Quantifying spatio-temporal variation in malaria transmission in near elimination settings using individual level surveillance data

Isobel Routledge

Supervisors: Dr Samir Bhatt, Dr Azra Ghani

Imperial College London

Department of Infectious Disease Epidemiology

Thesis submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Infectious Disease Epidemiology

Abstract

As countries move towards malaria elimination, tracking progress through quantifying changes in transmission over space and time is key. This information is necessary to effectively target resources to remaining 'hotspots' (high-risk locations) and 'hotpops' (high-risk populations) where transmission remains, decide if and when it is appropriate to scale back interventions, and to evaluate the success of existing interventions. However, as countries approach zero cases, it becomes difficult to measure transmission. Traditional metrics, such as the prevalence of parasites in the population, are no longer appropriate due to small numbers and increasingly focal distributions of cases over space and time.

In order to address this, this thesis developed Bayesian network inference approaches to utilise information about the time and location of cases showing symptoms of malaria to jointly infer the likelihood that a) each observed case was linked to another by transmission and b) that a case was infected by an external, unobserved source. This information was used to calculate individual reproduction numbers for each reported case, or how many new cases of malaria are expected to have resulted from each case. In elimination settings, quantifying the distribution of individual reproduction numbers provides useful information about how quickly a disease may die out, and how the introduction of new cases through importation may affect ongoing transmission. These estimates were incorporated into additive regression models as well as geostatistical models to map how malaria transmission varied over space and time as well as considering timelines to elimination and the likelihood of resurgence of transmission once zero cases is achieved. This approach was applied to previously unanalysed individual-level datasets of malaria cases from China and El Salvador.

Declaration of Originality

I declare that the work presented here is my own work, and all work carried out by others is appropriately referenced or described below:

In Chapter 3, Samir Bhatt provided template R-INLA code which I adapted for the spatial analysis of reproduction numbers and mapping of risk of Rc>1.

In Chapters 4 and 5, Samir Bhatt provided code for implementing the described algorithm in Tensorflow which I adapted and used in my analysis. Manuel Gomez Rodriguez provided advice and described how to derive the survival and hazard functions from my adapted algorithm, which I then re-derived on my own.

Parts of Chapter 1 have been published in an altered form in:

Routledge I., Watson O.J., Griffin J.T., Ghani A.C. (2018**)** Predictive Malaria Epidemiology, Models of Malaria Transmission and Elimination. In: Kremsner P., Krishna S. (eds) Encyclopedia of Malaria. Springer, New York, NY

Chapter 3 has been published in an altered form as:

Routledge, I., Chevéz, J.E.R., Cucunubá, Z.M., Rodriguez, M.G., Guinovart, C., Gustafson, K.B., Schneider, K., Walker, P.G., Ghani, A.C. and Bhatt, S., 2018. Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting. *Nature communications*, *9*(1), p.2476.

Chapter 4 has been published in an altered form as a preprint (bioArxiv DOI: https://doi.org/10.1101/628842)**:**

Routledge, I., Lai, S., Battle, K.E., Ghani, A.C., Gomez-Rodriguez, M., Gustafson, K.B., Mishra, S., Proctor, J.L., Tatem, A.J., Li, Z. and Bhatt, S., 2019. Tracking progress towards malaria elimination in China: estimates of reproduction numbers and their spatiotemporal variation. *bioRxiv*, p.628842.

Declaration of Copyright

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are many people who the work presented here owes a great deal to. I firstly would like to thank Sam for being such a wonderfully supportive supervisor, providing me with insightful guidance when needed, yet putting trust in me and giving me the freedom to take ownership and direct my work. His contagious enthusiasm for machine learning and statistics has gave me fresh perspectives and opened up new avenues for my research. Thanks to Azra, my second supervisor for her astute mentorship, advice and perceptive comments on manuscripts and thesis drafts.

Zulma Cucunuba, Katerina Guinovart, Kammerle Schneider and Patrick Walker provided help with scoping and arranging meetings with the Ministry of Health in El Salvador. Eduardo Romero Chevez and MINSAL in El Salvador shared the data analysed in Chapter 3. I am also grateful to China CDC and Shengjie Lai for their generosity in providing the line-list data analysed in Chapter 4. Bobby Reiner shared the Swaziland data which was used in Chapter 5. The datasets I was given access to represent a huge amount of work and dedication from many people on the ground collecting and collating the data, and the work carried out in this thesis would be impossible without them.

Manuel Gomez Rodriguez, Seth Flaxman, Josh Proctor provided useful discussions which shaped the methods development earlier on in my doctoral studies. I would also like to acknowledge Bryan Greenhouse, Isabel Rodriguez Barranquer and both of their labs at UCSF who hosted me for a month. They were so generous with their time, space and data and, importantly, introduced me to the delights of La Torta Gorda. I also am grateful to Alex Perkins and his lab at Notre Dame for hosting me, being generous with their time and asking insightful questions which improved my research. The malaria group at DIDE has also provided invaluable support and insight throughout my PhD.

This PhD would be much harder (and considerably less fun) without a wonderful support network. I am especially grateful to my Wadham friends, the Roundhouse collective, especially the Throwdown cast, and all the wonderful people in the London houseboating community. I also want to thank my fellow PhD students and friends in the department especially Finlay Campbell, Lucia Cilloni, Amy Dighe, Matt Dixon, Julia Dunn, Lily Geidelberg, James Hay, Joel Hellewell, Robin Schaefer, OJ Watson and Charlie Whittaker. I feel grateful to have been surrounded by such lovely people and be guaranteed to laugh every lunchtime. I have felt so many moments of joy, belonging and support throughout my doctoral studies as a result of these friendships and communities, and am so thankful for them.

Thanks to my parents, Carolyn and Bruce for their unwavering support and acceptance of me and my path in life. I am so grateful for them instilling a sense of curiosity in me, and for taking my thoughts and questions seriously and making me feel they had worth. Finally, thanks to my honorary sister and closest friend, Fez. Thank you for being my biggest supporter, inspiration, and source of belly laughs for nearly ten years.

Abbreviations

ACT Artemisinin-based Combination Therapies

AIC Akaike Information Criterion

BFGS -B Broyden-Fletcher-Goldfarb-Shanno (Bounded) Algorithm

CI Confidence Interval

CQ Chloroquinine

DDT dichloro-diphenyl-trichloroethane

G6PD glucose-6-phosphate dehydrogenase

GAM Generalised Additive Model

GMEP Global Malaria Eradication Programme

GP Gaussian Process

IPTp Intermittent preventive treatment in pregnancy

IRS Indoor Residual Spraying

IVM Integrated Vector Management

LLINs Long Lasting Insecticide Treated Nets

MAE Mean Absolute Error

MCMC Markov Chain Monte Carlo

MINSAL Ministerio de Salud de El Salvador (El Salvadorian Ministry of Health)

MLE Maximum likelihood estimation

PQ Primaquinine

PCR Polymerise Chain Reaction

RACD Reactive Active Case Detection

RDT Rapid Diagnostic Test

RMSE Root Mean Squared Error

SI Serial Interval /Suscebtible – Infected

SIR Susceptible – Infected – Recovered

SMC Seasonal Malaria Chemoprevention

UI Uncertainty Interval

WHO World Health Organisation

1
Introduction

In this chapter, I first review the epidemiology and burden of malaria. Following this, historical and current policy for malaria control and elimination is outlined, including shifts in the prioritisation of total elimination as opposed to control of the disease. The resulting epidemiological considerations which become relevant when aiming for elimination rather than control are then discussed. Next, I provide an overview of common metrics used to measure malaria transmission and the impact of control measures. Following from this, I introduce key mathematical models of malaria transmission dynamics and discuss how they have been used to estimate metrics of transmission and provide insights into the impact of interventions. I then discuss the potential for mathematical models to assist elimination efforts and the epidemiological challenges faced in elimination settings. Continuing the theme of measuring malaria in elimination contexts, I introduce the approaches developed in recent years in outbreak analysis to measure transmission using individual level surveillance data, and discuss how they may apply to endemic diseases in near elimination and elimination settings, with comparisons of such approaches to models developed within the machine learning and data science community. Finally, the problem and aims of the thesis are introduced.

1.1 Natural history of malaria

Malaria is a disease caused by *Plasmodium* parasites, spread to humans through the bites of mosquitoes in the *Anopheles* genus. There are 6 known species of malaria parasites which infect humans, of which *Plasmodium falciparum* and *Plasmodium vivax* are of most concern from a public health perspective. However, there has been growing concern about zoonotic *Plasmodium knowelsi* as this species has been associated with severe outcomes in certain areas of South East Asia (Millar and Cox-Singh, 2015). It is estimated that 70 *Anopheles* species are known to be capable of transmitting malaria to humans (Warrell and Gilles, 2002), of which 41 species have been identified as dominant vector species, meaning they are sufficiently competent vectors of the disease to be relevant to public health (Hay *et al.*, 2010). These different *Anopheles* species have various geographical distributions, vectoral capacities and ecologies which affect malaria epidemiology, transmission and the efficacy of interventions targeting them.

1.1.1 Malaria lifecycle

Plasmodium parasites present in the blood of the human host are ingested by the mosquito during blood meals as male and female gametocytes (Figure 1.1). Sexual reproduction occurs in the mosquito midgut whereby the male and female gametes fuse, producing a zygote. This zygote then elongates and becomes a motile ookinete, travelling to the mosquito mid-gut to develop into an oocyst. After a sporogonic period of approximately 8-10 days, the oocysts burst to release sporozoites which travel to the salivary glands of the mosquito, where they are passed to the mammalian host through the next blood feed. The sporozoites rapidly migrate through the blood to the liver where they invade hepatocytes which subsequently group together to form schizonts. Merozoites develop within the schizonts and, approximately 7-10 days after initial infection, merozoites are released into the bloodstream. During the blood-stage infections, merozoites infect the red blood cells (erythrocytes) where asexual reproduction occurs. This invasion of red blood cells, which occurs with a periodicity of 24-36 hours, produces most of the negative disease related outcomes seen in individuals with clinical malaria (Ménard *et al.*, 2013). After a period of approximately 10 days, a subset of the asexual parasites differentiate into the gametocyte (male and female) stages, which are ingested by a female mosquito and completing the cycle.

The lifecycle of *P. vivax* differs slightly to the *P. falciparum* lifecycle in several ways. One of the most epidemiologically relevant differences is the ability for *P. vivax* to lie dormant in the liver of infected individuals as hypnozoites, which can cause relapses in blood stage malaria by re-invading the bloodstream. This can occur several weeks (short relapse) or months (long relapse) after initial infection and the relative incidence of both relapse lengths varies between temperate and tropical regions, with short-relapse patterns generally occurring in tropical areas and long relapse patterns occurring in temperate areas in relation to seasonality (Battle *et al.*, 2014; White *et al.*, 2016).

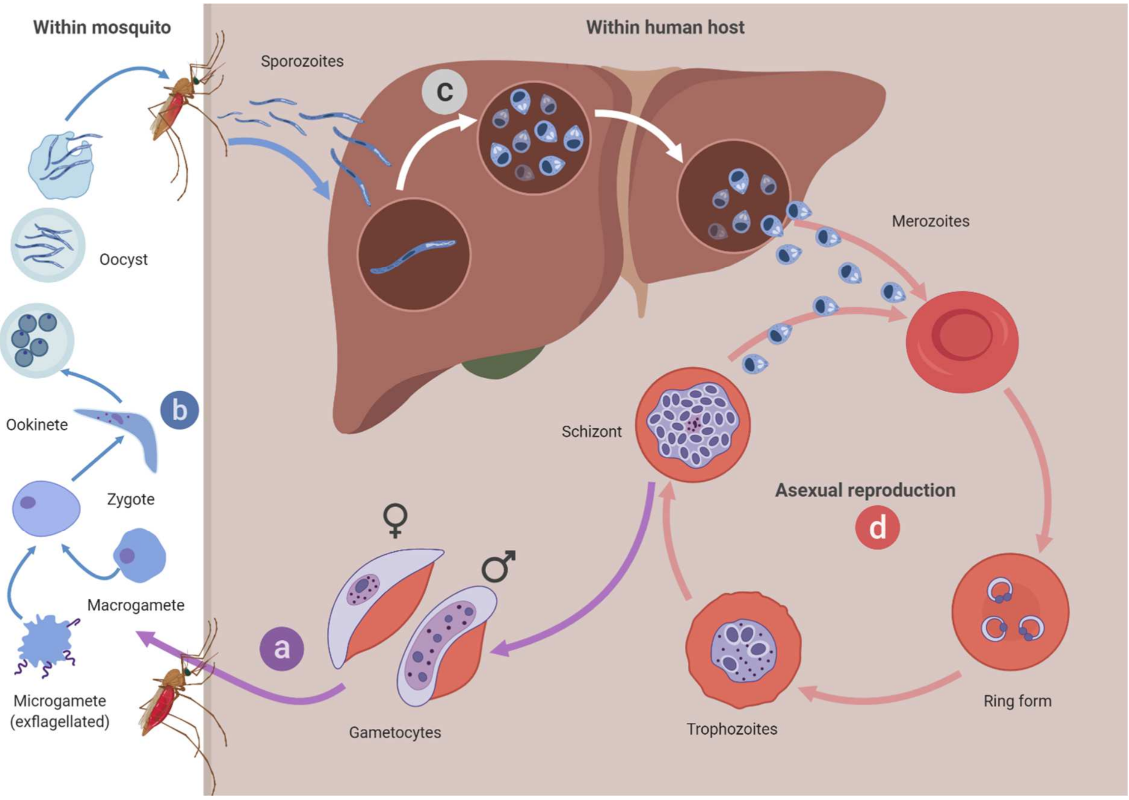


Figure 1.1 Plasmodium falciparum lifecycle *showing a) Gametocyte production and ingestion during a bloodmeal. b) sexual reproduction and developmental stages within the mosquito c) inoculation of sporozoites and liver stage infection within the human host d) asexual reproduction and blood stage infection. Drawn using BioRender (www.biorender.com)*

Malaria infections can vary in their severity and in their impact on the lives of those infected. Clinical, symptomatic malaria is initially characterised by fever, aches and chills. If left untreated, the patient can progress to severe malaria, resulting in severe anaemia, respiratory problems, or cerebral malaria, all of which can result in death. Under 5s are particularly at risk (WHO, 2018a). In areas where there is still a high burden of malaria, it is common for naturally acquired immunity to develop following repeated exposure to malaria and for asymptomatic infections to occur.

The age distributions of malaria infection also vary by species and transmission intensity. In high transmission settings, many members of the population become exposed and infected at a young age. As a result, they develop an immune response to malaria, meaning the age profile of individuals with clinical symptoms and often the prevalence of asymptomatic parasitaemia is higher in these settings. When naturally acquired immunity is not present in the population (due to reduced exposure) individuals are often older when exposed, and the age profile is a wider range of ages. Similarly, lower population level immunity means individuals exposed to malaria, for example through imported cases, are more likely to become infected and be at risk of clinical disease.

