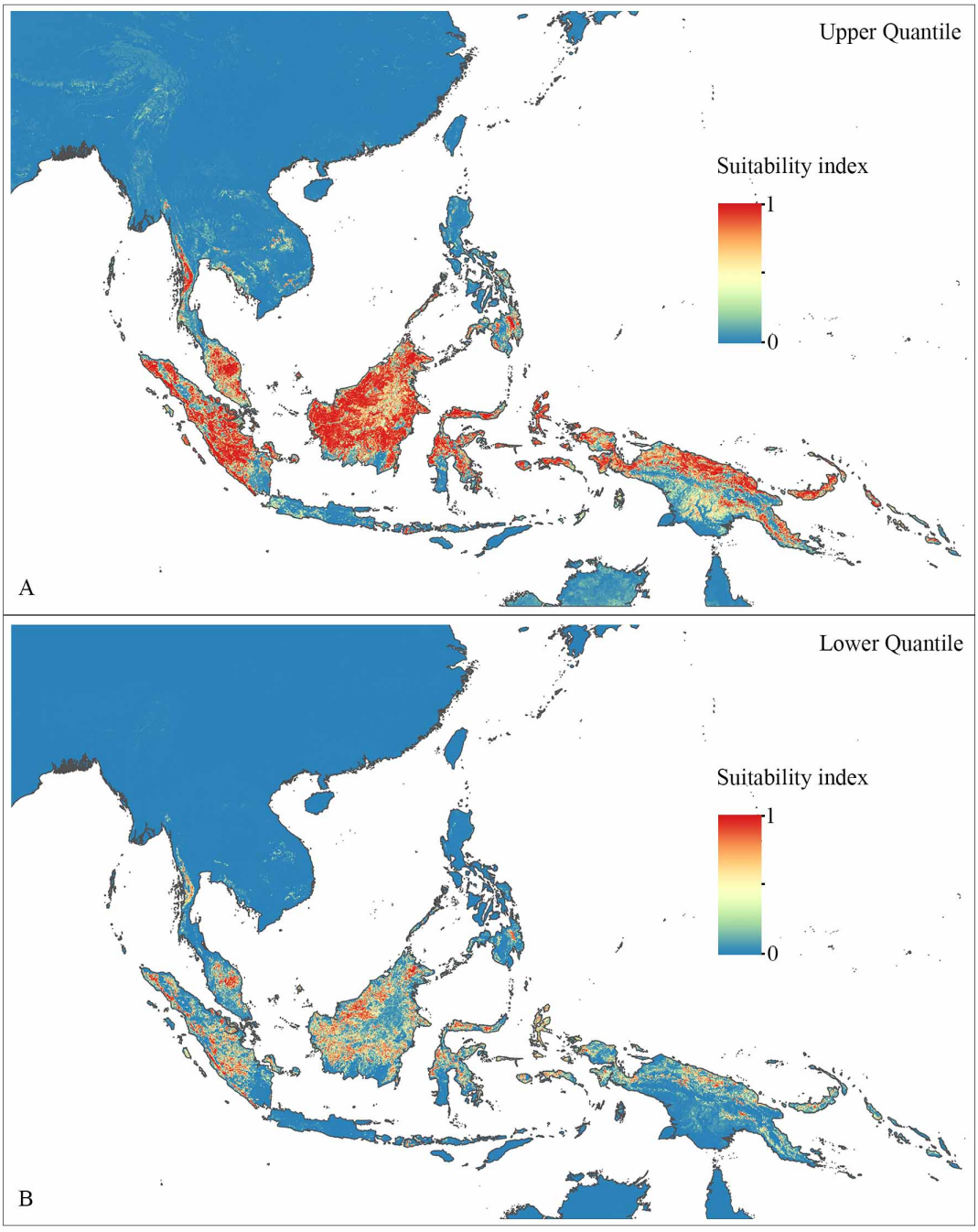


Figure A.4: Uncertainty plots for *Macaca nemestrina* model ensemble. A. Upper (0.975) and **B.** lower (0.025) quantiles for *Macaca nemestrina* model ensemble

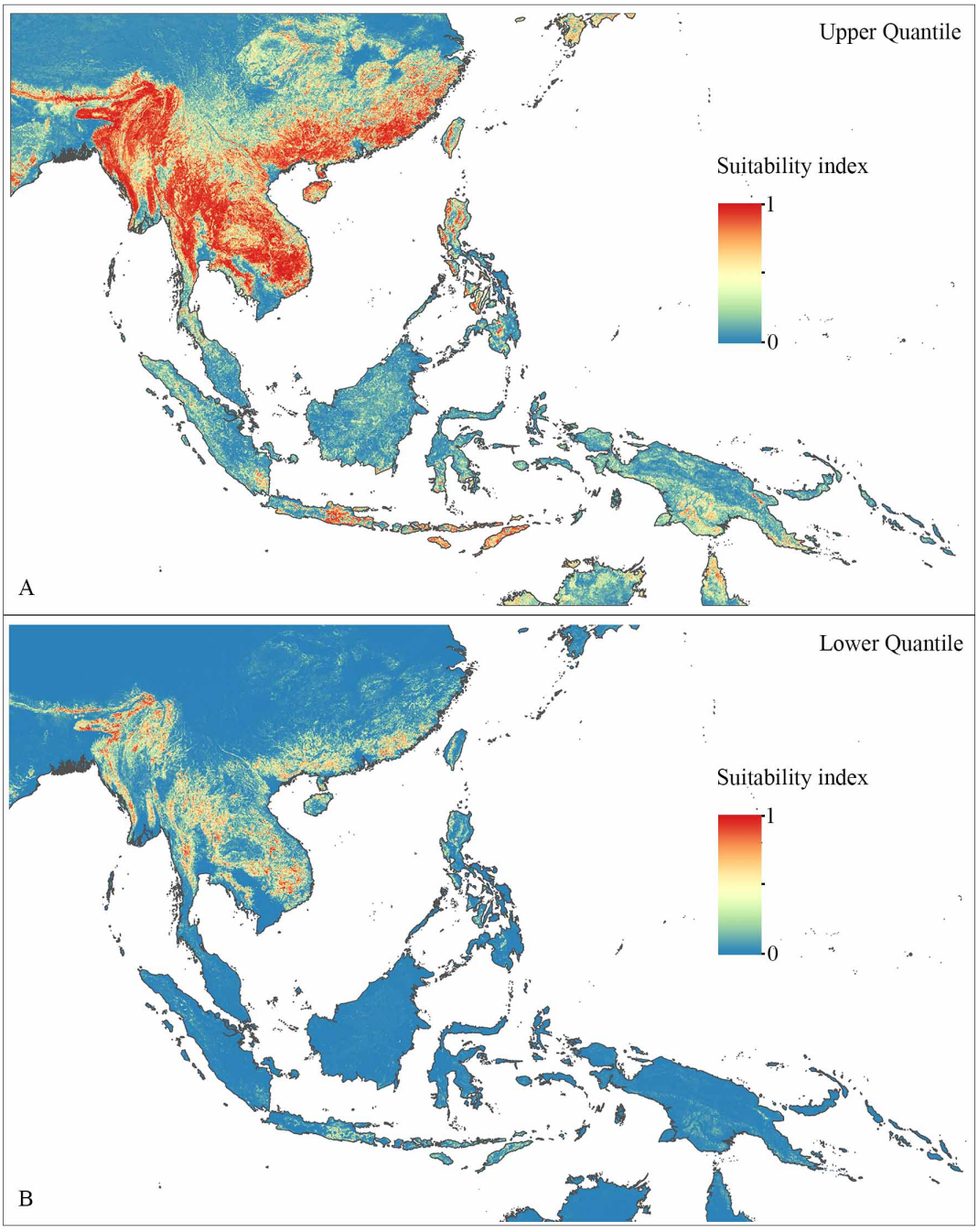


Figure A.5: Uncertainty plots for *Macaca leonina* model ensemble. **A.** Upper (0.975) and **B.** lower (0.025) quantiles for *M. leonina* model ensemble

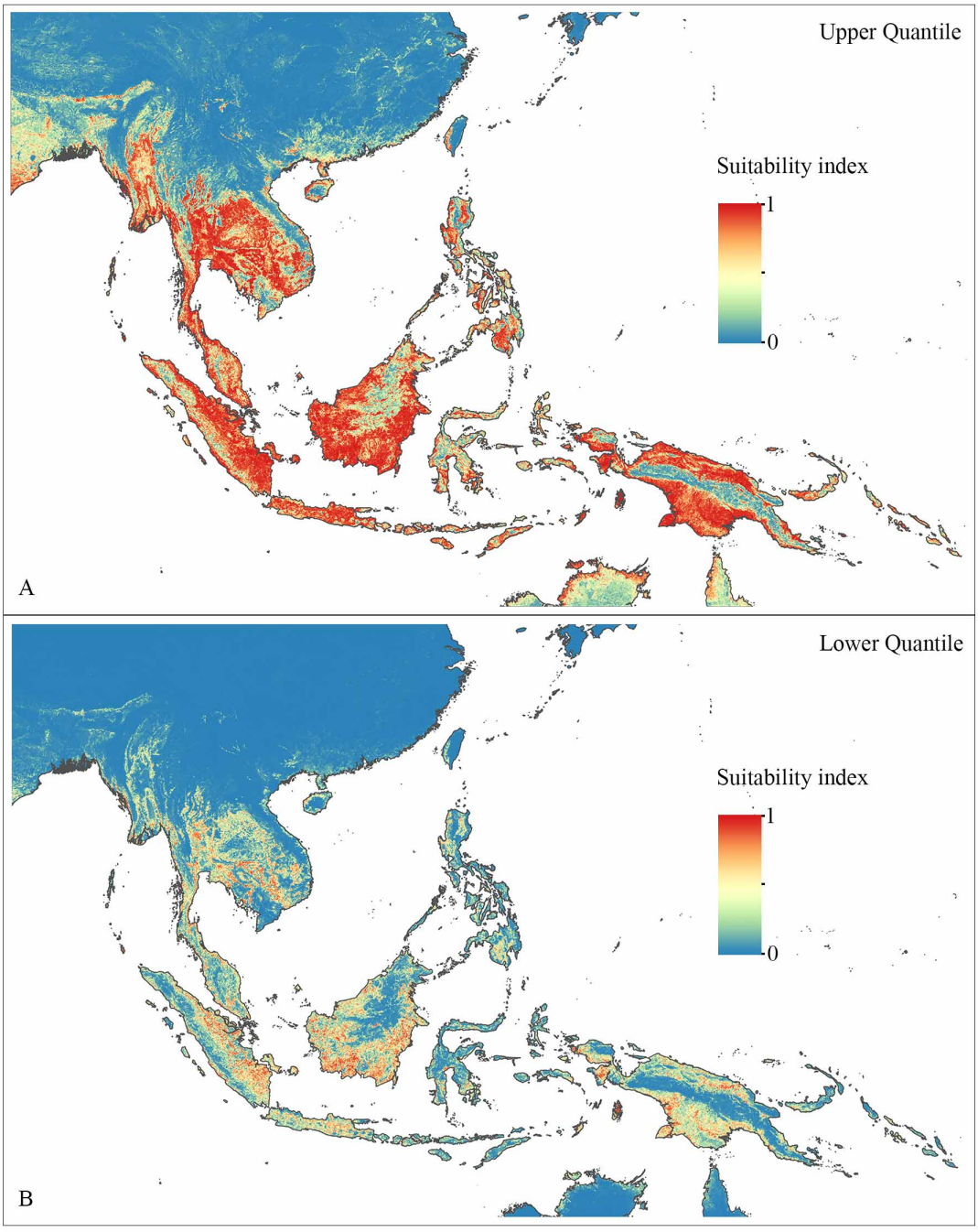


Figure A.6: Uncertainty plots for *Macaca fascicularis* model ensemble. **A.** Upper (0.975) and **B.** lower (0.025) quantiles for *M. fascicularis* model ensemble