1.2 Burden and epidemiology of malaria

According to the World Health Organisation (WHO) World Malaria Report 2018 (WHO, 2018a), an estimated 219 million cases (95% Confidence Interval (CI) = 203 – 262 million) of malaria occurred globally in 2017. Of these, 3.4% are attributed to *P. vivax*, with the rest attributed to *P. falciparum*. However, this proportion reached 74.1% and 37.2% in the WHO Americas and South East Asia regions respectively. In the same year the disease was responsible for an estimated 435 000 deaths worldwide, with the majority of mortality (93%) occurring in the WHO Africa region, and in children under 5 years old (61%). An estimated 99% of the mortality was caused by infection with *P. falciparum*. In 2017 92% of all malaria cases were thought to have occurred in the WHO African region, with 5% occurring in the WHO South-East Asia Region, 2% occurring in the WHO Eastern Mediterranean region, and less than 1% occurring in each of the remaining WHO areas (WHO, 2018a). However, there is a significant burden of *P. vivax* outside of sub-Saharan Africa. *P. vivax* has a much wider geographic range than *P. falciparum*, in part due to its ability to lie dormant, allowing transmission to be sustained following seasons which are unsuitable for vectors (Howes *et al.*, 2016; Battle *et al.*, 2019; Weiss *et al.*, 2019). In contexts outside of sub-Saharan Africa, *P. vivax* often appears to persist when control measures reduce the burden of *P. falciparum*.

1.3 History of malaria control and elimination policy

In 1955, the WHO launched a global campaign with the aim of eradicating malaria globally, known as the Global Malaria Eradication Programme (GMEP). This decision was made following promising results from pilots using dichloro-diphenyl-trichloroethane (DDT) to kill malaria vectors and results from mathematical models by Ross and MacDonald demonstrating the value of adult vector control on reducing malaria transmission (MacDonald, 1956; MacDonald, 1957). The main control measure used during this time was indoor residual spraying with DDT and other insecticides. Moderate successes were achieved; 37 countries eliminated malaria, some of which without evidence of resurgence in the decades following the end of the programme (Nájera, González-Silva and Alonso, 2011). However, many areas where transmission had been interrupted did observe resurgences during the 1960s, as well as transmission occurring in areas where it had not previously occurred (Chiyaka *et al.*, 2013; Smith *et al.*, 2013). As the GMEP programme did not achieve its aims for eradication, it was abandoned in 1969. This resulted in a decline in funding for malaria control and elimination. During this time many countries experienced economic and political challenges, insecticides became more expensive and attention was shifted away from malaria. There was also increased exploitation of countries' natural resources, with mining, logging and other forms of land-use change increasing (Nájera, González-Silva and Alonso, 2011). A systematic review found that the resurgences seen during this time were strongly correlated with reductions in IRS control in Latin America, civil/cold war in Europe and Asia and generally with weakening or cessation of control programmes (Cohen *et al.*, 2012).

In 2000, African leaders gathered at the Roll Back Malaria Summit in Abuja to sign a declaration which committed to halving malaria mortality by 2010 (Global Partnership to Roll Back Malaria, 2000). This represented a strong commitment from the global health community and leaders of malaria endemic countries to investing in malaria control. In the same year, the Bill and Melinda Gates Foundation was formed, which made a strong financial and political commitment to malaria elimination and eradication (Roberts and Enserink, 2007). Renewed commitment to malaria elimination also was reflected in the Millennium Development Goals, with the aim of halting and beginning to reverse the global incidence of malaria by 2015 (United Nations, 2015). This political and financial investment has resulted in clear successes in malaria control. Since 2000, great strides have been made in reducing malaria incidence, prevalence, mortality and progress towards elimination. There was an estimated 41% reduction in global malaria incidence between 2000 and 2015, with a reduction of 21% between 2010 and 2015 (WHO, 2018a). Between 2000 and 2015 an estimated 1.2 billion cases and 6.2 million deaths have been averted, and the global malaria incidence rate has fallen by an estimated 37% (WHO, 2018a). However, these gains have been found to have stalled in recent years. Although the global incidence rate of malaria declined by

18% from 72 to 59 cases per 1000 at risk between 2010 and 2015, this lower incidence rate remained constant in 2016 and 2017, with no further reductions achieved (WHO, 2018a).

In 2016, the WHO listed 21 countries who aimed to eliminate malaria by 2020 (WHO, 2016). The feasibility of eradication remains controversial, and whilst there has been renewed optimism in achieving elimination at country level in many parts of the world (Feachem, Phillips and Targett, 2009; Mendis *et al.*, 2009; Tatem *et al.*, 2010), the most recent report from the WHO Strategic Advisory Group on Malaria Eradication, concluded that eradication will not be possible using current tools, and significant investment in new tools, strengthened healthcare systems and improved surveillance and response will be required in order to reach eradication. Nonetheless in many spheres of malaria governance and policy, eradication remains the ultimate goal (Feachem *et al.*, 2019). This history is important in understanding elimination goals in the present day. Between 1987 and 2007, no countries were certified as eliminated (WHO Global Malaria Progamme, 2016). However, since 2007, eleven countries have been certified by the WHO as having eliminated malaria: Algeria, Argentina, Armenia, Maldives, Morocco, Kyrgyzstan, Paraguay, Sri Lanka, Turkmenistan, United Arab Emirates and Uzbekistan. Two additional countries, El Salvador and China, reported no locally acquired cases in 2018. However, there are many parts of the world, where new tools and new strategies will be required to reduce transmission and further areas such as Venezuela (Grillet *et al.*, 2019) where resurgences in transmission have been observed.

1.4 Principal methods of malaria control

Malaria control and elimination interventions can be broadly divided into anti-malarial measures and anti-vectorial measures. Current WHO policy (WHO, 2015) for treatment of uncomplicated *P. falciparum* (apart from pregnant women in their first trimester, who are recommended 7 days of quinine and clindamycin) is 3 days of with Artemisinin-based combination therapies (ACTs). For uncomplicated *P. vivax*, chloroquine or ACTs are recommended, unless the area is known to have chloroquine resistance, in which case ACTs alone are recommended. If GP6D deficiency status of patients is known, they may be given primaquine to reduce the chance of *P. vivax* relapse. For severe malaria, intravenous artesunate is recommended, followed by standard ACT treatment.

The two principal vector control strategies employed are through the distribution of long lasting insecticidal nets (LLINs), commonly known as bed nets, and where appropriate Indoor Residual Spraying (IRS), delivered as part of integrated vector management (IVM) programmes . The massive scale up of LLINs distribution since 2000 (2% of children slept under a bed net in 2000 compared to 68% in 2015, however this has stalled in recent years (WHO, 2018a)) has been attributed as a major contributor to observed declines in malaria prevalence and incidence over the past 15 years (Bhatt *et al.*, 2015). Nonetheless, the effectiveness of vector control varies by mosquito species and context. A particular challenge to vector control where outdoor-biting vector species are present, as they are not affected by IRS and LLINs, the mainstay of vector control. In these contexts there has been development of interventions such as spatial repellents and attractive-toxic sugar baits (Beier *et al.*, 2012; Maia *et al.*, 2018), however the effectiveness of such interventions has not been conclusively demonstrated and therefore not recommended by WHO at the current time (WHO, 2019a). Resistance to insecticide is also a concern, with evidence that between 2010 and 2016, the frequency of pyrethroid resistance increased in frequency by 32% *An. funestus s.l* and by 13% in *An. gambiae s.l.* (WHO, 2018b; WHO, 2019a).

There has also been a great deal of investment in developing a vaccine for malaria. The furthest along of these vaccines, RTS,S, has begun pilots in select sites in Ghana, Kenya and Malawi after stage 3 clinical trials found four doses of the vaccine had an efficacy of 39% against clinical malaria over the 4 years that patients were followed post vaccination - leading to an estimated 1,774 cases of malaria averted per 1,000 children vaccinated (RTS,S Clinical Trials Partnership, 2015). Given these levels of efficacy, the vaccine has potential as a tool to complement rather than replace existing control strategies such as LLINs and IRS.

Preventative chemotherapies have also been introduced as control interventions in select contexts and in high-risk groups. Seasonal Malaria Chemoprevention (SMC) and Intermittent Preventive Treatment in pregnancy (IPTp) both are techniques in which anti-malarial drugs are delivered over a defined period of time, either the transmission season in the case of SMC, or for the duration of pregnancy in IPTp (WHO, 2015; WHO, 2017a). Mass drug administration (MDA) is also recommended for implementation in select contexts – namely in elimination settings to interrupt transmission of *P. falciparum* malaria, in the Greater Mekong subregion to prevent spread of spread of multi-drug resistance, during malaria epidemics, and in exceptional complex emergencies (WHO, 2015).

In elimination settings, there is a shift towards surveillance, rapid detection and treatment of cases, identification and targeting of foci and in some contexts continued vector control (WHO, 2016). There have been several targeted approaches which have been piloted included treating individuals within the same household, vector control in surrounding areas and in some contexts targeting villages with intensive interventions and monitoring. Where importation is an issue, border surveillance and screening are also key.

1.5 Methods to measure malaria transmission and the impact of control

1.5.1 Metrics of malaria transmission

Malaria transmission varies greatly between populations, demographic groups, over space, over time and in response to control measures. A wide variety of methods and measures (Table 1.1) have been developed to quantify, understand and predict these differences in transmission.

The Entomological Inoculation Rate is the measure of the rate of infectious bites received per person. The EIR, long used as a key measure of malaria transmission is a measure of exposure of humans to infectious mosquitoes. It is defined as the number of infectious bites a human receives over a given time period, and is the product of the human biting rate, *ma* and the sporozoite prevalence, Z/M. Whilst traditionally one of the mainstays in measuring transmission, the EIR can vary greatly between contexts and seasons, and requires large sample sizes and time-consuming sampling methods to get accurate measures. The Force of Infection is the rate at which individuals become infected. This can be determined from transmission models, however, increasingly has been estimated by fitting catalytic models to antibody data. The proportion of a surveyed population with malaria parasites in blood at a given time and location, PrPf/PrPv, is a mainstay of malaria measurements and widely collected through cross sectional surveys. It has been incorporated into global maps which have been key in risk mapping for malaria and tracking declines in the burden of malaria transmission (Bhatt *et al.*, 2015; Battle *et al.*, 2019; Weiss *et al.*, 2019).

Incidence, or the number of cases occurring per 1000 over a given time period is also used. Two key metrics which are measures of incidence. Annual Parasite Index is an annual measure of incidence and is the annual sum of cases occurring per 1000 individuals at risk in a given location over a year. Due to the seasonality and heterogeneity in malaria incidence which can occur, by looking at year on year trends, some of the heterogeneity is removed compared to looking at more fine grain measures of incidence, however of course this level of aggregation also removes useful or interesting patterns and given that seasonality and heterogeneity is a key part of malaria epidemiology, is not as useful for designing the timing of intervention deployment for example.

The Slide Positivity Rate (SPR) is the proportion of malaria-positive slides of all slides examined. This value can be biased if the individuals sampled are not representative of the population as a whole, which is generally the case if being used to diagnose fevers, however if the sampling strategy remains constant year on year, SPR is a useful metric for measuring changes in transmission over time.

The basic reproduction number, R_0 ,or how many secondary cases expected from an index case in a well-mixed, fully susceptible population is another key metric, which will be discussed in more detail in relation to mathematical modelling of malaria. When above one, we expect transmission to continue, and when below one we expect to die of its own accord, however this may take long periods of time if there is importation, or if there are some individuals with reproduction numbers above one even if the mean is below one.

In many situations we may be interested in measuring transmission in the context of ongoing interventions or in a population with some existing immunity, and therefore not fully susceptible. In such cases the effective reproductive number, *R*, is calculated. In some cases R_c is used to denote R under control measures, however this is not used in this thesis as R_c is also used to represent the case or cohort reproductive number - how many cases on average will a case infected at time *t* go onto infect.

Table 1.1: Metrics used to measure malaria transmission

Notation	Measure	Definition
R_O	Basic reproduction number	How many secondary cases expected from an index case in a well-mixed, fully susceptible population
R	Effective reproduction number	Reproductive number when assumption of fully susceptible population is not met
R_C	Case or cohort reproduction number	For a given case or cohort infected at a given time, how many people they go onto infect
EIR	Entomological Inoculation Rate	Rate of infectious bites per person
FOI	Force of infection	The rate at which individuals become infected
$PfPr/PvPr$	Parasite prevalence	The proportion of a surveyed population with malaria parasites in blood at a given time and location
API	Annual Parasite Index	The number of malaria-positive patients per 1,000 inhabitants
SPR	Slide Positivity Rate	The proportion of malaria-positive slides of all slides examined

1.5.2 Mathematical models of malaria transmission

Mathematical models of malaria transmission are tools which provide insight into the dynamics of malaria transmission and assist in the design and evaluation of malaria control and elimination programs. They range from simple sets of equations, through to complex individual-based simulations. Models also have provided key metrics to quantify transmission and progress towards elimination, such as the basic reproduction number. In this section I will summarise the history of malaria transmission models and then explore their potential contribution to elimination efforts, highlighting the epidemiological challenges for malaria in elimination and very low transmission settings (where prevalence is under 1%, (WHO, 2017b)).

1.5.3 Looking back: malaria transmission models in the 20th century

The first mathematical model of malaria transmission was published in 1908 by Ronald Ross after being tasked with recommending methods for the prevention of malaria in Mauritius (Ross, 1908). This model was based on an *a priori* description of how the prevalence of malaria was causally related to the ratio of mosquitoes to humans, *m*. Ross used the model to argue that only a proportion of a mosquito population would need to be killed to prevent transmission, which led to the formulation of a critical mosquito density, *m'*, above which transmission would be sustained. The parameters involved (summarised in Table 1.2) have now been standardised (Smith *et al.*, 2012): *m* is the ratio of mosquitoes to humans, *a* is the proportion of mosquitoes that feed on humans each day, *b* is the proportion of bites by infectious mosquitoes that infect a human, *c* is the probability a mosquito becomes infected after biting an infected human, *r* is the daily rate each human recovers from infection, *v* is the number of days from infection to infectiousness in the mosquito, and *g* is the instantaneous death rate, which also can be expressed as $-\ln p$, where *p* is the probability of an adult mosquito surviving one day, to give the following interpretation of Ross' formula:

$$m' > \frac{gr}{a^2bce^{-gv}}$$

Table 1.2: Parameters in Ross MacDonald model

Parameter	Definition
$\boldsymbol{m}$	The ratio of mosquitoes to humans
$\boldsymbol{a}$	The rate at which a mosquito takes human blood meals
$\boldsymbol{b}$	The probability that a bite by an infectious mosquito infects a human
$\boldsymbol{c}$	The probability a mosquito becomes infected after biting an infected human
$\boldsymbol{r}$	The daily rate each human recovers from infection
$\boldsymbol{v}$	The number of days from infection to infectiousness in the mosquito
$\boldsymbol{g}$	The instantaneous death rate of a mosquito, also expressed as $-\ln(\boldsymbol{p})$
$\boldsymbol{p}$	The probability of an adult mosquito surviving one day

The original Ross model of malaria transmission was simulated using discrete time steps equal to one month. This use of a fixed time step represents one of two broad classes of numerical implementation methods within mathematical models: discrete and continuous time. The latter class was used in formulating the second dynamic model of malaria in 1911 (Ross, 1911), which utilised a pair of differential equations (parameterisation standardised as before (Smith *et al.*, 2012)) to describe how the number of infected humans, X, and the number of infectious mosquitoes, Z, change over time as follows:

$$\frac{dX}{dt} = ma\frac{Z}{M}(H - X) - rX$$

$$\frac{dZ}{dt} = ac\frac{X}{H}(M - Z) - gZ$$

Here H is the total number of humans in the population of focus. Similarly, M is the total number of mosquitoes, and then $m = M/H$. These differential equations do not incorporate the delay from infection to infectiousness in the mosquito.

The Ross model of malaria transmission was developed further by George MacDonald during the Global Malaria Eradication Programme (GMEP) between 1955 and 1969 (MacDonald, 1957). The resulting Ross-MacDonald model was used to provide insight into the efficacy of using the insecticide DDT as a malaria control strategy, with the explicit aim of eliminating the parasite. This model illustrated the impact that could be achieved by reducing mosquito longevity and the subsequent demonstration of the non-linear relationship between increasing mosquito death rates and decreasing sporozoite positivity rates (MacDonald, 1956). MacDonald's work also led to the development of a quantitative theory for malaria control that was explained by the impact of interventions on the entomological transmission of malaria. This resulted in the first formulation of the basic reproductive number, R_0, for malaria.

In the 1970s, the World Health Organisation sponsored an investigation within the Garki district of Nigeria to assess if malaria could be eliminated using a combination of treating cases effectively with chloroquine and IRS (Molineaux and Gramiccia, 1980). Although the project in Garki ultimately failed to eliminate malaria within the study region, the parameter fitting of the model enabled a quantitative relationship between both entomological and parasitological prevalence and incidence to be formulated. This enabled estimation of key malaria indices concerning the vectorial capacity below which malaria could not be maintained at an endemic level. Lastly, the model moved towards a more nuanced understanding of the range of possible endemic levels, and arguably was the first attempt to use mathematical models to predict how control interventions could lead to reductions in vectorial capacity and subsequent changes in the severity of endemicity.

1.5.4 Transmission Metrics: R_0 and Vectorial Capacity

The basic reproductive number, R_0, or the number of cases expected to arise from a single infected individual in a well-mixed, fully susceptible population, has become a fundamental concept in the study of infectious disease dynamics. Borrowed from demography (Dublin and Lotka, 1925) to quantify population growth, R_0 was first used in an infectious disease context by MacDonald to quantify malaria transmission (Macdonald, 1952). The classical equation for R$_0$ for malaria is illustrated visually in Figure 1.2 and written as follows:

$$R_0 = \frac{ma^2bc}{gr}e^{-gv} = \frac{ma^2bc}{(-\ln\ p)r}p^v$$

R$_0$ laid the foundation for a quantitative approach to designing and evaluating malaria control and eradication schemes, especially in estimating the impact of targeting adult mosquito vectors. R$_0$ allowed epidemiologists to quantify two key concepts: 1) the effect size of an intervention and 2) an epidemic or endemic threshold. According to theory, an R$_0$ of one is the threshold below which an epidemic disease will not invade a susceptible population and an endemic disease, such as malaria will eventually die out. Therefore, establishing a measure of R$_0$ and aiming to reduce it to below one provides a simple framework for malaria elimination. In recent years there has been some debate about whether reducing R$_0$ below one is sufficient to eliminate malaria (Breban, Vardavas and Blower, 2007; Li, Blakeley and Smith, 2011). Simple models suggest a bi-stable equilibrium – suggesting that even with R$_0$ below one, malaria can persist indefinitely in a population (Smith *et al.*, 2013). However when Griffin (2016) explored the relationship between R$_0$ and Entomological Inoculation Rate (EIR) for biologically realistic models which incorporated the effects of immunity and were calibrated to a wide range of datasets, the bi-stable equilibrium did not appear to exist for *P. falciparum* malaria.

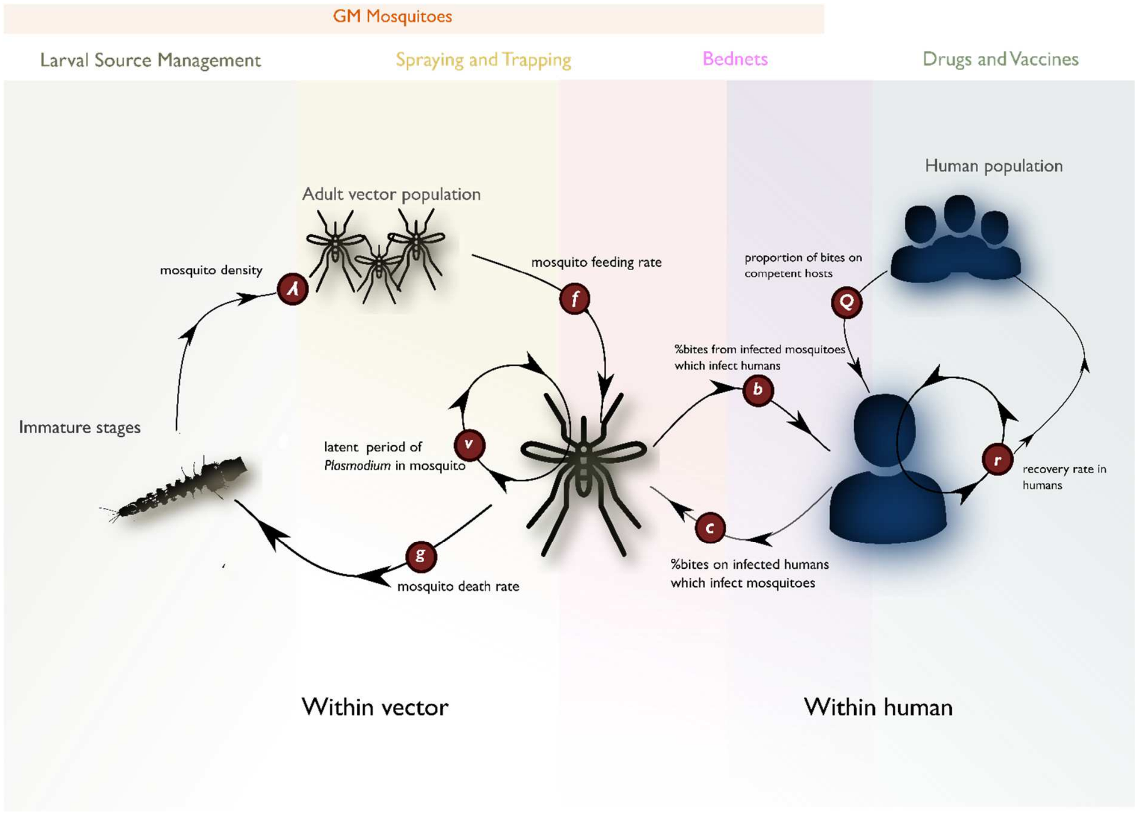


***Figure 1.2 : Visualisation of parameters in R0 equation**. The sections of the diagram shaded in different colours correspond to the different interventions which effect each parameter/part of life cycle*

The ratio of R_0:R or the *effect size* is a measure of what has already achieved with existing interventions. When R remains above one, further interventions will be required to achieve eventual elimination of malaria. Quantifying the effect size was key in supporting decisions to target the mosquito vector in malaria control strategies. This is because interventions which affect adult mosquitoes are likely to have a larger effect size than interventions which reduce parasite density, thereby illustrating the importance of interventions which target the vector. Returning to the equation for R_0, we can see that R_0 has a linear relationship with mosquito density, *m*; infectivity of mosquitoes to humans, *b*; the infectivity of humans to mosquitoes, *c*; and the infectious period, *1/r*. However it increases quadratically with increases to human feeding rates *a*. Depending on *v*, the extrinsic incubation period of the mosquito, R_0 increases more or less cubically with increases in the mosquito death rate *g* (Smith *et al.*, 2012). Therefore, transmission intensity is highly sensitive to mosquito survival – meaning if adult mosquito survival was reduced through interventions, a great impact on malaria transmission could be achieved. In addition, many interventions which reduce mosquito survival will also reduce mosquito density and, if also a mosquito repellent, possibly affect human feeding rates. Interventions targeting the parasite however only impact linear parameters within the equation: *b*, *c* and *r*.

Mathematical models also lead to the derivation of other important metrics of malaria transmission. The vectoral capacity, defined as the total number of potentially infectious bites to mosquitoes arising per fully infectious human per day (Garrett-Jones, 1964), or measure of mosquito exposure to infectious humans, has been a key measure of transmission potential. It is described by a four-parameter equation as follows, using the same parameter symbols in previous equations:

$$\mathrm{V} = \frac{\mathrm{ma}^2 \mathrm{e}^{-\mathrm{gv}}}{\mathrm{g}}$$

The entomological inoculation rate long used as a key measure of malaria transmission is the converse of the vectoral capacity : a measure of exposure of humans to infectious mosquitoes. It is defined as the number of infectious bites a human receives over a given time period, and is the product of the human biting rate, *ma* and the sporozoite prevalence, Z/M. Both measures have been widely used in quantifying transmission potential, designing effective transmission strategies and measuring reductions in malaria burden through vector control.

1.5.5 Looking forward: Modelling to support malaria elimination

Burden reduction is a key aim in malaria control, especially in higher transmission settings. However, the eventual goal for most malaria endemic countries is to achieve and sustain elimination, defined as the absence of locally acquired cases. Country level elimination is a key stepping stone towards global eradication. However, several key changes in malaria epidemiology occur as cases approach zero which must be overcome in order to accelerate, achieve and maintain elimination. Mathematical models and theory can provide useful insights in elimination planning and implementation to understand and overcome these challenges. Here I focus on six key features of malaria epidemiology and monitoring in elimination and near elimination settings: focality of cases; asymptomatic reservoirs; importation and relapse; "residual" transmission, resurgence and changes in the demography of those infected; and developing metrics to quantify progress towards, achievement and maintenance of elimination.

1.5.5.1 Focality of cases

Firstly, malaria prevalence and incidence is heterogeneous at multiple spatial scales (Bousema *et al.*, 2012; Clements *et al.*, 2013; Bejon *et al.*, 2014), with the disease becoming more focal as cases approach zero (Carter, Mendis and Roberts, 2000; Bousema *et al.*, 2012; Sturrock *et al.*, 2016). Because of the tendency for cases to cluster geographically approaching elimination, it has been suggested that targeting interventions to areas of higher transmission may provide a more efficient allocation of resources than implementing control measures homogenously (Carter, Mendis and Roberts, 2000; Bousema *et al.*, 2012). Spatial targeting of resources also may accelerate elimination efforts by allowing a more effective and rapid response to outbreaks. However, there is uncertainty as to how stable hotspots of transmission are, what their contribution to overall malaria dynamics is, whether their stability varies at different spatial scales, and whether this stability can be predicted. There is evidence that at some spatial scales hotspots of febrile malaria may be highly temporally variable, whereas asymptomatic parasitaemia seemed to be much more stable (Bejon *et al.*, 2010, 2014), however other studies have found also found stability in hotspots of febrile malaria (Ernst *et al.*, 2006). These issues inform how and at what spatial scale to target elimination efforts. In order to address this, spatially explicit models of malaria transmission are required.

1.5.5.2 Asymptomatic reservoirs

Measuring progress towards elimination is further complicated by asymptomatic reservoirs. Reviews (Okell *et al.*, 2012; Bousema *et al.*, 2014) of prevalence surveys find a non-linear relationship between prevalence by PCR and by RDT or microscopy, suggesting that even in low transmission settings the prevalence detected by PCR is often higher than that detected by other methods. However the amount of sub-patent and/or asymptomatic malaria in low transmission contexts is highly heterogeneous (Okell *et al.*, 2012). It is also unclear what the contribution of asymptomatic reservoirs are on ongoing transmission. Quantifying the asymptomatic reservoir and modelling its impact on malaria transmission remains a major challenge in modelling elimination scenarios.

1.5.5.3 Importation and relapse of cases

As countries reach low numbers of locally acquired cases, the role of imported cases in sustaining transmission or reintroducing malaria to a country is a concern (Cotter *et al.*, 2013; Churcher *et al.*, 2014). Increases in international travel, migration and connectivity all can lead to more importation. In addition, movement between neighbouring malaria endemic countries, documented or undocumented, as the result of socio-political, environment or economic changes can be important drivers of local transmission via importation or internal movement (Chuquiyauri *et al.*, 2012). Examples of such changes in transmission, often related to land-use change such as mining and logging, have been observed in a variety of countries in SE Asia and the Americas including Cambodia (Sluydts *et al.*, 2014; Guyant *et al.*, 2015; Siv *et al.*, 2016), Thailand (Dondorp *et al.*, 2009), Indonesia (Surjadjaja, Surya and Baird, 2016) and Peru (Rosas-Aguirre *et al.*, 2016). As cases get closer to zero, transmission becomes more evidently focal, with small areas of more intense transmission intensity (Oesterholt *et al.*, 2006; Bejon *et al.*, 2010, 2014; Sturrock *et al.*, 2016).

1.5.5.4 "Residual" transmission, resurgence and changing demographics

It has been noted that in some areas, even with high coverage of interventions, transmission of malaria is sustained. Understanding the causes of residual transmission and targeting resources to these foci of residual transmission is required. There is a high economic cost of surveillance and response, and therefore targeting resources in the most efficient and effective way is important.

Characterising communities at risk, or driving transmission is also key due to specific needs of the communities or strategies which may be required, for example different strategies may be required to control malaria in adult migrant workers than in young children (Cotter *et al.*, 2013). Outside of sub-Saharan Africa, the burden of malaria often shifts from *P. falciparum* to *P. vivax* in lower transmission settings (Gething *et al.*, 2011; Cotter *et al.*, 2013). Because *P. vivax* can remain dormant and lead to relapses in illness, understanding *P. vivax* epidemiology as cases approach zero becomes increasingly important. *P. vivax* also proves more resistant to traditional control measures. The ability to spend long periods of time in dormancy in the liver allows the parasite to survive in contexts where *P. falciparum* would be unable to do so (Battle *et al.*, 2014) .

Most countries which successfully eliminated malaria have managed to maintain elimination, despite importation (Smith *et al.*, 2013). However great resurgences have been seen in countries which reduced levels of malaria but did not achieve elimination. This apparent stability was explored in a series of reviews and meta-analyses (Cohen *et al.*, 2010; Chiyaka *et al.*, 2013; Smith *et al.*, 2013) which concluded that stability in elimination was likely to be due to a combination of existing contextual factors such as economic development and vectorial capacity and self-reinforcing benefits of elimination efforts such as improved surveillance and health systems. However, there is still a great deal of uncertainty in the mechanism behind elimination stability (Smith *et al.*, 2013).

1.5.5.5 Challenges in defining and measuring elimination

Finally, quantifying the impact of interventions on malaria transmission is important to inform the design of optimal intervention strategies, and to evaluate the success of elimination programmes. This has been approached using a wide variety of methods, including mapping (Gething *et al.*, 2014; Bhatt *et al.*, 2015) and mechanistic modelling (Griffin *et al.*, 2010; Walker *et al.*, 2016). However traditional metrics for malaria burden and transmission are not appropriate for near elimination settings. Great strides have been made in mapping many aspects of malaria epidemiology, however these techniques require large numbers of cases to estimate values of interest such as prevalence with a reasonable amount of uncertainty or at an appropriately fine spatial scale. As a result, they cannot be easily applied to elimination settings where case counts are low. The majority of current metrics of transmission for malaria also are difficult to apply to very low-transmission settings (Sturrock *et al.*, 2016). Whilst the EIR is a mainstay in measuring transmission intensity, it is not suited to elimination contexts. The EIR is generally measured through human landing catches and captures a single point in time. Small numbers of infective mosquitoes and focality of transmission in low-transmission settings make accurate EIR values very difficult to obtain.

Increasingly serosurveys have been used to estimate malaria transmission and exposure to the parasite (Corran *et al.*, 2007; Pothin *et al.*, 2016; Biggs *et al.*, 2017; Greenhouse *et al.*, 2018). Although established methods have difficulty in identifying between very recent as opposed to less recent exposure in low transmission settings (Sturrock *et al.*, 2016,) or less abrupt changes in transmission with smaller sample sizes (Sepúlveda, Paulino and Drakeley, 2015), there are is increasing promise for multi-antibody assays which could provide increasingly detailed pictures of malaria exposure (Helb *et al.*, 2015; Greenhouse *et al.*, 2018). Nonetheless, selecting the most informative antibody responses to measure is dependent on context and in many contexts requires further research and development to identify informative antibody responses. Furthermore, these and current serological approaches are limited to locations where cross-sectional surveys have been carried out. Clinical incidence data, when of high quality offers potential for use in such settings, however fewer established methods exist within malaria research to make the most of routine incidence data and account for its potential biases, or for how best to combine with additional information, such as genetic, spatial or serological data. There is a need to develop modelling methods which make use of surveillance data, which is improving in quality in many low transmission contexts.