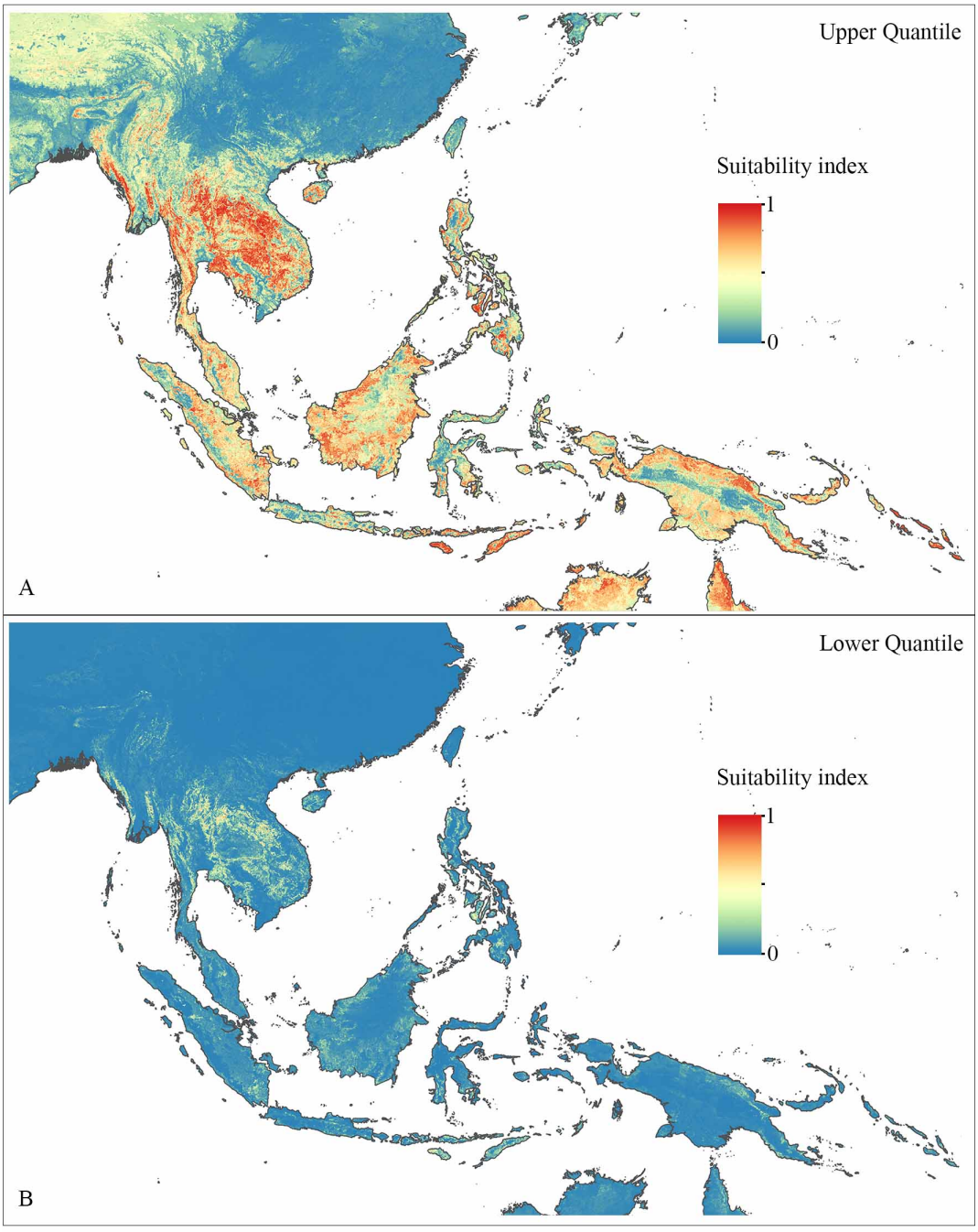


Figure A.7: Uncertainty plots for *Anopheles dirus* model ensemble.A. Upper (0.975) and **B.** lower (0.025) quantiles for *Anopheles dirus* model ensemble

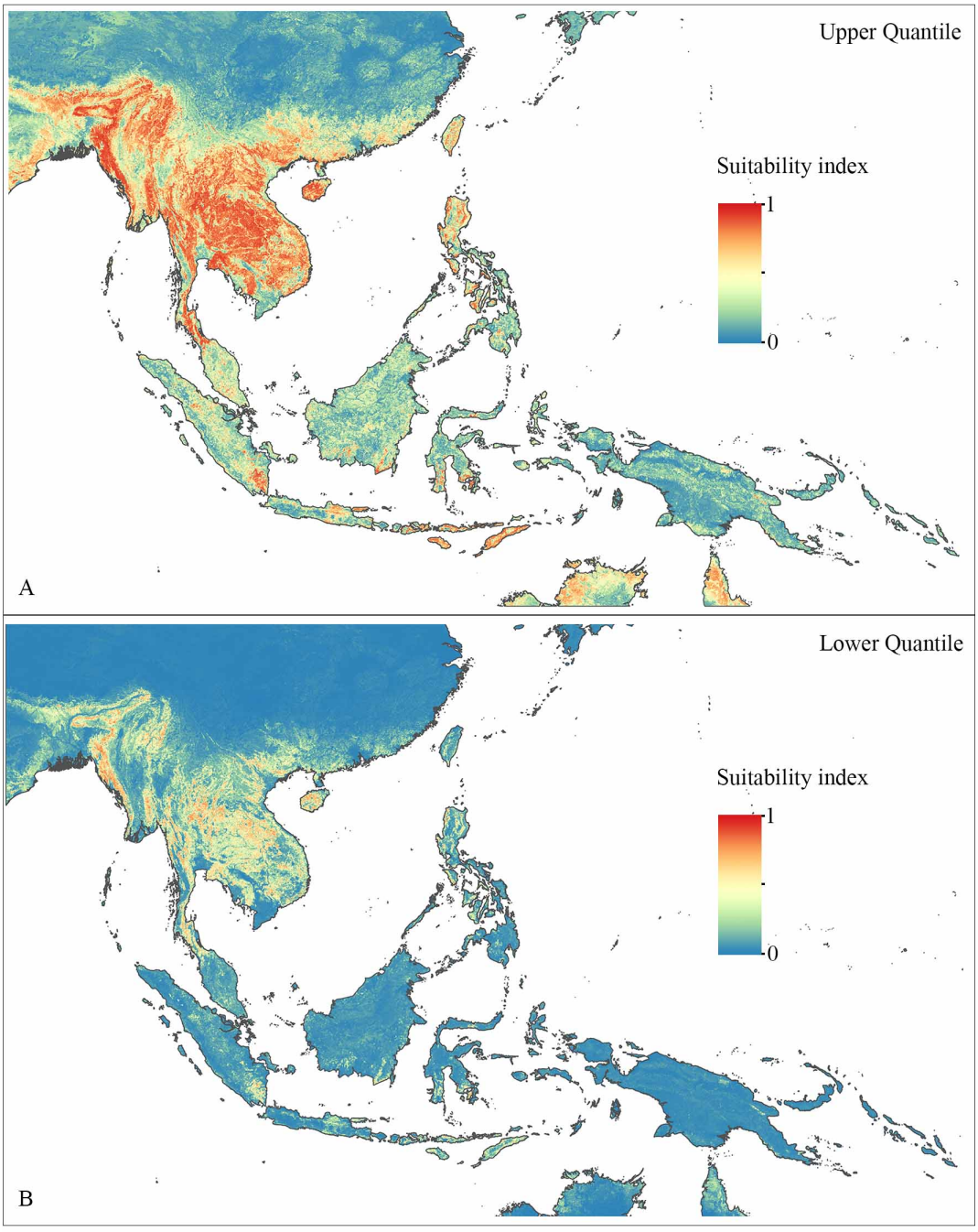


Figure A.8: Uncertainty plots for the Dirus Complex model ensemble. A. Upper (0.975) and **B.** lower (0.025) quantiles for the Dirus Complex model ensemble

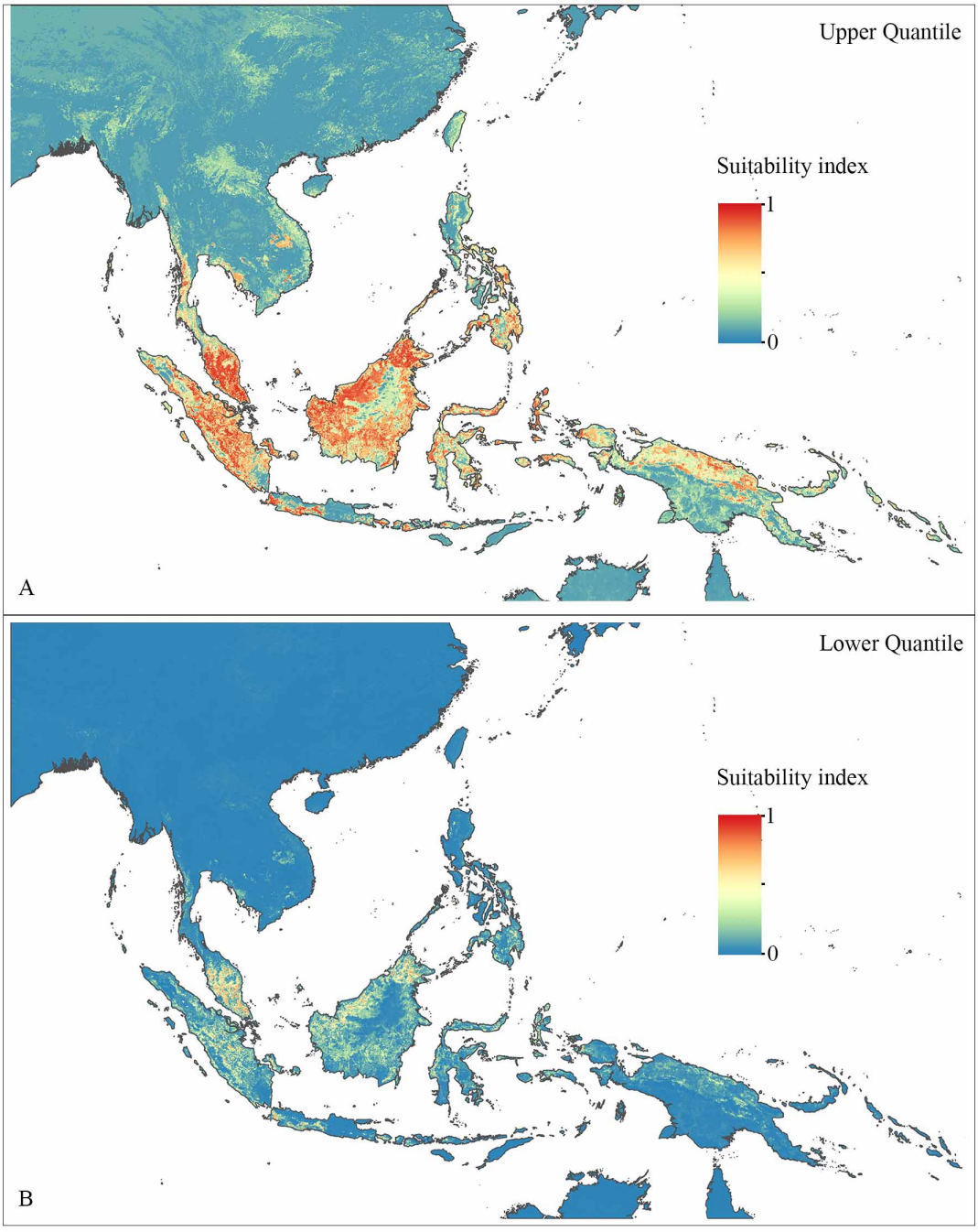


Figure A.9: Uncertainty plots for the Leucosphyrus Complex model ensemble. A. Upper (0.975) and **B.** lower (0.025) quantiles for the Leucosphyrus Complex model ensemble

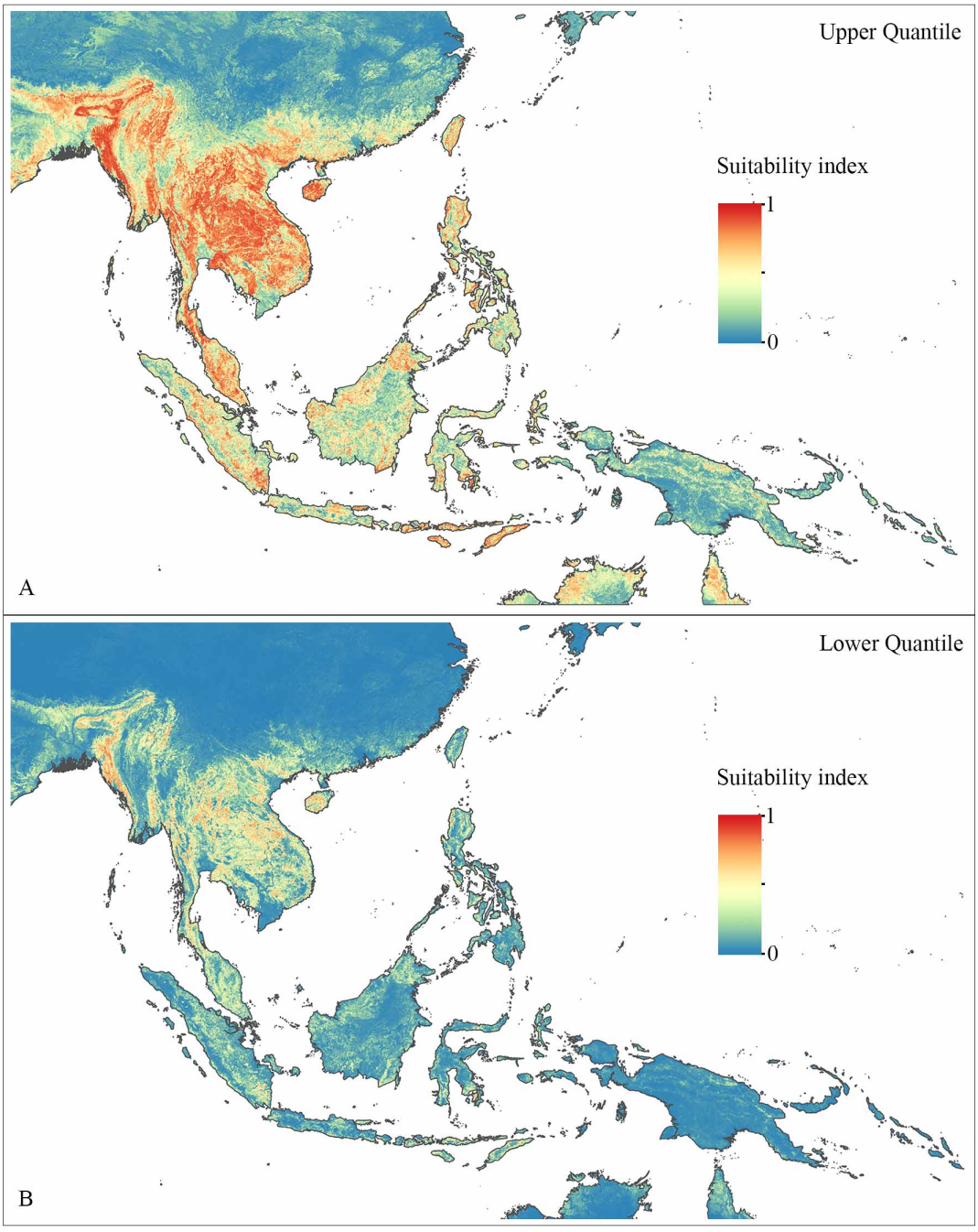


Figure A.10: Uncertainty plots for the Leucosphyrus Group model ensemble. A. Upper (0.975) and **B.** lower (0.025) quantiles for the Leucosphyrus Group model ensemble

A.3 *Plasmodium knowlesi* risk model uncertainty and top predictors

In this section, Figure A.11 displays the zone of zoonotic transmission of *P. knowlesi* malaria based on reservoir and vector species distribution model outputs. Figures A.12 to A.15 represent additional outputs for the *P. knowlesi* malaria risk model.

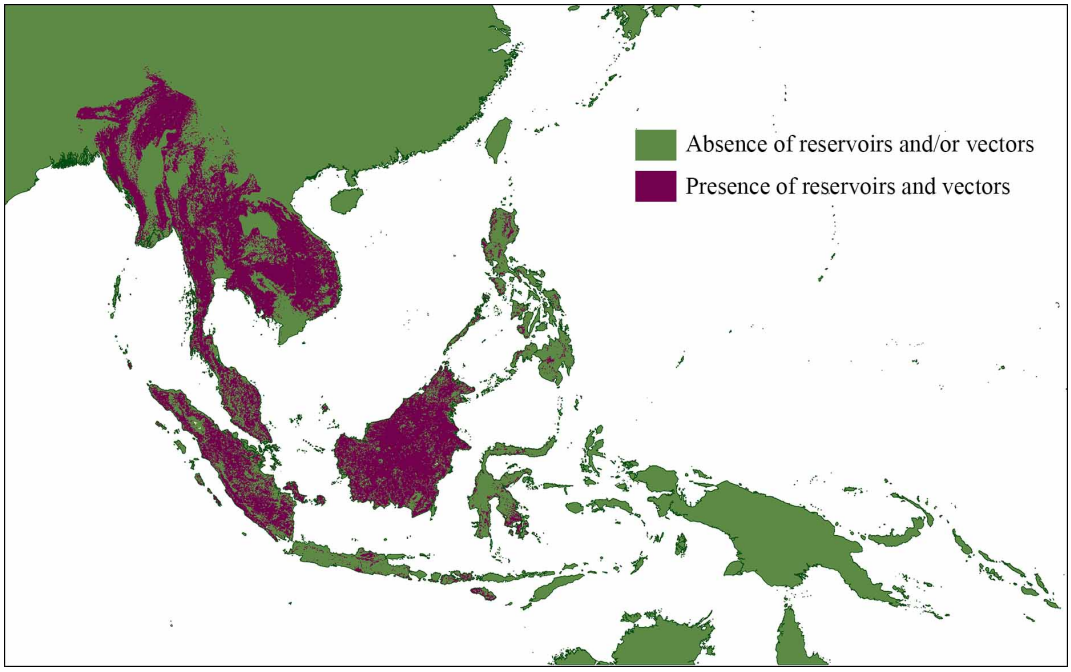


Figure A.11: Fine-scale map of presumed limits of zoonotic *Plasmodium knowlesi* malaria transmission. Indicates areas where *P. knowlesi* reservoir and vector species are predicted to co-occur, and areas where either reservoirs or vectors, or both, are predicted to be absent.