Challenges in defining elimination operationally also exist. Elimination defined is the interruption of local transmission of malaria within a geographic location, such as a country. The WHO certifies a country as eliminated when there are zero locally acquired cases for three years or more (Cohen *et al.*, 2010; Alonso, 2016). However, this criteria is very difficult for countries to fulfil when bordered by higher burden countries with importation occurring, as importation can lead to some locally acquired cases, even with effective control and surveillance measures (Churcher *et al.*, 2014). In addition, assessing the effectiveness of intervention in low transmission settings can be complicated by the effects of imported cases and *P. vivax* relapses. Relapses and importation can lead to outbreaks of local transmission although the initial source of infection may originate from a much earlier time point or distant point in space. This can reduce the apparent effectiveness of interventions (Churcher *et al.*, 2014).

One way models have addressed this is to develop methods (Churcher *et al.*, 2014; Reiner *et al.*, 2015) which quantify R_c values and how they vary over space and time, as well as modelling human movement (Ruktanonchai *et al.*, 2016). These approaches respectively model underlying malaria transmission potential ("receptivity") and importation risk ("vulnerability"), which together create an indication of overall malaria transmission risk ("malariogenic potential"). However, these methods are still in their infancy in application to malaria elimination.

Over the past 15 years, a wide range of methods have been developed (Ypma *et al.*; Ferguson, Donnelly and Anderson, 2001; Wallinga and Teunis, 2004; Cottam *et al.*, 2008; Chis Ster, Singh and Ferguson, 2009; Morelli *et al.*, 2012) in the context of epidemic disease to measure transmission by estimating individual case and time varying reproduction numbers, using surveillance data which may contain epidemiological, demographic, spatial and genetic distance. They are informed by mechanistic models and empirical data describing key aspects of the transmission cycle of the disease in question, for example the distribution of serial intervals, which is the time between a case showing symptoms and the case they infect showing symptoms (Fine, 2003) . Such methods making use of routine surveillance data have been rarely applied to vector borne diseases such as malaria, with notable exceptions (Reiner *et al.*, 2015; Salje *et al.*, 2016) but will become increasingly relevant as countries reduce malaria burden and increase the strength of their surveillance systems. These approaches can inform elimination policy and inform appropriate elimination strategies.

1.6 Key measures of transmission for outbreaks: applicable to malaria elimination contexts?

In outbreak situations and studies of diseases which take on epidemic dynamics, a wide range of techniques have been developed to determine key epidemiological parameters from surveillance data. There are interesting parallels between malaria in elimination settings and outbreak scenarios which could mean similar approaches will be useful and applicable.

Malaria often takes on epidemic dynamics nearing elimination. Furthermore, the individual line list data produced by control programmes and ministries of health is often similar in structure to outbreak contexts. As in emerging outbreaks, there are often smaller numbers of cases but more detailed information available about each individual case (often in the form of a line list). Furthermore, in both contexts importation can have important effects on disease dynamics. In addition, in both contexts the epidemic is partially but not fully observed and there can be changes in immunity over time and wave-like incidence patterns are seen.

Two key measures of transmission will be explored here: the generation time distribution and reproduction numbers. One key parameter is the generation time distribution of a disease. The generation time of an infection is defined as the average time between an individual becoming infected and passing the infection on to a new individual (Fine, 2003). The distribution of generation times for an infection in a population can provide useful information about its spread and have a wide range of applications to epidemiology, control and elimination strategies. Generation time distributions have been used to infer likely chains of transmission (Wallinga and Teunis, 2004; Cauchemez *et al.*, 2016), explore changes in transmissibility over time (Fraser, 2007; Cori *et al.*, 2013), explore the impact of interventions (Ster, Singh and Ferguson, 2009; Walker *et* *al.*, 2012) or social/environmental factors (Salje *et al.*, 2016), and understand drivers of remaining locally-acquired transmission (Perkins *et al.*, 2015; Reiner *et al.*, 2015).

Often the time of infection is not known, but rather the date symptoms begin, or date care was sought. As a result, the serial interval (SI), or the time between a primary and secondary case presenting with clinical symptoms, is often used in epidemiological analysis in place of generation time distributions. For directly transmitted diseases, SI distributions can be estimated through contact tracing or household studies (Cauchemez and Donnelly, 2009). However, the generation time and serial interval distribution for indirectly transmitted diseases such as malaria can be difficult to estimate because they involve several events which are poorly observed or characterized, such as the time between becoming infectious and being bitten by a mosquito. The first attempts to characterize the generation time of malaria were by Macdonald in 1956, who suggested that a SI of 36 days was a minimum (MacDonald, 1956). There have been several studies since which characterize the serial interval and/or generation time of malaria. Churcher and colleagues (2014) suggest that the expected serial interval for malaria is has a mean of 33 days with treatment and 102 days without. Huber and colleagues (2016) used a combination of empirical data and mathematical models to estimate distributions of all the key processes contributing to the serial interval of malaria the liver emergence period (LEP), the human-to-mosquito transmission period (HMTP), the extrinsic incubation period (EIP), the mosquito-to-human transmission period (MHTP), and the infection-to-detection period (IDP). This work found that there was a great deal of variability in untreated or asymptomatic malaria, due to a long tail in distribution of human to mosquito transmission period. Their work estimated a mean of 48 days with treatment and 102 without, with the discordance between their estimates and previous estimates being the result of different assumptions surrounding the delay between symptom onset

and seeking treatment.

Reproductive numbers and how they vary through time and space are also key measures which epidemiologists attempt to establish. In a malaria elimination context, the reproductive number is a key measure for understanding whether malaria is likely to persist and for how long (Churcher *et al.*, 2014). As discussed in Section 1.5.1, the assumptions inherent in the basic reproductive number, R_0 are rarely present in real disease transmission, and as a result the effective reproductive number, R can be used to measure transmission when the population is not fully susceptible, either as a result of control and/or immunity from prior exposure.

1.7 Estimating key transmission characteristics from epidemiological surveillance data

1.7.1 Why estimate transmission routes and reproductive numbers?

Infectious diseases can be described statistically as point processes where events (such as the onset of disease symptoms) are fully or partially observed but the processes generating them are not. In recent years several statistically rigorous methods have been developed to infer chains of transmission from epidemiological surveillance data. Understanding the transmission network of an outbreak as well as key parameters such as R is highly relevant to the dynamics of malaria in low transmission settings. In these contexts, disease dynamics resemble epidemics or outbreaks, due to low immunity within the population and importation events where parasites are introduced into areas where suitable vectors are present. Inferring the most likely routes of transmission between individuals or groups can provide a range of useful insights, such as covariates associated with infectors and/or infectees, transmission kernels and modes of transmission (Morelli *et al.*, 2012). Transmission chain reconstruction has proved valuable in informing control and intervention policy, with the first application of this approach following the 2001 Foot and Mouth epidemic in the UK (Ferguson, Donnelly and Anderson, 2001; Keeling *et al.*, 2003) .

Since this period, transmission chain reconstruction and estimation of reproduction numbers have been found to be useful in real-time and retrospective studies of outbreaks and epidemics, such as the 2003 SARs outbreak (Wallinga and Teunis, 2004) and global Influenza pandemics (Ghani *et* *al.*, 2009). A key example of the utility of such approaches is during the response to the 2014 Ebola epidemic. Responding in real-time to line-list data, analyses using approaches to quantify reproduction numbers using a mixture of contact tracing and inference methods (Faye *et al.*, 2015) revealed key epidemiological information to inform targeted containment and control strategies.

1.7.2 Use of networks in epidemiological modelling

There has been a large body of work within epidemiology and infectious disease modelling exploring the structure of populations and modelling this structure through networks of social contacts or interactions (Welch, Bansal and Hunter, 2011). Individuals in the population are represented as nodes and their potential contacts for disease transmission are represented as edges. The focus of this work is generally on directly transmitted disease, especially sexually transmitted infections (Keeling and Eames, 2005), however a wide variety of diseases have been explored, including indirectly transmitted diseases (Reiner *et al.*, 2015; Salje *et al.*, 2016). Whilst much of this work has been in exploring the effect of network structure on disease dynamics and the impact of control measures (Cauchemez *et al.*, 2006; Walker *et al.*, 2012), there has also been a great deal of work carried out in developing rigorous statistical methods for inferring contact structure or transmission trees. Many methods used in epidemiology to infer transmission chains build upon an approach popularized by Wallinga and Teunis in 2004 (Wallinga and Teunis, 2004), which allows the inference of most likely transmission routes using incidence time series data and a serial interval/generation time distribution. Consider an individual i, infected (or shows symptoms) at time t_i. The probability of infection from an individual/member of cohort j which was infected at time t_j is determined by a function w, which can be the generation time or serial interval distribution, normalised by the likelihood of any other candidates infecting i .

$$P_{ij} \; = \; w(t_i t_j)/\sum_{ik} w(t_i t_k)$$

The case reproductive number is the sum of all likely transmissions resulting from a case or cohort of interest infected/showing symptoms at time j. In other words, it describes how many individuals on average an individual infected at time j will go on to infect.

$$R_c = \sum_i P_{ij}$$

1.7.3 Extensions and developments

The Wallinga and Teunis method provided a useful tool to quickly derive important measures from epidemiological surveillance data and, with suitable prior information, estimate most likely transmission routes. Developed to assist in the analysis of the 2003 SARs outbreak in Singapore, this and similar approaches allowed for rapid real time quantification of key epidemiological parameters from limited surveillance data (Wallinga and Teunis, 2004).

However, there are several limitations which have been explored through a variety of approaches and extensions, summarized in Table 1.3. Often there can be uncertainty in both the date of symptom onset or infection and also in the proportion of unobserved cases. Unreported cases may shape inference of transmission by linking cases which occurred further away from each other in time, leading to slower apparent rates of transmission between cohorts or individuals. Previous work (Ferguson, Donnelly and Anderson, 2001; Walker *et al.*, 2010, 2012) has considered uncertainty in dates of symptom onset by treating symptom dates as nuisance parameters within a Bayesian framework. Data augmentation methods have also been used to explore the impact of unreported cases(Ferguson, Donnelly and Anderson, 2001; Salje *et al.*, 2016). In many contexts, timing of infection alone is not suitable to accurately reconstruct chains of transmission. There may be many candidates with similar likelihoods of transmission if many cases occur in a short space of time or if the SI distribution is wide. As a result a wide variety of extensions have been introduced which incorporate genetic data (Ypma *et al.*; Cottam *et al.*, 2008; Morelli *et al.*, 2012), spatial data via estimation of a spatial kernel (Morelli *et al.*, 2012; Walker *et al.*, 2012; Salje *et al.*,

2016) and demographic data such as age and sex (Salje *et al.*, 2016). These extensions have provided key insights into transmission dynamics and the impact of control measures for a wide range of diseases. However, there are often strong assumptions on the generation time interval and little formal inclusion of error or variation generated by the wide variety of factors which may affect likelihood of transmission between transmission pairs. This is particularly true for diseases with long and variable generation times, such as untreated malaria.

1.7.4 Approaches from other fields of study

Within applied statistics and machine learning research there has been a rich body of work using information on timing of node "activation" to reconstruct networks. These networks often represent flows of information between individuals, for example through online social networks. The generic problem of knowing (or being able to infer) times of contagion infecting nodes, but not observing the process of transmission, is highly applicable to infectious disease outbreaks. A class of models known as independent cascade models, introduced by Kempe in 2003 (Kempe, Kleinberg and Tardos, 2003) were first proposed to solve a problem known as influence maximization - to identify the most influential "nodes" in a network through which information is propagated (for example the posting of viral videos by individuals in a social network). The independent cascade model can be thought of as a generalized Susceptible - Infected (SI) model.

The basic assumptions of the independent cascade model (Kempe, Kleinberg and Tardos, 2003) are:

1) Infections are binary; an individual is infected or is not. Intensity of infection is not modelled.
2) Infections along edges occur independently of each other.
3) Infection propagates through network via diffusion. There are no external sources of infection.
4) Cascades of infection propagate independently of each other.
5) A node is infected only by the action of one parent node. Cascades map onto transmission trees

Following these basic assumptions, algorithms to estimate network diffusion rates and structure have been developed, such as *NetInf* and *NetRate* (Rodriguez, Balduzzi and Schölkopf, 2011; Gomez Rodriguez *et al.*, 2014) and feature-enhanced methods (Wang, Ermon and Hopcroft, 2012). *NetInf* infers networks through observations of the timing of "cascades" of infection events. The algorithm assumes equal weights or α values on edges between nodes in an unobserved network. This assumes that all connected nodes in the network infect their neighbours with the same probability. Following from Kempe (Kempe, Kleinberg and Tardos, 2003) the submodularity properties of the independent cascade model were exploited in this algorithm, meaning the likelihood of a given cascade (or transmission network) can be defined as the sum of all the pairwise likelihoods of transmission between each node in that cascade. An extension to this method, *NetRate* (Rodriguez, Balduzzi and Schölkopf, 2011), removes the assumption of constant hazards of infection, allowing estimation of varying relationships between infection hazard and time. This better captures the complex factors beyond time (e.g. age, sex, location, immunity, rainfall) which may affect probability of transmission occurring between nodes. *NetRate* additionally casts the network diffusion as a survival likelihood parameterised by hazard functions. These together form a function describing how the likelihood of transmission varies over time. The parametric form of the hazard, survival and likelihood depends upon hypothesised mechanisms of transmission, and

There have been a variety of real-world applications which diffusion network approaches have been used for, mainly surrounding analysis of the spread of information and influence along online social and media networks. For example, this approach was used to reconstruct the spread of particular memes and hashtags to better understand the way in which information travels between blogging sites and mainstream media outlets, and comparing how this varies for population-wide events such as civil uprise in Syria during the Arab Spring, compared to unexpected news events which may generate large amounts of attention for a shorter period of time, such as the death of singer Amy Winehouse (Gomez Rodriguez *et al.*, 2014).

1.7.5 Comparison of approaches

Whilst there are many commonalities in the approaches introduced here, there are several key differences in approaches and features of each algorithm, summarised Table 1.3. A key feature of the information diffusion/independent cascade approaches is that they were designed to be generic and flexible to different problems and contexts, whereas within infectious disease epidemiology generally the approaches were designed to be specific to a particular disease and dataset. The advantage of a specific approach is that it is easier to tailor to the biology and particular features of interest for a particular disease or dataset, however the broader applicability and accuracy of approach in different contexts is then harder to determine. It also may be not obvious how to include new, additional sources of information.

The way that the likelihood is constructed, and therefore how inference is performed also varies between approaches. Some inference frameworks jointly infer multiple parameters within single inference framework and likelihood, whereas others have been multi-staged and more heuristic (Table 1.3). Previous approaches have either allowed all possible connections in a particular network structure (Wallinga and Teunis, 2004), sampled from the likelihood (Ferguson, Donnelly and Anderson, 2001) or explored a limited number of pathways (Salje, Cummings and Lessler, 2016). Instead, the information diffusion approaches introduced in this chapter either find the most likely underlying transmission network given the timing of symptom onset for a set of k transmission events linking cases using a greedy algorithm, (Rodriguez and Schölkopf, 2012), or in the case of *NetRate* the transmission tree and all possible linkages between cases are considered, but, as will be described further in Chapter 2, the introduction of a survival term penalises unlikely connections, meaning sparsity is encouraged and the transmission tree log likelihood can be estimated as the sum of all hazards and survivals for each case, meaning that even for large numbers of cases the likely connectivity between cases can be feasibly estimated.