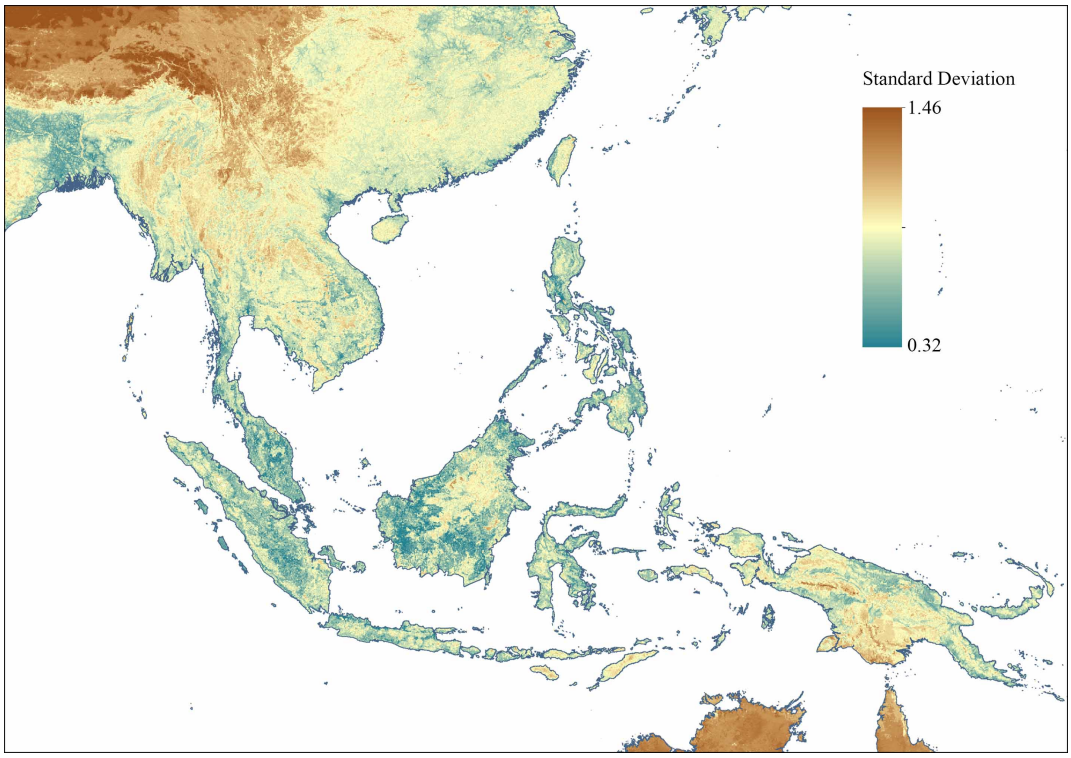


Figure A.12: *Plasmodium knowlesi* malaria risk model uncertainty. Standard deviation values for each pixel were calculated across the model ensemble. Areas from lower to higher standard deviation values are shown.

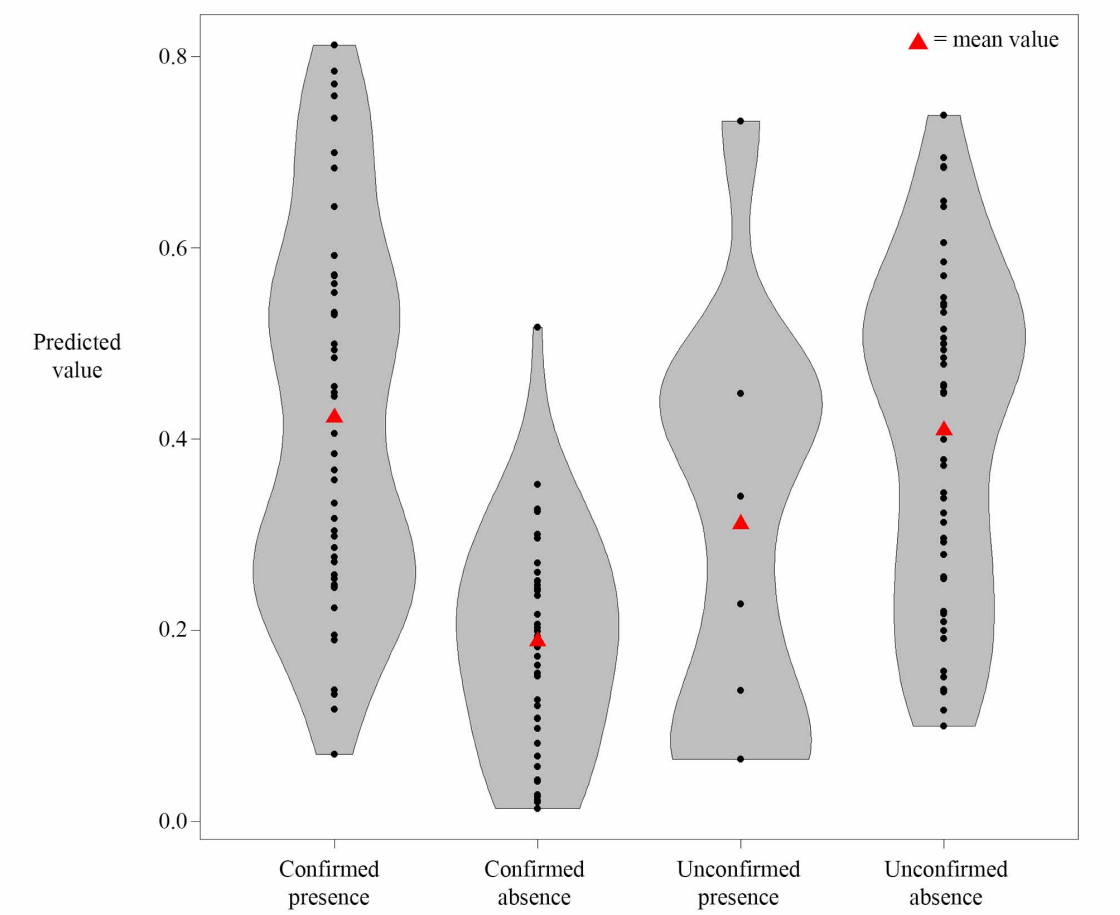


Figure A.13: Predicted disease risk values at locations with confirmed/unconfirmed presence and absence from outside Malaysia, Brunei and Singapore. The black dots represent the predicted values of confirmed/unconfirmed *P. knowlesi* presence and absence points and violin plots showing the density of points at each predicted value are shown in grey. Reports of *P. knowlesi* that were not supported by results from a second independent group working in the same region were classified as unconfirmed.

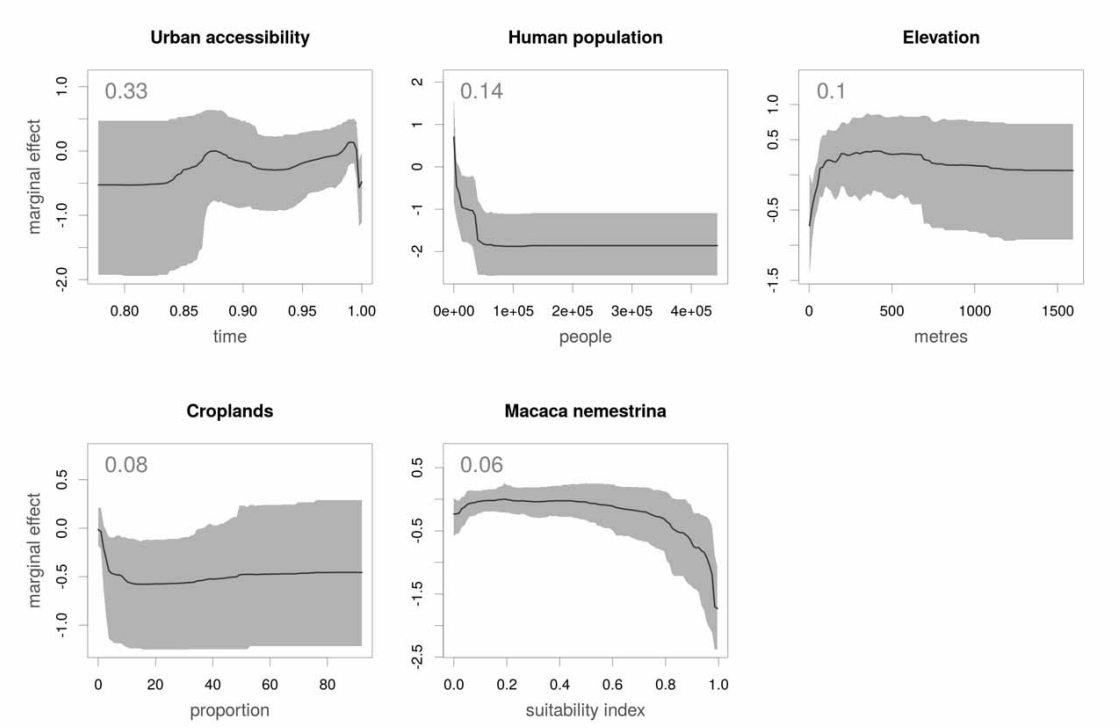


Figure A.14: Marginal effect plots for the most influential covariates for the *Plasmodium knowlesi* malaria risk model. The black line represents the mean marginal effect and grey envelopes the associated 95% quantiles. The mean relative contribution is displayed in the top left corner of each plot. The most influential predictors were defined as those whose relative influence was greater than 1.0 divided by the total number of covariates (since the sum of the relative influence values is 1.0). Marginal effect plots visualise the effect of each variable on the response after averaging the effects of all other variables.

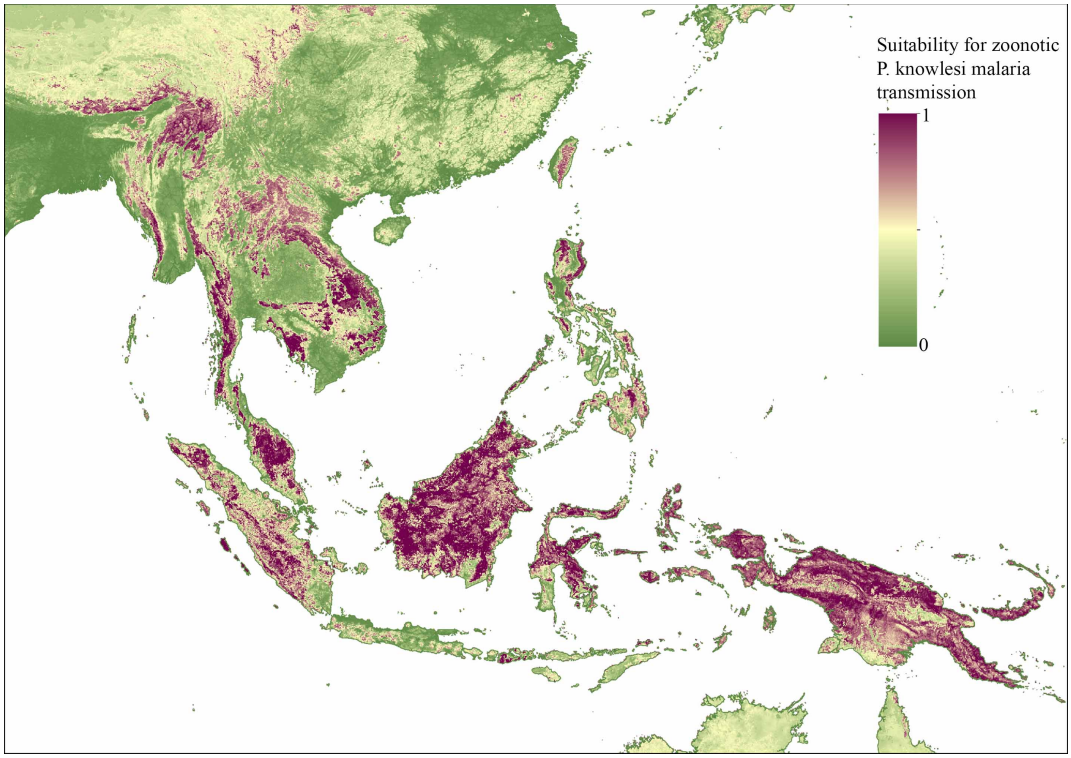


Figure A.15: Unmasked mean *Plasmodium knowlesi* malaria risk model output. Predicted risk of zoonotic *P. knowlesi* infection for a hypothetical scenario where known reservoir and vectors species have expanded their ranges to reach all suitable environments.