Some approaches were developed within a Bayesian framework whilst others were implemented within a frequentist framework (Table 1.3). Working within a Bayesian framework is helpful when there is prior information or a range of possible parameter values, to incorporate both prior knowledge and uncertainty. However, frequentist approaches are sometimes simpler and easier to implement quickly in outbreak scenarios.

One major difference in structure is that independent cascade models were designed for observations of "multiple cascades" of transmission, where the same node or individual in the network potentially being observed multiple times, e.g. spread of two different hashtags, two periods of time. This is generally not the case for infectious disease. Although multiple and repeated malaria infection is common in high transmission settings, in settings where this type of approach is useful and appropriate it is very unlikely, we will see repeated infections in individual level datasets over reasonable observation windows. Although *NetInf* in particular, this could reduce the ability to accurately reconstruct disease transmission networks, compared to previous applications to problems with multiple cascades available. Therefore, testing on simulated data is recommended to determine the impact of this.

The methods used here were chosen because of their flexibility, potential for incorporating multiple data types within a single inference framework, convex likelihoods, encouragement of sparsity, estimation of full transmission tree, scalability and how well cited and applied such approaches have been within machine learning communit

Table 1.3 Comparison of transmission network approaches.

Reference	Temporal	Spatial	Genetic	Demographic	Uncertainty in infection dates	Unobserved infection	Imported	Variable transmission rates	Single estimation framework
(Ferguson, Donnelly and Anderson, 2001)	Y	Y	N	N	Y	Y	N	N	Y
(Britton and O ’ Neill, 2002)	Y	N	N	N	N	N	N	N	Y
(Wallinga and Teunis, 2004)	Y	N	N	N	N	N	N	N	N
(Ghani *et al.*, 2009)	Y	N	N	Y	Y	Y	Y	Y	Y
(Chis Ster, Singh and Ferguson, 2009)	Y	Y	N	Y	Y	Y	N	N	Y
(Cauchemez *et al.*, 2011)	Y	N	N	Y	Y	N	N	N	N
(Walker *et al.* 2012)	Y	Y	N	N	Y	N	N	Y	N
(Ypma *et al 2010.*)	Y	N	N	N	Y	N	N	N	Y
(Morelli *et al.*, 2012)	Y	N	Y	N	Y	Y	N	N	Y
(Jombart *et al.*, 2014)	Y	N	N	N	Y	N	Y	N	Y
(Reiner *et al.*, 2015)	Y	Y	N	N	Y	N	Y	N	N
(Salje *et al.*, 2016)	Y	Y	N	Y	Y	Y	N	N	N
(Kempe, Kleinberg and Tardos, 2003)	Y	N	N	N	N	N	N	N	N
(Rodriguez, Balduzzi and Schölkopf, 2011)	Y	N	N	N	N	N	N	N	Y
(Rodriguez and Schölkopf, 2012)	Y	N	N	N	N	N	N	Y	Y
(Wang, Ermon and Hopcroft, 2012)	Y	N	N	N	N	N	N	Y	Y

1.8 Problem Statement

Malaria elimination at the national level, where local transmission of malaria is no longer sustained (Cohen *et al.*, 2010), is increasingly a goal in global malaria policy. However, as countries approach elimination, changes in malaria epidemiology can pose challenges to reaching zero cases (Cotter *et al.*, 2013). Understanding these changes is important in designing optimal elimination strategies. Challenges also arise in measuring the success of elimination (Cotter *et al.*, 2013; Churcher *et al.*, 2014), both in understanding the stability of elimination (Chiyaka *et al.*, 2013; Smith *et al.*, 2013) and assessing the impact of control measures in low transmission settings, especially in the presence of importation. This information is important when deciding if, when and how to scale back interventions and change surveillance methods (Chiyaka, *et al.*, 2013). This can also inform policy surrounding certification of elimination, which can have significant impacts on countries. For regions which have set ambitious targets for elimination, understanding changes in epidemiology over space and time approaching elimination are highly pressing for designing effective strategies to reach and maintain zero cases. In hyper and meso-endemic settings current methods developed to measure changes in transmission have been effective. However, in low-transmission settings new tools are required. Methods traditionally applied to outbreak data are one such promising tool. In low transmission and elimination settings, malaria cases are infrequent, immunity is lower, known cases tend to be treated and surveillance is often stronger. When linked with covariates of interest and spatial information, reconstructed transmission chains and R_c values can be mapped (Reiner *et al.*, 2015). They provide information about changes in transmissibility over time (Cori *et al.*, 2013), reveal heterogeneities in transmission between individuals and/or cohorts (Cauchemez *et al.*, 2011), and can be linked to both environmental/demographic factors and interventions to explore their role (Salje *et al.*, 2016).

1.9 Aims and approach

My thesis aims to introduce a new approach to quantifying malaria transmission in near elimination settings by extending, adapting and applying statistically rigorous methods from independent cascade family of models to estimate individual level reproduction numbers. I then analyse these estimates, time series analysis and geostatistical approaches to quantify how they vary over space and time and uncertainty in these estimates. I aim to use these methods to retrospectively explore the dynamics of malaria transmission in several elimination settings, and in doing so provide useful evidence to support decision making around elimination certification and policy.

Chapter 2 describes the methodological approach used in chapters 3-5 and rationale for its development and use**.** describing the novel extensions and applications made to independent cascade models to apply them in the context of this thesis, namely carrying out work within a Bayesian framework, considering missing cases, and incorporating spatial information. This chapter also includes the results of testing methods on simulated data. **Chapter 3** illustrates an application of one such extension to a previously unanalysed dataset from El Salvador, and further timeseries analysis and geospatial analysis is used to explore how malaria transmission has varied over space and time as the country approaches elimination and explore the impact of imported cases on malaria transmission. **Chapter 4** illustrates a further application of a different extension, allowing joint Bayesian inference of the connectivity between all cases, scalable to large datasets and illustrates the application of this method to an individual-level dataset from Yunnan province China which has previously only been analysed descriptively. As in Chapter 3, geostatistical and additive regression models are used to further analyse the estimated spatiotemporal changes in transmission. **Chapter 5** illustrates the version of the model adapted to include spatial information, applied to four line-list datasets from diverse economical, demographical and ecological contexts in elimination settings. In **Chapter 6** I summarise and discuss the results and approach as a whole, considering key findings, limitations and future directions.

2
Methodology

2.1 Introduction

As introduced in Chapter 1, this thesis aims to develop and apply methods to measure malaria transmission and its spatiotemporal variation in very low transmissionand elimination settings, where high quality individual-level surveillance data are available but case numbers are relatively low and metrics such as parasite prevalence are no longer informative. Here I introduce and derive the core algorithms and approaches used in Chapters 3-5 to estimate reproduction numbers. I then present the results of testing the methods used in Chapter 4 on simulated data to explore its ability to accurately estimate reproduction number distributions in different contexts with different amounts of missing data.

All methods used were adapted from a family of algorithms, introduced in Chapter 1, which model the diffusion of a contagion along latent networks, where the time and or location of some signal (such as symptom onset, or posting a tweet) are known, but the transmission process itself and the underlying network is unobserved. I chose to adapt, extend and apply these algorithms to malaria case data for several reasons. Firstly, a network diffusion approach addresses the generic problem of observing timings of transmission or diffusion events across networks, where the transmission process itself is unobserved, and has been extensively tested on both simulated and real datasets (Rodriguez, Balduzzi and Schölkopf, 2011; Rodriguez and Schölkopf, 2012; Wang, Ermon and Hopcroft, 2012; Gomez Rodriguez *et al.*, 2014). Furthermore, this approach shares similarities with other network-based approaches which are increasingly used to understand infectious disease dynamics (Wallinga and Teunis, 2004; Cori *et al.*, 2013), but rarely applied to malaria and other vector borne diseases (with the exception of Reiner *et al.*, 2015). In addition, due to information diffusion algorithms being designed with a general focus, they are more flexible and adaptable than many approaches, allowing the incorporation of further data sources and functional forms within a single estimation framework, as will be further explored and discussed in Chapter 5 and the Discussion (Chapter 6). They also have provably convex solutions, meaning there is a single global optimal solution which can be estimated from gradient descent optimisation of the log-likelihood. They also encourage sparsity, meaning many parameters shrink to zero and overfitting is penalised, which is advantageous for this type of inference when multiple parameters are estimated from small to moderate numbers of cases. These algorithms have been widely cited, used and tested on a variety of real and simulated datasets (Rodriguez, Balduzzi and Schölkopf, 2011; Rodriguez and Schölkopf, 2012; Wang, Ermon and Hopcroft, 2012; Gomez Rodriguez *et al.*, 2014) and provide a flexible approach to leverage diverse datatypes within a single inferential framework.

Due to the aims of this thesis, namely, to quantify spatiotemporal variation in malaria transmission in near elimination and elimination settings, I do no aim to specifically infer who infected whom. Instead, this thesis aims to produce temporally and spatially sensitive estimates of transmission as measured by reproductive numbers, as well as quantify uncertainty in these estimates. However, these frameworks do estimate transmission likelihoods and therefore transmission pathways can be constructed. Therefore, there is potential to utilise these methods, especially if supplemented with contact tracing data and/or genetic data to explore reconstructed networks and their properties. In order to facilitate understanding of the approach and the process, the technical derivation of the core algorithms as well as their extensions and the rationale behind their choice are included here. For ease of reading, a simplified description of the relevant methods are also included in the methods sections of Chapters 3-5.

Before deriving and describing each approach separately, it is useful to consider what all methods share. All methods require a line list of individual cases and symptom onset times as a minimum, however can incorporate additional information, such as imported/local classification based on epidemiological investigation or travel history and location of residence or health facility. In this thesis I do not incorporate information such as genetic distance due to a lack of data available in the contexts of focus, however in theory, any appropriate distance matrix could be incorporated within the framework presented here.

All approaches take a prior on hyperparameters defining a serial interval distribution and then estimate the connectivity between cases, or the likelihood that each case infected the others. In all methods this inference is on the whole transmission tree, rather than solely considering pairwise transmission. All methods have been shown to provide convex and sparse solutions, allow for missing infectors, and use estimates of connectivity to estimate individual reproduction numbers for each case.

There are several assumptions inherent to all approaches in this thesis. The implications of these assumptions are discussed in detail in relation to each dataset and context in Chapters 3, 4 and 5. Briefly, all approaches do not explicitly model reinfection/relapse, although do allow for unobserved sources of infection, which could be due to either of these processes. They assume that classification of cases as imported or locally acquired by elimination programmes is correct, and that therefore cases classified as imported can infect other cases but not be infected by other cases themselves. All assume that infection and symptom onset is in chronological order – i.e. cases will always show symptoms after their infectors, and therefore cases can only be infected by those which showed symptoms earlier than they did.

2.2 Algorithm 1: Submodular inference from multiple trees (Implemented in Chapter 3)

To infer the most likely pathways of transmission linking observed cases, I extended and adapted *Multitree* (Rodriguez and Schölkopf, 2012), a method based upon the independent cascade model introduced in Chapter 1 (Kempe, Kleinberg and Tardos, 2003). This algorithm exploits the submodular properties of the objective function, which in this case is the negative log-likelihood function we aim to optimise. This submodularity, or the ability to calculate the negative log likelihood by calculating the pairwise likelihoods, means one can use pairwise likelihoods of transmission events occurring based on time of symptom onset and serial interval distributions and a greedy algorithm to iteratively build the most likely tree connecting observed cases for a given serial interval distribution and assumption about likelihood of infection by external source. This approach was specifically designed for small numbers of data points (Rodriguez and Schölkopf, 2012). In order to allow the inference of probabilities of transmission rates and estimate uncertainty in the estimates of the network connectivity, priors for the hyperparameters shaping the serial interval distribution were defined. By drawing many times from a prior distribution of hyperparameters governing a serial interval distribution and varying the value of epsilon, the parameter determining the likelihood of an unobserved source of infection infecting a case, it is possible to generate estimates of uncertainty in estimates transmission links and the corresponding reproduction number estimates calculated from them.

There are some important limitations to this approach. Firstly, the choice of cut-off point in marginal gain in likelihood for invoking additional edges in the network is somewhat heuristic. I address this in Chapter 3 by carrying out a sensitivity analysis and inspecting where the marginal gains in likelihood begin to asymptote. Secondly, for each network, the edges are defined as 1 or 0, there is no measure of the importance or likelihood of each edge for a single network. However, the marginal gain in likelihood that each edge provides can be used as a measure of importance for each edge of the network. Furthermore, by adapting the model to have a pseudo-Bayesian approach with priors which are drawn from many times, we can also average the network and subsequent reproduction number estimates to obtain estimates of uncertainty.

2.2.1 Data and parameter inputs

This method assumes a dataset consisting of a time series of symptom (fever) onset of malaria cases $t \in \{t_1, \dots, t_n\}$, time ordered such that $t_1 < t_2, \dots, < t_n$. While the times of symptom onset are known, what is not known is who infected whom. The goal of the model is to infer the most probable network structure, $\mathcal{G}$, connecting these n cases. We infer $\mathcal{G}$ solely from the symptom onset times t, a serial interval distribution, and hyperprior probability distributions for the serial interval distribution parameters.

2.2.2 Serial interval distribution

The serial interval is the time between a given case, j, showing symptoms and the appearance of symptoms in a case i infected by the earlier case, such that $t_j < t_i$ (Fine, 2003). The serial interval distribution specifies a normalised pairwise transmission likelihood, or the likelihood of case *j* infecting case *i*, given the time between symptom onsets, $t_i - t_j$. The model allows flexibility to define a range of prior distributions for possible serial interval distributions to allow for possible variation in transmission dynamics. For example, even in contexts where malaria transmission is extremely low and surveillance is high quality there remains a possibility of a small number of asymptomatic or undetected and therefore untreated infections contributing to ongoing transmission, which will take on a longer serial interval. Defining a prior for the shape parameter of a serial interval distribution accounts for some of this uncertainty. The specific parameter specifications used for serial intervals in particular contexts are described separately in Chapters 3-5.

In all the applications explored in this thesis, a shifted Rayleigh distribution is used for the serial interval distribution, which is a special case of a Weibull distribution. Used widely in modelling propagation events and the serial intervals of other infectious diseases (Brookmeyer, Gail and Gail, 1994; Virlogeux *et al.*, 2015), when shifted to include an incubation period it is very similar in density to modelled malaria serial intervals (Huber *et al.*, 2016).