B

Yellow fever vaccination coverage

B.1 Yellow fever vaccination coverage by age group 1970-2010

In this section, Figures B.1 to B.5 display the estimated proportion of individuals within each age group that have ever received a YF vaccine (vaccination coverage) for each country at risk of YFV transmission, based on the *untargeted*, *unbiased* vaccination-targeting scenario. Separate plots are shown for the beginning of each decade from 1970 to 2010.

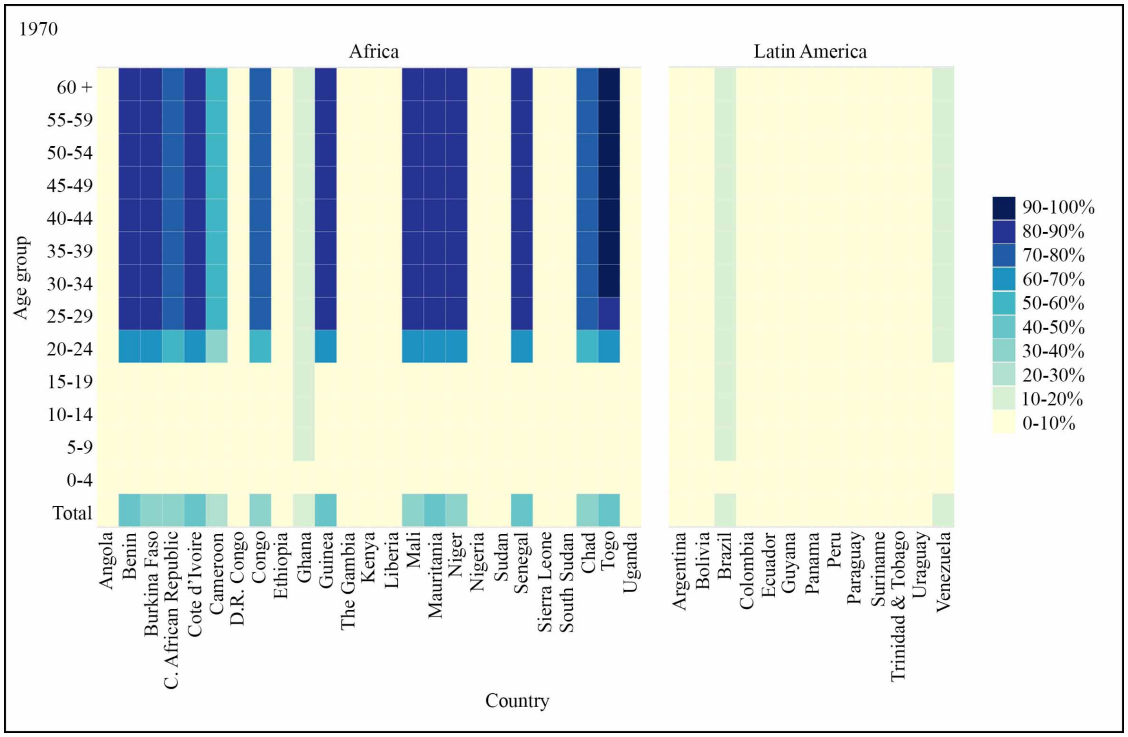


Figure B.1: Estimated yellow fever vaccination coverage by age group in 1970.

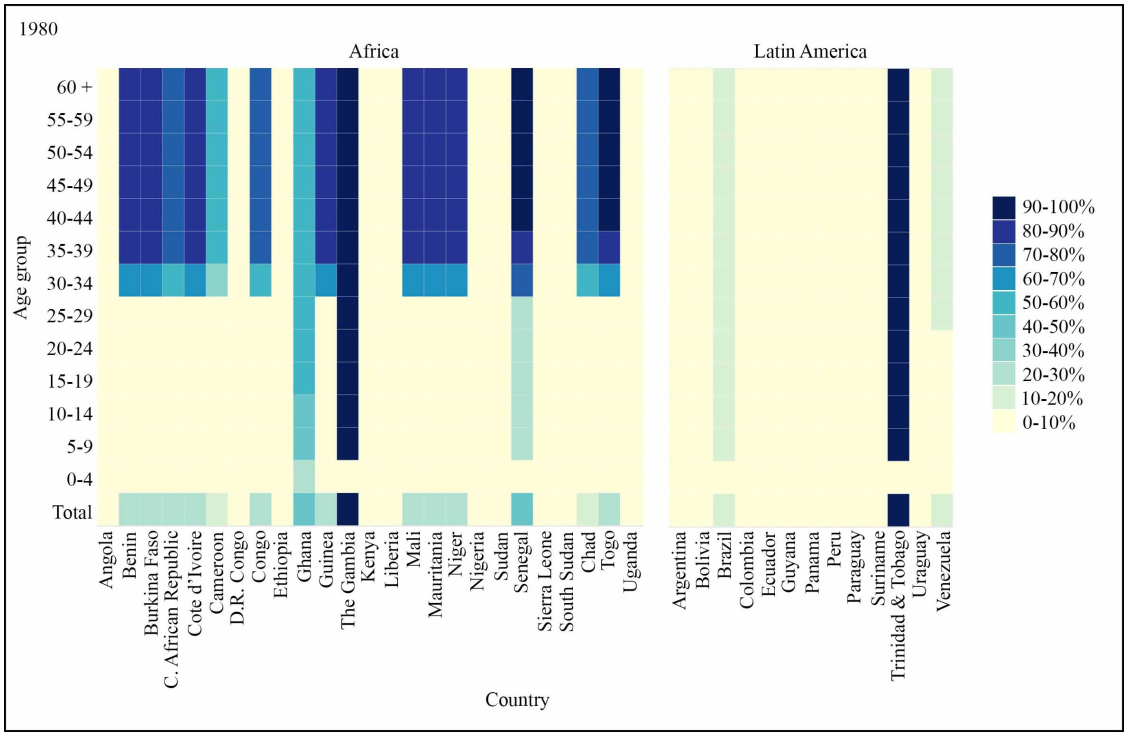


Figure B.2: Estimated yellow fever vaccination coverage by age group in 1980.

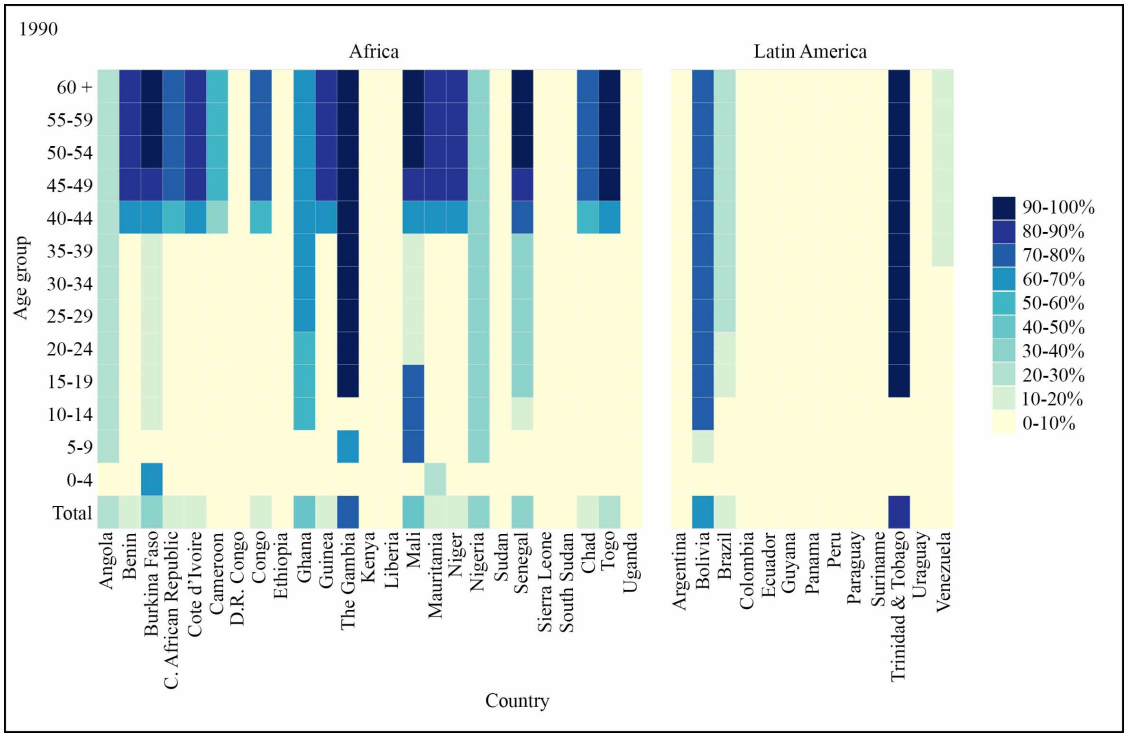


Figure B.3: Estimated yellow fever vaccination coverage by age group in 1990.