2.2.3 Algorithm derivation

Due to no evidence of *P. vivax* relapse in the transmission contexts explored in this thesis, I assume that a case can only be infected once by a case which has shown symptoms earlier in time. For a possible transmission tree $\mathcal{T}$ connecting cases with a set of transmission events or edges linking cases, $\mathcal{E}_{\mathcal{T}}$, the likelihood of observing symptom onset times conditional on a given $\mathcal{T}$ is: $f(t|\mathcal{T}) \propto \prod_{(u,v)\in\mathcal{E}_{\mathcal{T}}} f(t_u|t_v;\alpha,\gamma)$. Given this likelihood on a single transmission pathway $\mathcal{T}$, the underlying graph is found by considering all possible transmission pathways supported by a given network $\mathcal{G}$: $f(t|G) \propto \sum_{\mathcal{T}\in T(\mathcal{G})} f(t|\mathcal{T})\mathbb{P}(\mathcal{T}|\mathcal{G})$ where $T(\mathcal{G})$ is the set of all the possible transmission pathways for $\mathcal{G}$. By imposing a flat prior on $\mathbb{P}(\mathcal{T}|\mathcal{G})$ and as a consequence of the assumptions of a single parent node with an earlier symptom onset date the likelihood simplifies to

$$f(t|\mathcal{G}) \propto \sum_{\mathcal{T}\in T(\mathcal{G})} \prod_{(u,v)\in\mathcal{E}_{\mathcal{T}}} f(t_u|t_v;\alpha,\gamma) \qquad (1)$$

My derivation until this point is the same as that introduced by Wallinga and Teunis (Wallinga and Teunis, 2004) and extended to a wide variety of contexts by others (Morelli *et al.*, 2012). However, methods based on Wallinga and Teunis make the strong simplifying assumption that the likelihoods of all spanning trees on $\mathcal{T}$ and therefore $\mathcal{G}$ are constant. Thus, they fundamentally do not infer the most probable underlying network structure or jointly consider all infection times at once. In contrast, by following the approach introduced by Gomez-Rodriguez and Shölkopf (Rodriguez and Schölkopf, 2012), one can solve the optimisation problem $G = \max_{|G|\leq k} f(t|G)$ for a set of at most k edges, or transmission events linking cases. The two fundamental challenges with solving this optimisation problem are (a) the sum $\sum_{\mathcal{T}\in T(\mathcal{G})}(\cdot)$ is evaluated over all directed spanning trees in $\mathcal{G}$, which can be super-exponential in n, and (b) $\max_{|G|\leq k} f(t|G)$ is a special case of the maximum coverage problem which has been proven to be NP-hard (Khuller, Moss and Naor, 1999) and therefore unsolvable without searching all possible transmission trees with brute force. Following previous approaches (Gomez-Rodriguez, Leskovec and Krause, 2010; Rodriguez and

Schölkopf, 2012), challenge (a) can be solved by observing that the resulting matrix $f(t_u|t_v; \alpha, \gamma)$ for all $(u, v) \in \mathcal{E}_{\mathcal{T}}$ pairs is an upper triangular connectivity matrix. From Tutte (2003) and Gomez and Shölkopf (2012) the connectivity matrix can be expressed as a determinant, which for an upper triangular matrix is the product of the diagonal elements. Therefore, the likelihood in equation (1) becomes tractable and can be evaluated in quadratic time as:

$$f(t|\mathcal{G}) \propto \prod_{t_i \in t} \sum_{t_j \in t, t_j < t_i} f(t_i|t_j; \alpha, \gamma) \text{ (2)}$$

Equation (2) can be evaluated on a log scale

$$F(t|\mathcal{G}) \propto \sum_{t_i \in t} \log\left(\sum_{t_j \in \mathrm{t}^c : \mathrm{t}_j \leq \mathrm{t}} f(t_i|t_j; \alpha, \gamma)\right) \text{ (3)}$$

For challenge (b) it can be proved (Rodriguez and Schölkopf, 2012) that, while finding an optimum to solve $\max_{|G| \leq k} F(t|G)$ is NP-hard, the structure of $F(t|G)$ is submodular. Submodularity in the structure of $F(t|G)$ yields a natural property of diminishing returns. That is, the incremental value that a single edge makes when added to $\mathcal{G}$ decreases as the size of the graph increases. Optimising submodular functions is possible using the greedy algorithm with provable and near-optimal performance guarantees (Nemhauser, Wolsey and Fisher, 1978). To implement the greedy algorithm, we start with an empty graph, $\mathcal{K}$, and then add edges sequentially such that the *marginal gain* from each iteration is maximised. Formally, this means one starts with $\mathcal{G} = \mathcal{K}$ a and then each iteration (m) evaluates the edge $e_m \in \{i, j\}\ \forall j < i$ that yields the best marginal gain, $e_m = \max_{e \in \mathcal{G} \setminus \mathcal{G}_{m-1}} F(\mathcal{G}_{m-1} \cup \{e\}) - F(G_{m-1})$, and add this edge to the graph $\mathcal{G} = \mathcal{G} \cup \{e_m\}$. Edges continue to be added and stop when $\mathcal{G} = \{e_1, \dots, e_k\}$ edges is reached. Due to submodularity, the solution quality on increases with each additional edge, however, the marginal gain quickly asymptotes, thus ensuring sparse solutions. The number of edges, *k*, can be determined by setting a cut-off point for when the marginal gain in likelihood of adding edges falls below a certain value.

Two modifications were made to the above optimisation algorithm. Firstly, to incorporate edges known to be importations, I constrain child/infectee, i, edges in $e \in \{i, j\} \forall j < i$ to be only non-imported infections. This ensures that local infections cannot infect imported infections but imported infections can infect any node. Secondly, to account for variation in the serial interval distribution, I run the above greedy scheme for prior samples of α, γ, as discussed above. This approach naturally lends itself to Bayesian formulations. As it currently is applied in this thesis, this formulation uses a proportional likelihood optimised by exploiting submodularity.

2.2.4 Accounting for missing cases

Assuming all cases reaching community health workers or health facilities are recorded, missing cases may be generated by two processes. Symptomatic cases may be missed by not seeking care or being found through active case detection. On the other hand, cases may be asymptomatic and therefore unlikely to seek care or be detected. They may have densities of parasites in their blood which are too low to be detectable by microscopy if active case detection occurs. These reasons for missed detection apply to both imported cases and locally acquired cases. We assume the pool of asymptomatic cases in the country is low and has a small contribution to ongoing transmission. To explore the amount of cases which may be going undetected within the independent cascade framework, we consider additional edges π, that represent unobserved individuals who can infect any observed individual, i, in a transmission chain. Every observed individual i can get infected by unobserved individuals, π, with an arbitrarily small probability ε. This so called ϵ-edge is connected to every node in the network and do not, by design, participate in the diffusion propagation. The ϵ-edges prevents breaks in the network diffusion cascade where the likelihood of transmission between observed cases is sufficiently low, the case is linked to an external source. Additionally, ϵ-edges ensure the likelihood is monotonic, that is, converting an ϵ-edge to a network edge in $\mathcal{G}$ only increases the likelihood. The addition of ϵ-edges was achieved by augmenting the pairwise transmission likelihood as follows: $f(t_i|t_j; \alpha, \gamma) = \epsilon^{-1}\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}$

The specific value of ε should be set to balance between false positives and false negatives when linking cases by infection events. The higher the value of ε, the larger the number of nodes that are assumed to be infected by an external source.

2.2.5 Estimating R_c

In solving $G = \max_{|G|\leq k} F(t|G)$ via the greedy algorithm we estimate k edges $e_m \in \{i,j\}\forall j < i$ by iteratively maximising the marginal gain in the log transmission likelihood of that edge over all other edges $e_m = \max_{e\in\mathcal{G}\setminus\mathcal{G}_{m-1}} F(\mathcal{G}_{m-1} \cup \{e\}) - F(G_{m-1})$. We therefore can calculate a $(n - q) \times n$ matrix, $\mathcal{M},$ for n total infections and q imported infections of $k \leq (n-q) \times n$ marginal gains edges. The rows of the upper triangular matrix $\mathcal{M}$ are therefore the infectees and the columns the infectors. Because the solution is a positive and monotonically increasing function and $F(t|G)$ is submodular, these marginal edge gains asymptote, thereby creating sparse solutions and diminishing gains for each additional edge.

By normalising the rows/infectees of $\mathcal{M}$ and creating a normalised matrix $\mathcal{R} = \mathcal{M}_{[i,\cdot j]}/\sum_{j=1}^{n} \mathcal{M}_{[i,j]}\ \forall\ \{i = 1,..,(n-q)\}$ we get a matrix that represents both which infector edges are connected to infectees and the normalised marginal gain of that edge. Intuitively then, by taking the row sums of $\mathcal{R}$ we get the (fractional) number of secondary infections and therefore a point estimate of the time varying reproductive number $R_c(t_j) = \sum_{j=1}^{n-q} \mathcal{R}_{[\cdot,j]}$. This reflects for an individual, how many people they are likely to have gone onto infect. When multiple individuals have been infected at a given time and/or place, we can take the mean individual $\mathcal{R}_c$ and uncertainty in this value as an indicator of reproductive numbers for a given time and/or location.

2.3 Algorithm 2: Inference of network transmission rates (implemented in Chapter 4)

The submodular approach (Algorithm 1) was suited to the dataset and context to which it was applied in Chapter 3, where there are very few and sparsely distributed cases. However, in some contexts it is advantageous to utilise a less heuristic approach within a single, fully Bayesian optimisation framework. Furthermore, to widen the utility of the approach, a more flexible framework was required which could be modified to include spatial information or other data sources such as genetic or demographic information. The α_{ij} term can be decomposed into constituent variables or multiplied by an additional function, both allowing incorporation of additional sources of information. There was also a need to devise a method which could be easily run on a larger dataset without a need for many computational resources in order to increase the utility of any approach by malaria elimination programmes. As a result, an approach was adopted which jointly infers separate transmission rates for each edge connecting potential infectors and infectees. This has been widely tested on simulated and real datasets, and is advantageous in both having a convex likelihood, meaning global optimal values can be estimated, and in encouraging sparse solutions by penalising non-zero values of α_{ij} through the survival function.

There are several key extensions and adaptations which I developed, considering applications to malaria surveillance data in elimination settings . Firstly, epsilon edges, ε, were added to allow for unobserved sources of infection, acting as competing hazards with observed cases. Secondly, the algorithm was implemented in a Bayesian framework to incorporate uncertainty and prior knowledge about the serial interval distribution and proportion of unobserved cases.

Additional versions of the algorithm were developed and coded in different coding languages to increase speed of computation and facilitate the analysis of larger datasets. I will first introduce the derivation of the general version and then explore the extensions and variations considered.

2.3.1 Data and parameter inputs

Data consist of a set of n infections/nodes $\boldsymbol{I} \in (I_1, \dots, I_n)$ with associated times $\boldsymbol{t} = \{t_1, \dots t_n\} \in \mathbb{R}^+$ and binary yes/no importation status $\boldsymbol{\pi} = \{\pi_1, .., \pi_n\} \in \{1,0\}$

2.3.2 Serial interval distribution

The serial interval distribution of malaria, defining the probability individual I_j infected individual I_i at times $t_i > t_j$ is defined through a shifted Rayleigh distribution $f\left(t_i \middle| t_j; \alpha, \gamma\right) = \alpha\left(t_i - t_j - \gamma\right)e^{-\frac{1}{2}\alpha\left(t_i - t_j - \gamma\right)}$ for shaping parameters α and γ (Routledge et al., 2018).

2.3.3 Algorithm derivation

If we assume that infections are conditionally independent given the parents of infected nodes, then the likelihood of a given transmission chain can be defined as

$$f(\boldsymbol{t};\boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} f(t_i | t_1, \dots, t_n \backslash t_i; \boldsymbol{\alpha}) \ (1)$$

Where $\boldsymbol{\alpha}$ is a parameter matrix. Computing the likelihood of a given transmission chain thus involves computing the conditional likelihood of the infection time of each infection (t_i) given all other infections ($t_1, \dots, t_n \backslash t_i$). If we make the assumption that a node gets infected once the first parent infects it (Kempe, Kleinberg and Tardos, 2003) and define a survival function

$$S\left(t_i \middle| t_j; \alpha_{j,i}\right) = 1 - \int_0^{t_i - t_j} f\left(t_i \middle| t_j; \alpha_{j,i}\right) dt \ (2)$$

as the probability that infection I_i is *not* infected by infection I_j by time t_i then one can simplify the transmission likelihood as

$$f(\boldsymbol{t};\boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} \sum_{I_j : t_j < t_i} f(t_i | t_j; \alpha_{j,i}) \prod_{I_k : t_k < t_i, I_k \neq I_j} S(t_i | t_k; \alpha_{k,i}) \ (3)$$

In this conditional likelihood the first term computes the probability the I_j infected I_i and the second term computes the probability that I_i was not infected by any *other* previous infections excluding I_j. This likelihood therefore accounts for competing infectors and finds the infector most likely to have infected I_i. To remove the $k \neq j$ condition makes the product independent of j and results in the likelihood

$$f(\boldsymbol{t};\boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} \prod_{I_k:t_k<t_i} S(t_i|t_k;\alpha_{k,i}) \sum_{I_j:t_j<t_i} \frac{f(t_i|t_j;\alpha_{j,i})}{S(t_i|t_j;\alpha_{j,i})} \quad (4)$$

In equation 4, $f(\cdot)/S(\cdot) = H$ is the hazard function or rate and represents the instantaneous infection rate between individuals I_i and I_j.

Similar to the submodular approach used in Chapter 3 (Algorithm 1), to account for unobserved infectors within this framework I include a time-independent edge that can infect any individual (Figure 2.1). The survival and hazard functions for this edge are defined as $S_0(\epsilon_i) = e^{-\epsilon_i}$ and $H_0 = \epsilon_i$. As we will see below, as a consequence of the optimisation problem these edges are encouraged to be sparse and only invoked if no other infectors can continue the transmission chain.

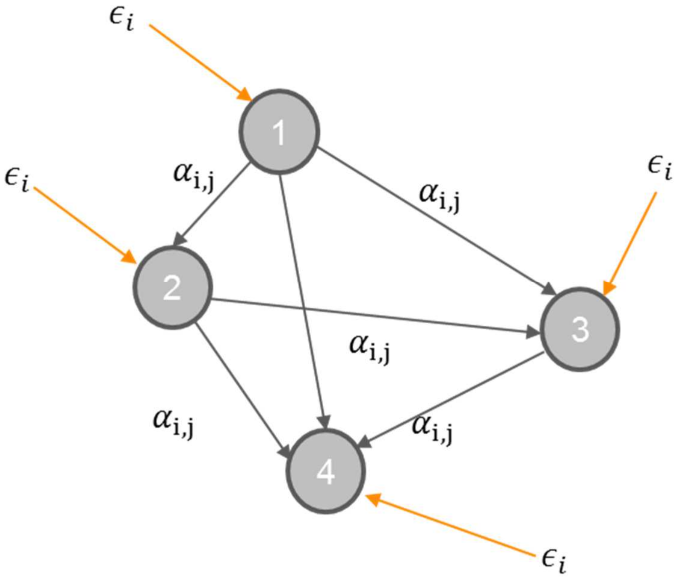


Figure 2.1: Diagram showing the parameters estimated by Algorithm 2. *The likelihood of transmission occurring between each pair of edges is determined by* α_{ij}*, representing a transmission rate/hazard and the* ε_i *estimated for each case, representing competing hazards from unobserved infectors.*

In addition to unobserved edges, we assume that observed imported infectors can infect other cases but cannot be infected themselves. The final likelihood incorporating these two modifications becomes

$$f(\boldsymbol{t};\boldsymbol{\alpha},\boldsymbol{\epsilon}) = \prod_{t_i \in \boldsymbol{t}} S_0(\epsilon_i) \prod_{I_k:t_k<t_i} S\left(t_i \middle| t_k;\alpha_{k,i}\right) \left(H_0(\epsilon_i) + \sum_{I_j:t_j<t_i} H\left(t_i \middle| t_k;\alpha_{k,i}\right)\right) \quad (5)$$

In order to find the optimal parameters for $\boldsymbol{\alpha},\boldsymbol{\epsilon}$ we minimize the following log likelihood subject to positive constraints on the parameters:

$$minimize_{\boldsymbol{\alpha},\boldsymbol{\epsilon}} - \log f(\boldsymbol{t};\boldsymbol{\alpha},\boldsymbol{\epsilon}) \qquad subject\ to\ \boldsymbol{\alpha},\boldsymbol{\epsilon} > 0\ \ \forall i,j \ (6)$$

This optimisation problem is convex and guarantees a consistent maximum likelihood estimate (Gomez Rodriguez *et al.*, 2014).

To prevent biologically implausible serial interval distributions, we impose a truncated normal prior probability distribution on $\boldsymbol{\alpha}$ ~Normal(0.003,0.1) [0,0.01]. When optimising the likelihood, I include this prior probability and therefore evaluate the Bayesian Maximum-a-Posteriori estimate.

2.3.4 Estimating R_c

We can establish individual reproduction numbers for each case by creating a matrix where each column represents a potential infector and the rows represent a potential infectee, describing which infector edges are connected to infectees and the normalised likelihood of the cases being connected by a transmission event. Intuitively then, by taking the row sums we get the (fractional) number of secondary infections and therefore a point estimate of the time varying reproduction number $R_c(t_j)$ This reflects for an individual, how many people they subsequently infect. When multiple individuals have been infected at a given time and/or place, we can take the mean individual R_c and uncertainty in this value as an indicator of reproduction numbers for a given time and/or location.

In contrast to traditional methods based on Wallinga and Teunis (2004) using the method in this way encapsulates not only the pairwise likelihood of transmission between two cases, but conditions this likelihood on the impact of competing edges in the inferred network (the survival of an edge). The estimates of R_c therefore consider the overall transmission tree in optimisation and allow for missing cases within the tree.

2.3.5 Alternative versions

Throughout the development of this work, several versions of the algorithm were devised and tested, mainly with the aim of adapting of the methods to suit larger datasets, or contexts with varying levels of uncertainty/information around key model inputs, where a Bayesian framework may be useful.

The parameters were estimated both within a frequentist framework by Maximum Likelihood Estimation (MLE) using a bounded Broyden-Fletcher-Goldfarb–Shanno (BFGS-B) algorithm to optimise the negative log likelihood and within a Bayesian framework using Hamiltonian Markov-Chain Monte Carlo methods (Duane *et al.*, 1987) in Stan, a C++ based language designed for efficient Hamiltonian MCMC sampling which was implemented through the rStan package (Stan Development Team, 2016).

By working within a Bayesian framework, this approach allows the incorporation of prior knowledge around the serial interval, allowing better quantification of uncertainty, as for many outbreaks and infectious diseases there is some information about the serial interval from epidemiological investigation/natural history of the pathogen, but also a certain amount of variation and uncertainty.

For increased speed and computing efficiency, allowing the analysis of larger datasets, the model was rewritten and implemented in TensorFlow (Abadi *et al.*, 2015) both as a frequentist and Bayesian model, where the maximum-a-posteriori estimate was calculated. TensorFlow is an opensource library for numerical computation which is coded in the Python programming language but runs all numerical computation in C++.

2.4 Algorithm 3: Network transmission rate inference incorporating distance metrics (implemented in Chapter 5)

Following on from Algorithm 2, an approach was required which could incorporate additional information, such as Euclidian distance and accessibility information within one inference framework. In order to incorporate features other than time, I extended the method by introducing a second function, f_2, which describes the relationship between space (or distance of any kind) and likelihood of transmission. An appropriate function such as a power law distribution is decided and the parameters shaping that distribution, are estimated from the data. Together, the product returns a single function:

$$f\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) = f_1(t_i|t_j;\ \alpha_{i,j})\ f_2(x_i|x_j; \beta)$$

Determined by times t, spatial locations x, transmission rates $\alpha,$ spatial parameter(s) $\beta.$ The specific functions used in $f_1(t_i|t_j;\ \alpha_{i,j})$ and $f_2(x_i|x_j; \beta)$ impact the outcomes of results and therefore the assumptions inherent in these choices must be made explicit and linked to the mechanisms of transmission. In this thesis, two functions were used to model the relationship between space and the likelihood of transmission: Exponential and Gaussian Kernels.

Table 2.1: Summary of key equations for Algorithm 3. *Equations for f1, f2, hazard and survival for time -only, and spatial versions of the algorithm with Gaussian and Exponential transmission kernels*

	$\boldsymbol{f_1(t_i \vert t_j;\ \alpha_{i,j})}$	$\boldsymbol{f_2(x_i \vert x_j; \beta)}$	**Hazard**	**Survival**
Exponential	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}$	$e^{-\beta(x_i-x_j)}$	$\beta\alpha(t_i - t_j - \gamma)e^{-\beta(x_i-x_j)}$	$e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}\frac{1}{\beta}$
Gaussian	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}$	$e^{-\beta(x_i-x_j)^2}$	$\frac{2\sqrt{\beta}\alpha(t_i - t_j - \gamma)e^{-\beta(x_i-x_j}}{\sqrt{\pi}}$	$e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}\frac{\sqrt{\pi}}{2\sqrt{\beta}}$
Time only	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}$	n/a	$\alpha(t_i - t_j - \gamma)$	$e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}$

2.4.1 Derivation of hazard, survival and likelihood

The pairwise likelihood of a case showing symptoms at t_i and at residence location x_i being infected by a case showing symptoms at time t_j and at residence location x_j, becomes

$$f(x_i, t_i|x_j, t_j; \alpha_{i,j}, \beta) = \alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}e^{-\beta(x_i-x_j)} \quad (4)$$

The survival term is then the integral over all a time range and the real line of distances:

$$S(x_i, t_i|x_j, t_j; \alpha_{i,j}, \beta) = (\int_{x_j=0}^{\infty} \int_{t_j=0}^{t_i} \alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i-\mathrm{t_j}-\gamma)}e^{-\beta(x_i-x_j)} \ dt\, dx \quad (5)$$

Which simplifies to:

$$\mathrm{S}(x_i, t_i|x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)} \int_{x_j=0}^{\infty} e^{-\beta(x_i-x_j)} \quad dx \ (6)$$

$$\mathrm{S}(x_i, t_i|x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i-t_j-\gamma)}\frac{1}{\beta} \ (7)$$

Following on from this, as the hazard is equivalent to the likelihood divided by survival, $H = \frac{f(\cdot)}{S(\cdot)}$, it follows that

$$H(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \frac{\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}e^{-\beta(x_i - x_j)}}{e^{-\alpha(t_i - t_j - \gamma)\frac{1}{\beta}}} \text{ (7)}$$

Which simplifies to

$$H(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \beta\alpha(t_i - t_j - \gamma)e^{-\beta(x_i - x_j)} \text{ (8)}$$

For the Gaussian function, the pairwise likelihood of a case showing symptoms at t_i and at residence location x_i being infected by a case showing symptoms at time t_j and at residence location x_j is

$$f(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}e^{-\beta(x_i - x_j)^2} \text{ (9)}$$

The survival term is again determined by integrating the likelihood over all potential infection times and all distances:

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = (\int_{x_j=0}^{\infty} \int_{t_j=0}^{t_i - t_j} \alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}e^{-\beta(x_i - x_j)^2} \; dt \, dx \text{ (10)}$$

Integrating over time returns:

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} \int_{x_j=0}^{\infty} e^{-\beta(x_i - x_j)^2} \; dx \text{ (11)}$$

Integrating over all distances gives

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} \frac{\sqrt{\pi}}{2\sqrt{\beta}} \text{ (12)}$$

Following equation 12, the hazard is equivalent to

$$H\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) = \frac{\alpha\left(t_i - t_j - \gamma\right)e^{-\frac{1}{2}\alpha\left(t_i - t_j - \gamma\right)}e^{-\beta(x_i - x_j)^2}}{e^{-\frac{1}{2}\alpha\left(t_i - t_j - \gamma\right)}\frac{\sqrt{\pi}}{2\sqrt{\beta}}} \quad (13)$$

Which simplifies to

$$H\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) = \frac{2\sqrt{\beta}\alpha\left(t_i - t_j - \gamma\right)e^{-\beta(x_i - x_j)^2}}{\sqrt{\pi}} \quad (14)$$

2.4.2 Modelling missing cases using Ɛ *edges*

The vast majority of disease surveillance and outbreak response datasets will not be able to capture all cases due to asymptomatic infection, underreporting and movement of people in/out of the surveillance area. Therefore, it is important to consider the impact of missing information on results and identify potential missing sources of infection. In the work described in this chapter, as in chapter 2, we use Epsilon edges, ϵ_i , to identify potential sources of infection. Here, each hazard is estimated as a further competing edge of transmission from an unobserved source, $H_0(\epsilon_i)$. Depending on assumptions for the likelihood and extent of unobserved infection sources, the epsilon edge value can be set to a high or low value. When high, we assume high amounts of unobserved infection and unless two cases have a very high likelihood of being linked, we assume the case was from an unobserved source. When low, we assume little missing data and so cases are only linked to an outside source if they are very unlikely to be linked to an observed candidate infector.

Adding epsilon as a competing hazard and survival returns

$$f(\boldsymbol{t},\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\epsilon},\boldsymbol{\beta}) = \prod_{t_i \in \boldsymbol{t}} S_0(\epsilon_i) \prod_{I_k : t_k < t_i} S\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) \left(H_0(\epsilon_i) + \sum_{I_j : t_j < t_i} H\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right)\right) \quad (15)$$

The objective function is then

$$minimize_{\boldsymbol{\alpha},\boldsymbol{\epsilon}} - \log f(\boldsymbol{t},\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\epsilon},\boldsymbol{\beta}) \qquad subject\ to\ \boldsymbol{\alpha},\boldsymbol{\epsilon},\boldsymbol{\beta} > 0\ \ \forall i,j \quad (16)$$

2.5 Evaluation and comparison of methodologies

Simulations were carried out to explore the impact of various assumptions on the ability of the model to recover correct reproduction number estimates and serial intervals. Two approaches were used: firstly, simulating epidemics along explicit networks using a network based susceptible-infected model, and secondly using a stochastic susceptible-infected-recovered (SIR) model with a given R_c distribution to simulate line lists.

2.5.1 Simulation across networks

For the first simulation, data were simulated by generating small-world networks using the *igraph* (Csardi and Nepusz, 2006) package in R version 3.3 (R Core Team, 2016). Small world networks are hypothesised to reflect many real-life networks, which show both properties of regularity and randomness (Watts and Strogatz, 1998; Eubank *et al.*, 2004). The network generated for this analysis is illustrated in Figure 2.2. Then a susceptible-infected (SI) model was run along the network, where during each time step infected nodes infect their neighbours with probability β.

Under the SI model, at time zero $(t = 0)$, all nodes begin susceptible, bar a given number of seed nodes. For this simulation, initially one node was seeded with infection. At $t = 1$ the infected node can infect each neighbouring node which shares an edge with it (determined by the simulated network), with probability β. For this simulation β is constant, provided the infected node is connected by an edge to a susceptible node. At $t = 2$ if any new infections occur, the newly infected nodes then become able to infect their neighbours with probability β. The chain continues for a set horizon of time or until all nodes are infected. The incidence time series generated by this simulation was then input into a frequentist version of the algorithm.

Two factors were measured to explore the accuracy and effectiveness of the algorithm. Firstly, the mean α_{ij}value returned by the model, which is defined as the instantaneous hazard of infection, which for an exponential parameterisation is not time dependent. The true alpha value was assumed to be β, the hazard of infecting neighbours. Secondly, the corresponding likelihood functions calculated from the hazard value, determining likelihood of transmission over time, were also compared.

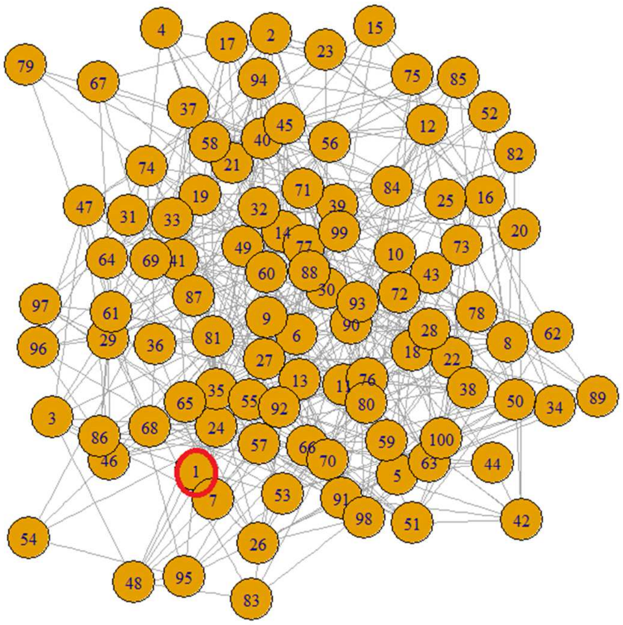


Figure 2.2 Network used for simulation. *Note the edges here represent potential connections and routes along which transmission could occur. Node 1, circled in red always seeded transmission.*

2.5.2 Stochastic SIR simulation of line lists

To further test assumptions in model, line lists with missing data were simulated using *EpiGen*R, an algorithm and R package which simulates transmission events and then samples from this to represent a final detected line list. This model implements a stochastic Susceptible-Infected-Recovered model over discrete time steps in the C++ language via the *Rcpp* package. Recovery is exponentially distributed, with rate parameter, γ. This parameter determines the time to infection of the next generation and in turn the serial interval distribution. Infectors infect a number of individuals, drawn from the offspring or reproductive number distribution, which is negative binomial with dispersion parameter K.

To reflect an elimination scenario, the distribution of individual reproduction numbers was defined as a negative binomial distribution with mean 0.5. I considered two values for the overdispersion parameter, K, as 0.1 (more over dispersed, more variance in R_c) and 1 (less over dispersed, less variance in R_c). For both values of K 100 outbreaks of minimum infected size 100 were simulated over 1000 days, with an exponentially distributed serial interval with a mean of 30 days. Then the ability of the algorithm to detect the underlying offspring distribution was measured. Each outbreak had 100 seed infectors in a fully susceptible population of 50 000, with no further importation occurring, to ensure the final sample size was large enough to measure R_c. As the simulator draws integers, for better comparison of model estimated results, the distribution of maximum-a-posteriori estimates for R_c estimates were rounded to the nearest integer and presented alongside the raw estimates. Both the histograms and means of simulated versus estimated results were compared.

To simulate missing cases, the fully observed dataset was sampled following a proportional approach where for each case the probability of observation was set at varying values between 1 and 0.3, and then each individual observation was determined by drawing from a binomial distribution with the given probability.

2.5.3 Simulation Results

Simulated data on a small world network found the inferred mean $\alpha_{i,j}$, or instantaneous hazards of transmission to be relatively similar to expected values, as shown in Figure 2.3. The corresponding likelihood of infection also closely resembled the true likelihood (Figure 2.4), assuming the same parametric form (an exponentially distributed likelihood, determined by mean α_{ij} and time).