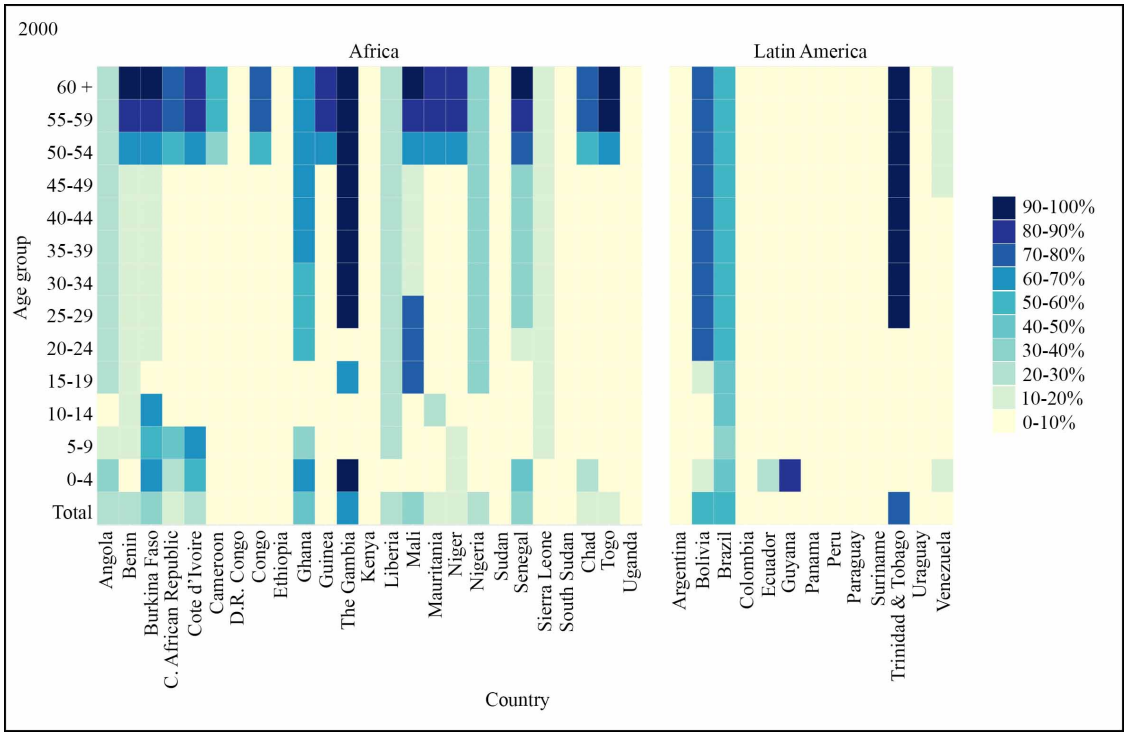


Figure B.4: Estimated yellow fever vaccination coverage by age group in 2000.

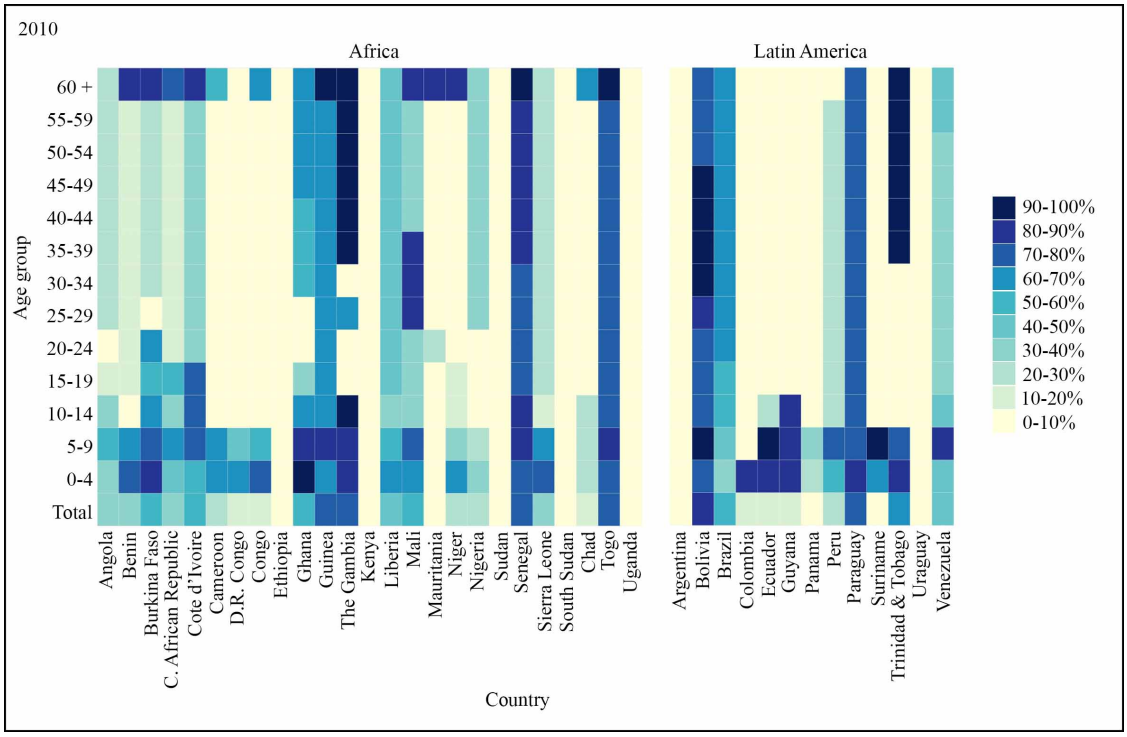


Figure B.5: Estimated yellow fever vaccination coverage by age group in 2010.

C

Modelling the existing and potential global infection risk zones of yellow fever

C.1 Supplementary plots of model inputs and outputs

In this section, Figures C.1 and C.2 display plots of land cover, environmental and biological covariates supplied to the model of YF risk. Figure C.3 displays plots of the environmental covariates supplied to the reservoir species distribution models (excluding those also supplied to the YF risk model). Figure C.4 displays the location of YF model integration points, and Figure C.5 shows average YF vaccination coverage over the study period used to estimate population immunity for YF risk mapping.

Additional model outputs are also included, with marginal effect plots of the most influential covariates for the YF risk model shown in Figure C.6, and model uncertainty plotted in Figure C.7.

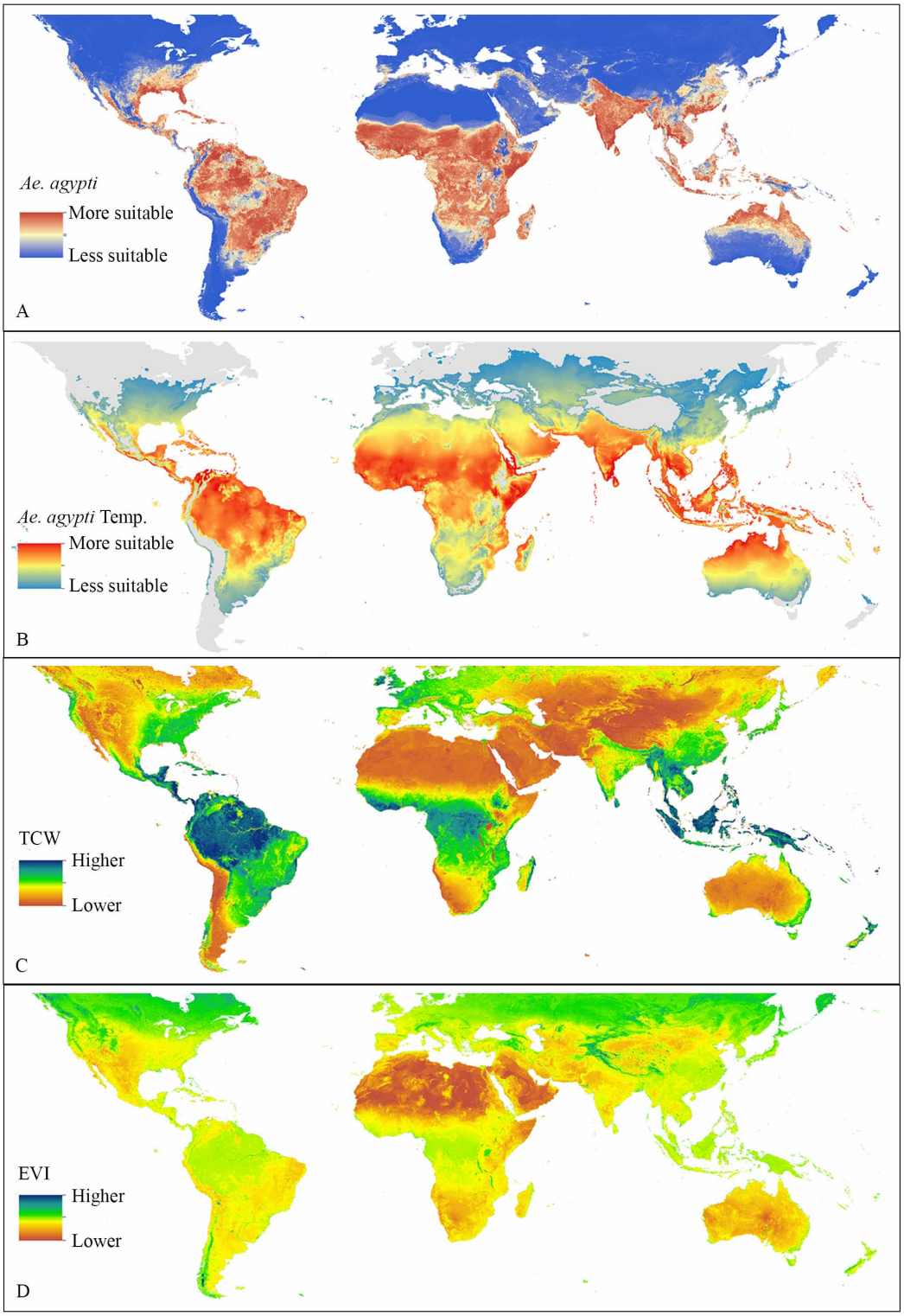


Figure C.1: Environmental and biological covariates supplied to the yellow fever risk model. A. Relative habitat suitability for vector *Aedes aegypti*. **B.** Index of temperature suitability for *Ae. aegypti*. Grey = unsuitable. **C.** Enhanced vegetation index. **D.** tasselled cap wetness - a measure of surface moisture.

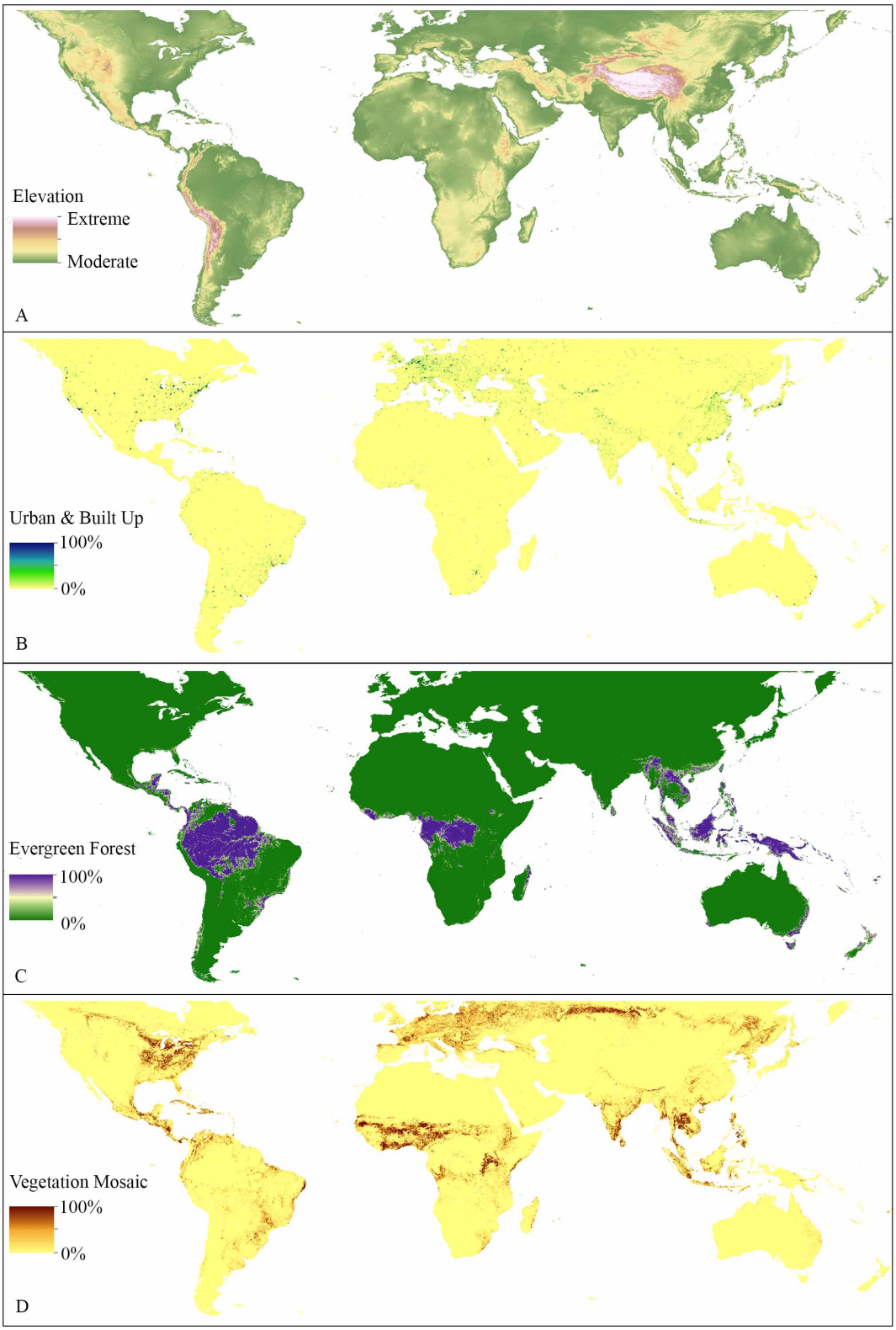


Figure C.2: Land cover covariates (and elevation) supplied to the yellow fever risk model. A. Elevation. **B-D.** Proportional cover for 2012 of urban and built-up landscapes, evergreen broadleaf forest, and cropland/natural vegetation mosaic, respectively.

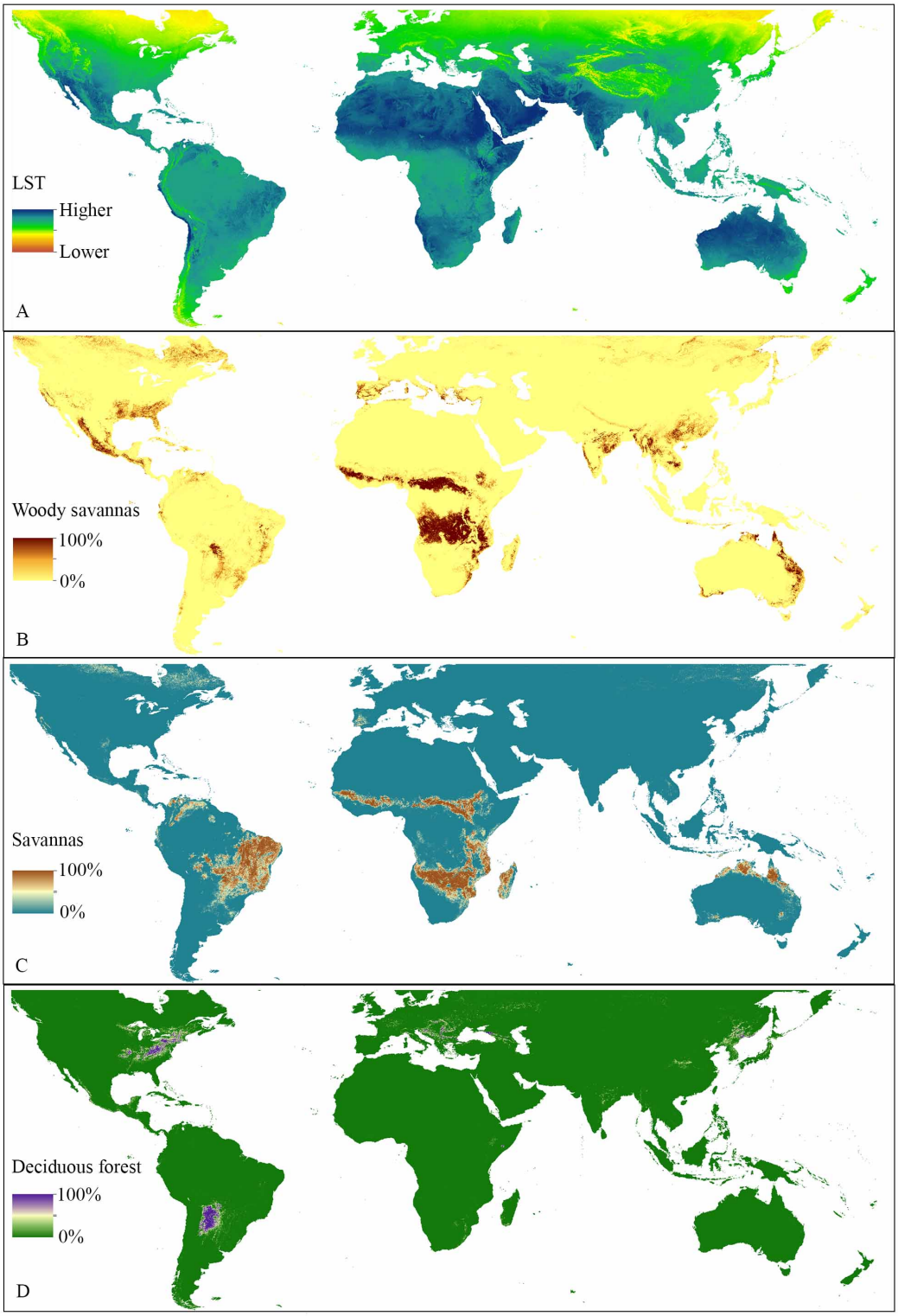


Figure C.3: Environmental covariates supplied to the YF reservoir species distribution models. **A.** Land surface temperature (mean). **B-D.** Proportional land cover for 2012 of woody savannas, savannas, and deciduous broadleaf forest, respectively.

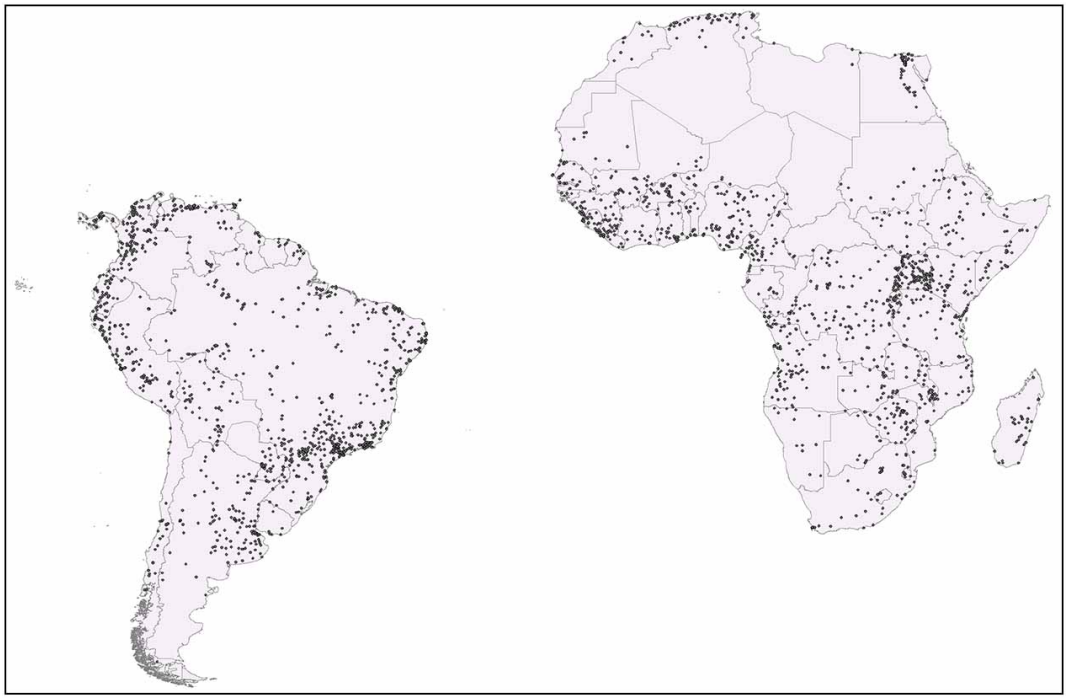


Figure C.4: Yellow fever (YF) model integration points. Location of integration points for YF disease model fitting, representing intensity of disease case reporting.

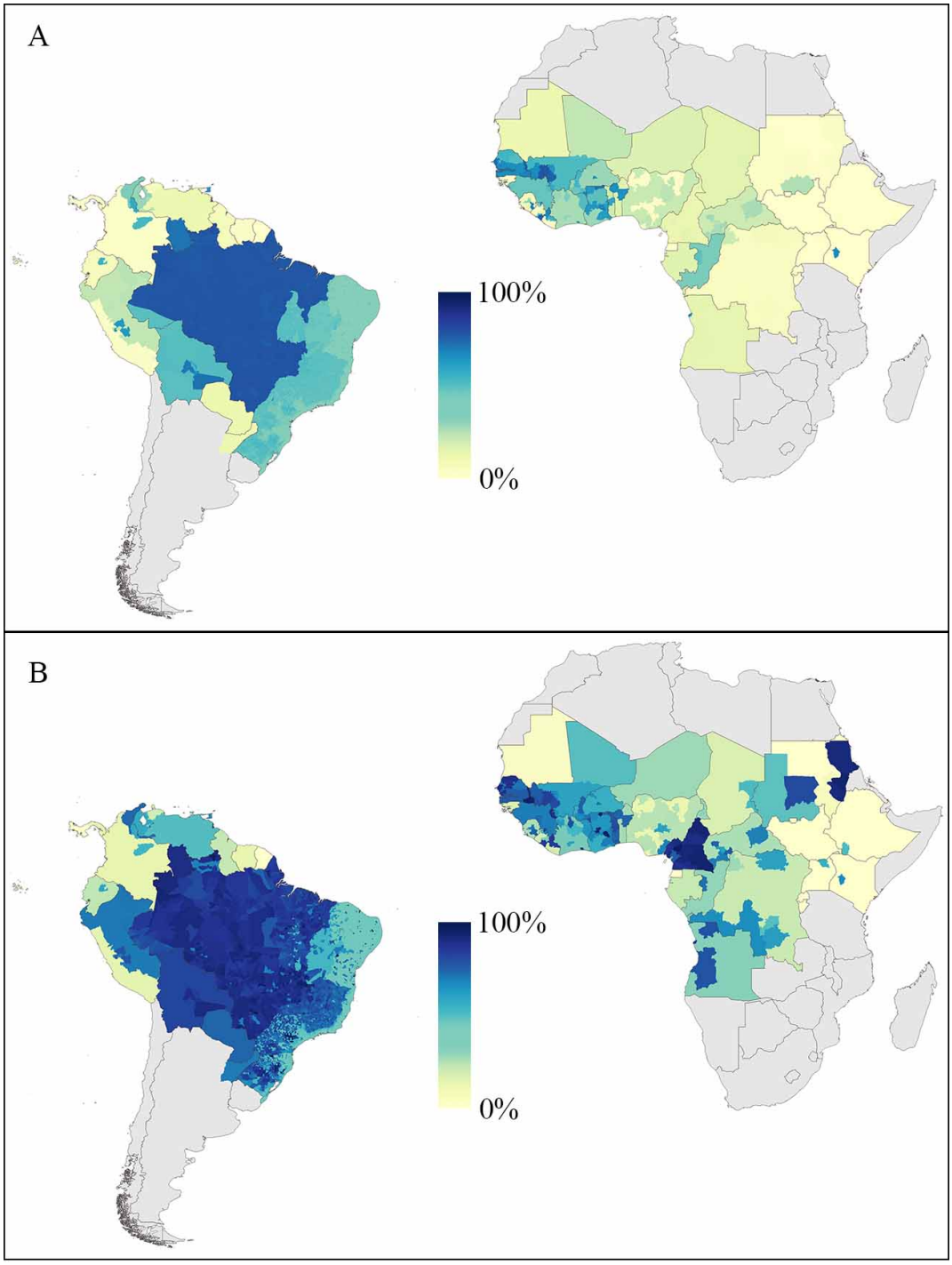


Figure C.5: Average and contemporary yellow fever (YF) vaccination coverage. A. Temporal mean YF vaccination coverage from 1970 to 2016, calculated using time series estimates of vaccination coverage developed in *Chapter 3*. **B.** Estimated YF vaccination coverage for 2016. Both maps are based on the *untargeted*, *unbiased* vaccination-targeting scenario described in *Chapter 3*.

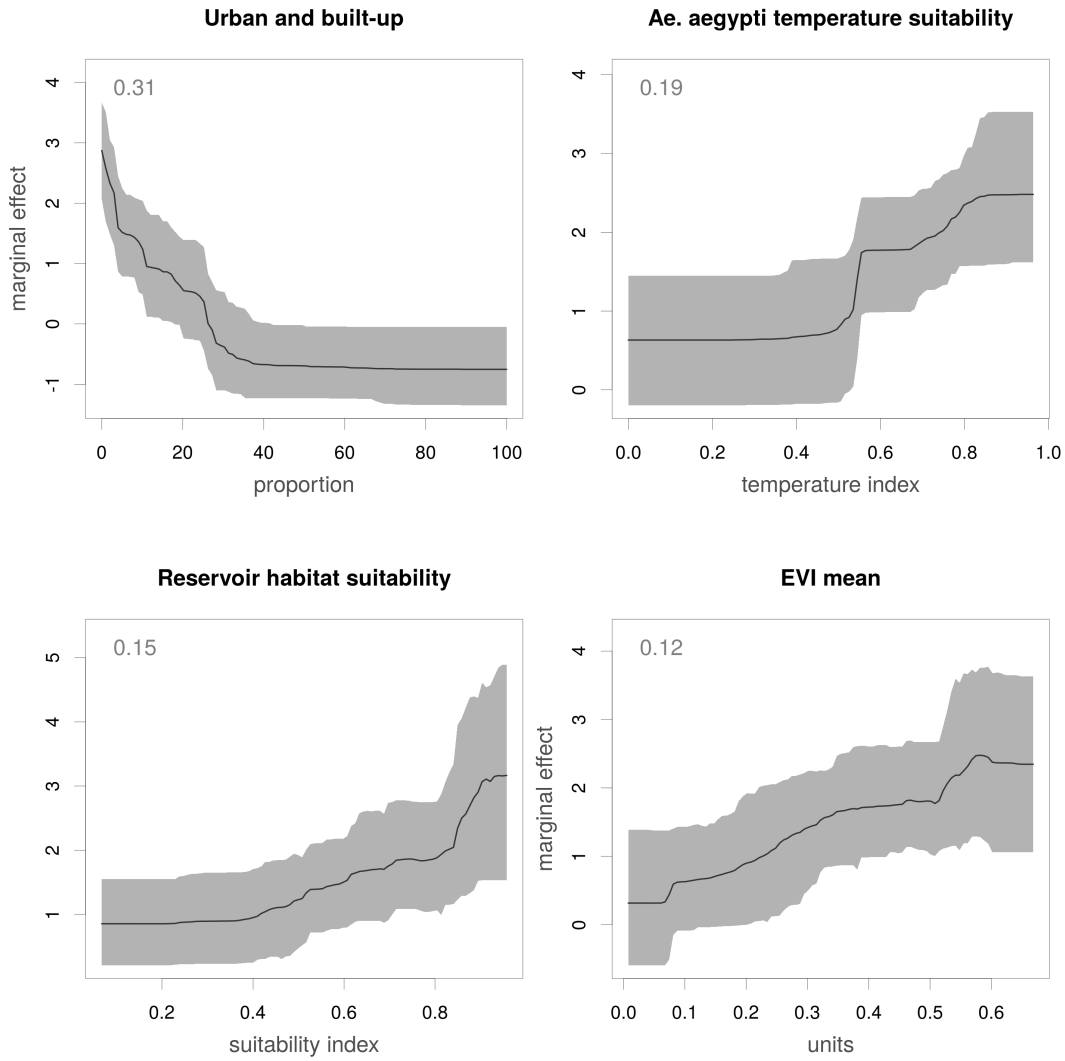


Figure C.6: Marginal effect plots for the most influential covariates for the yellow fever risk model. The black line represents the mean marginal effect and grey envelopes represent the associated 95% quantiles. The mean relative contribution is displayed in the top left corner of each plot. The most influential predictors were defined as those whose relative influence was greater than 100/total number of covariates (since the sum of the relative influence values is 100). Marginal effect plots visualise the effect of each variable on the response after averaging the effects of all other variables.

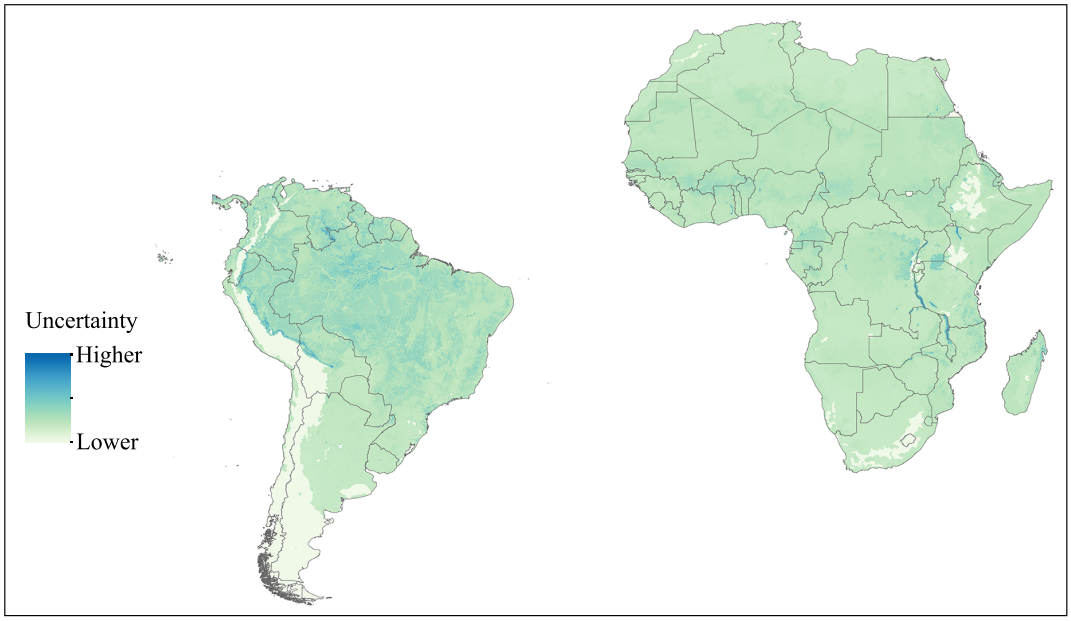


Figure C.7: Yellow fever (YF) model uncertainty. Standard deviation values for each pixel calculated across the YF model ensemble.