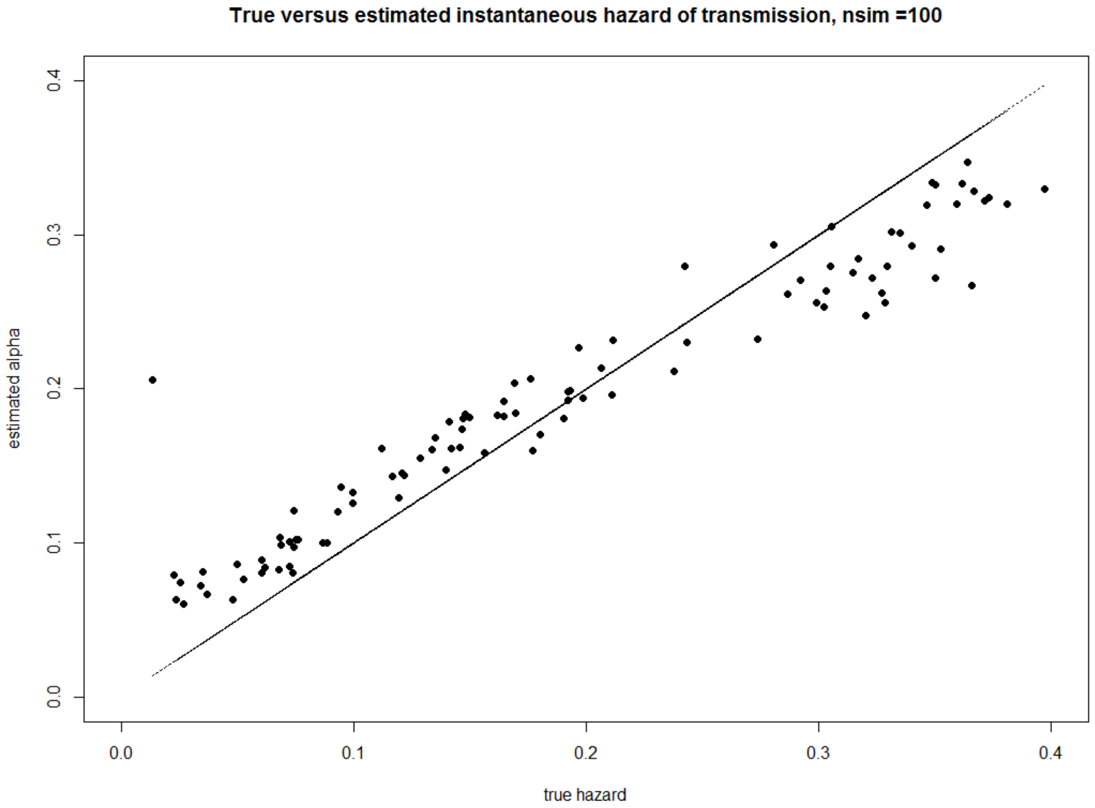


Figure 2.3 Plot of true transmission rate, β plotted against model estimated transmission rate (mean $\alpha_{i,j}$ or hazard) for 100 simulations of line lists with different values of β

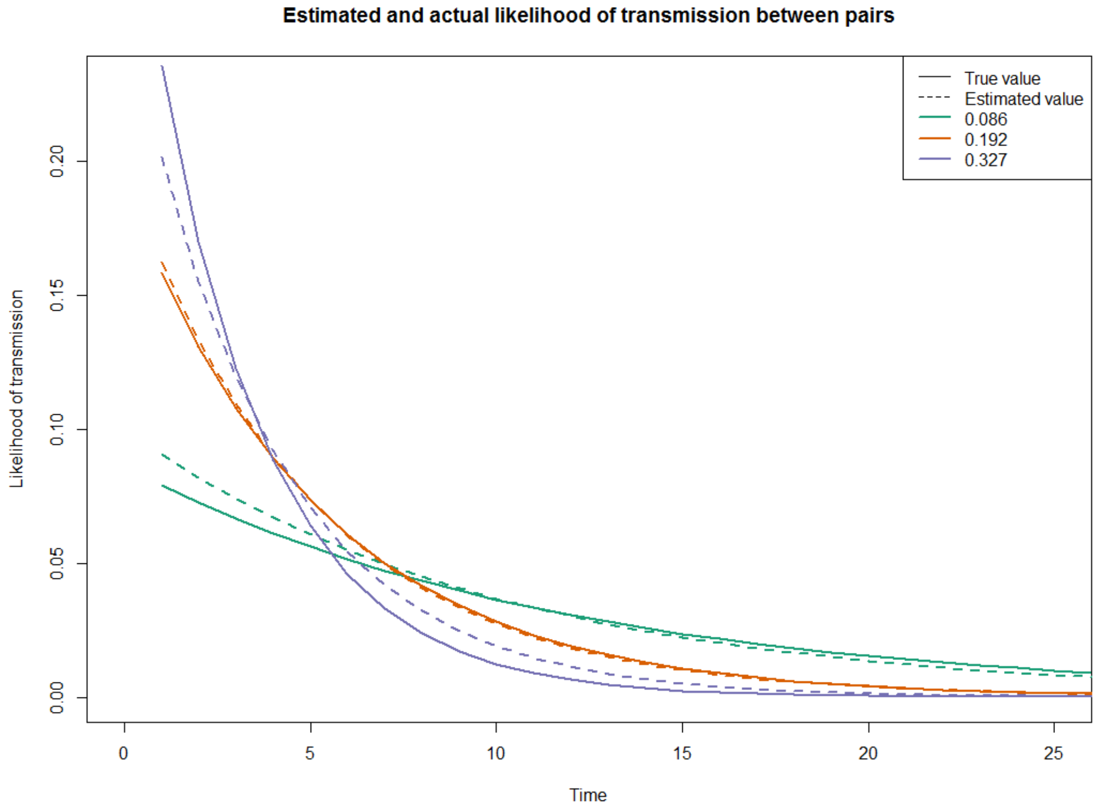


Figure 2.4 Three randomly drawn estimated and actual transmission rate/hazard values from figure 2.3 showing the corresponding estimated and actual transmission likelihoods they represent. *Colours show likelihoods of transmission over time for different values of actual α (solid line) and their corresponding estimated values (dotted line).*

2.5.3.1 *Simulations from a more over dispersed R distribution (K=0.1)*

When the probability of observing a case was 1, $P(case\ observed) = 1$, simulated line lists, simulated from a negative binomial R_c distribution of mean (μ) 0.5, with overdispersion parameter (K), of 0.1 ($R_c \sim Negative\ Binomial(\ \mu = 0.5, K\ = 0.1)$) had a true mean R_c of 0.56. When the prior for the ε edge was defined as having a Truncated Normal prior with mean = 0.001 and standard deviation = 1, ($prior(\varepsilon){\sim}\ Truncated\ Normal(\ \mu = 0.001,\ = 1)$) the algorithm returned a mean of 0.54 when results were rounded to the nearest integer and 0.6 when decimal values were not rounded (Figure 2.5). When the probability of observation was 90%, this value decreases to 0.52 (rounded) and 0.56 (decimal). The mean estimate continues to decrease with decreasing observations, but even with an average of 30% of cases observed, the mean R_c was estimated as 0.41 and 0.46 when R_c is a rounded integer or decimal estimate respectively (Figure 2.8).

2.5.3.2 Simulations from a less over dispersed R distribution (K=1)

When the probability of observing a case was 1, $P(case\ observed) = 1$, line-lists, simulated from a negative binomial R_c distribution of mean (μ) 0.5, with overdispersion parameter (K), of 1 ($R_c \sim Negative\ Binomial(\ \mu = 0.5, K = 1)$) had a true mean R_c of 0.59. When the prior for the ε edge was defined as having a Truncated Normal prior with mean = 0.00001 and standard deviation = 1, ($prior(\varepsilon) \sim Truncated\ Normal(\ \mu = 0.00001, K = 1)$), the algorithm returned a mean of 0.54 when results were rounded to the nearest integer and 0.53 when decimal values were not rounded (Figure 2.9). When the probability of observing a case was 0.9 ($P(case\ observed) = 0.9$), this value decreases to 0.49 (rounded) and 0.52 (decimal). When an accurate and informative prior for ε when ($P(case\ observed) = 0.9$) is chosen, the model accurately returns the mean R_c of 0.59 (Figure 2.11). With an average of 30% of cases observed, the mean R_c was estimated as 0.41 and 0.44 when R_c is a rounded integer or decimal estimate respectively. Observationally, the distribution of R_cs remain similar to the true value (Figure 2.11), however more quantitative analysis would be required to rigorously assess similarities in the distributions.

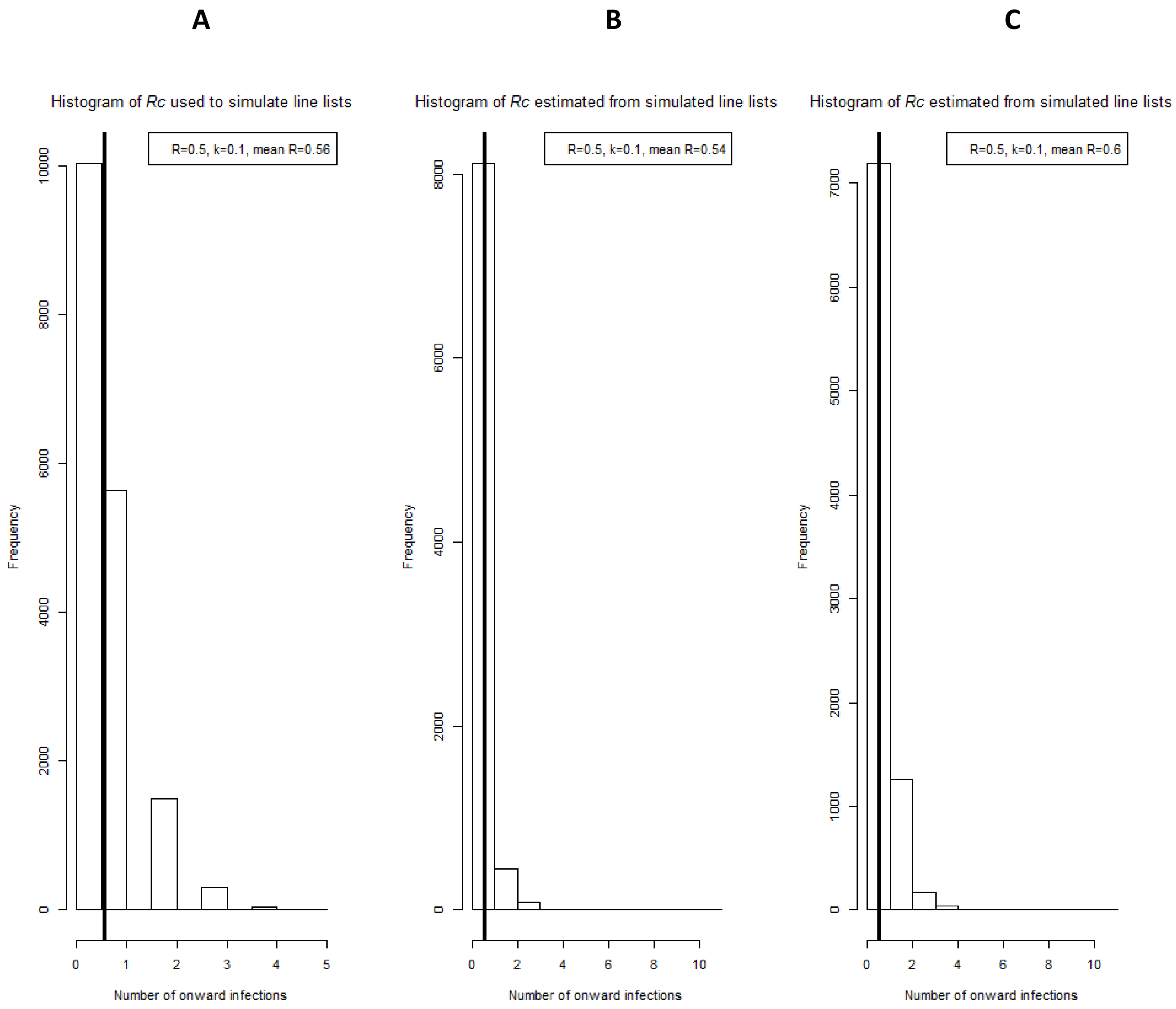


Figure 2.5: Histograms of simulated Rc and model-estimated Rc when P(observation) is 1 and K is 0.1

A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used

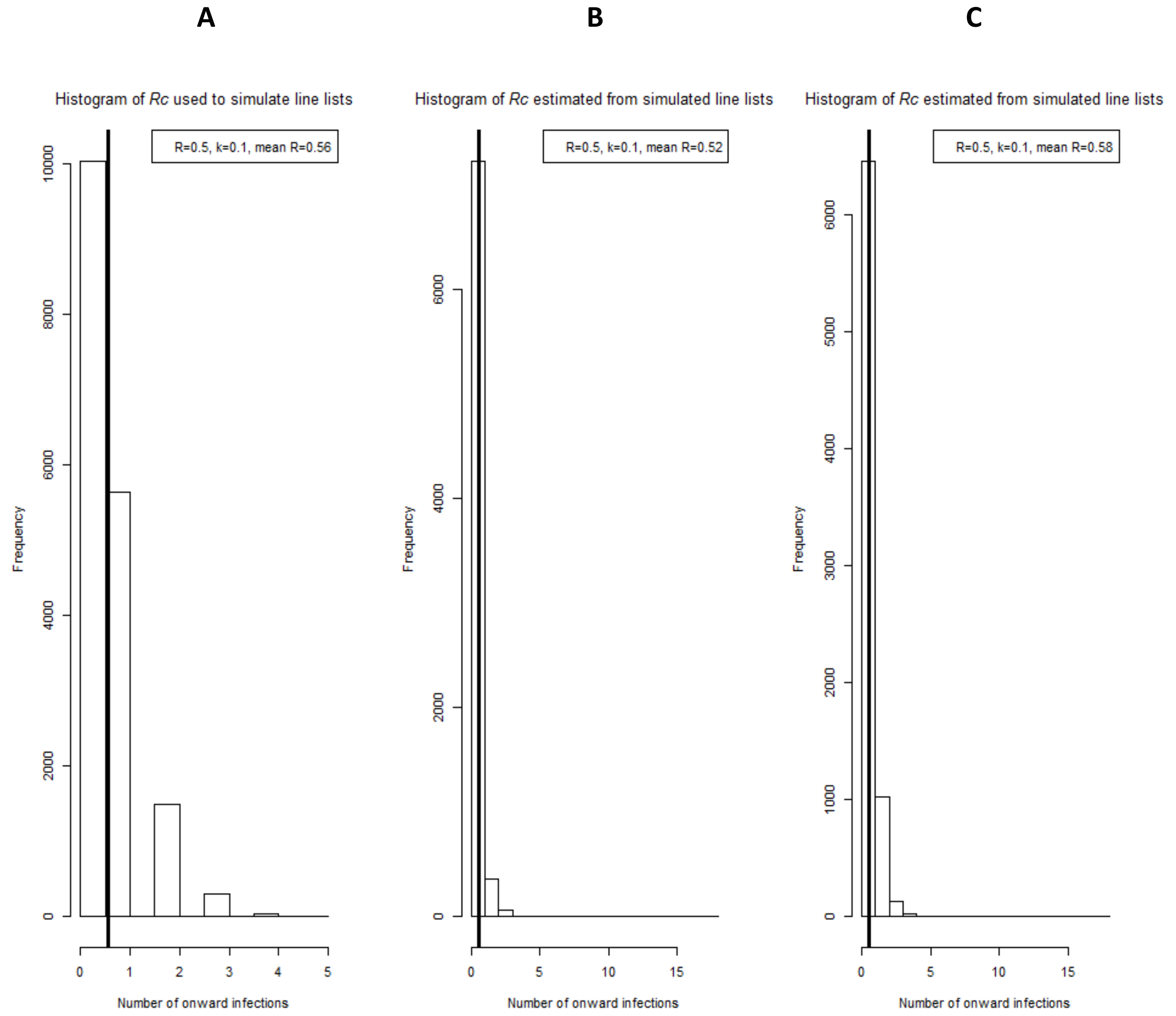


Figure 2.6 Histograms of simulated Rc and model-estimated Rc when P(observation) is 0.9 and K is 0.1

A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used

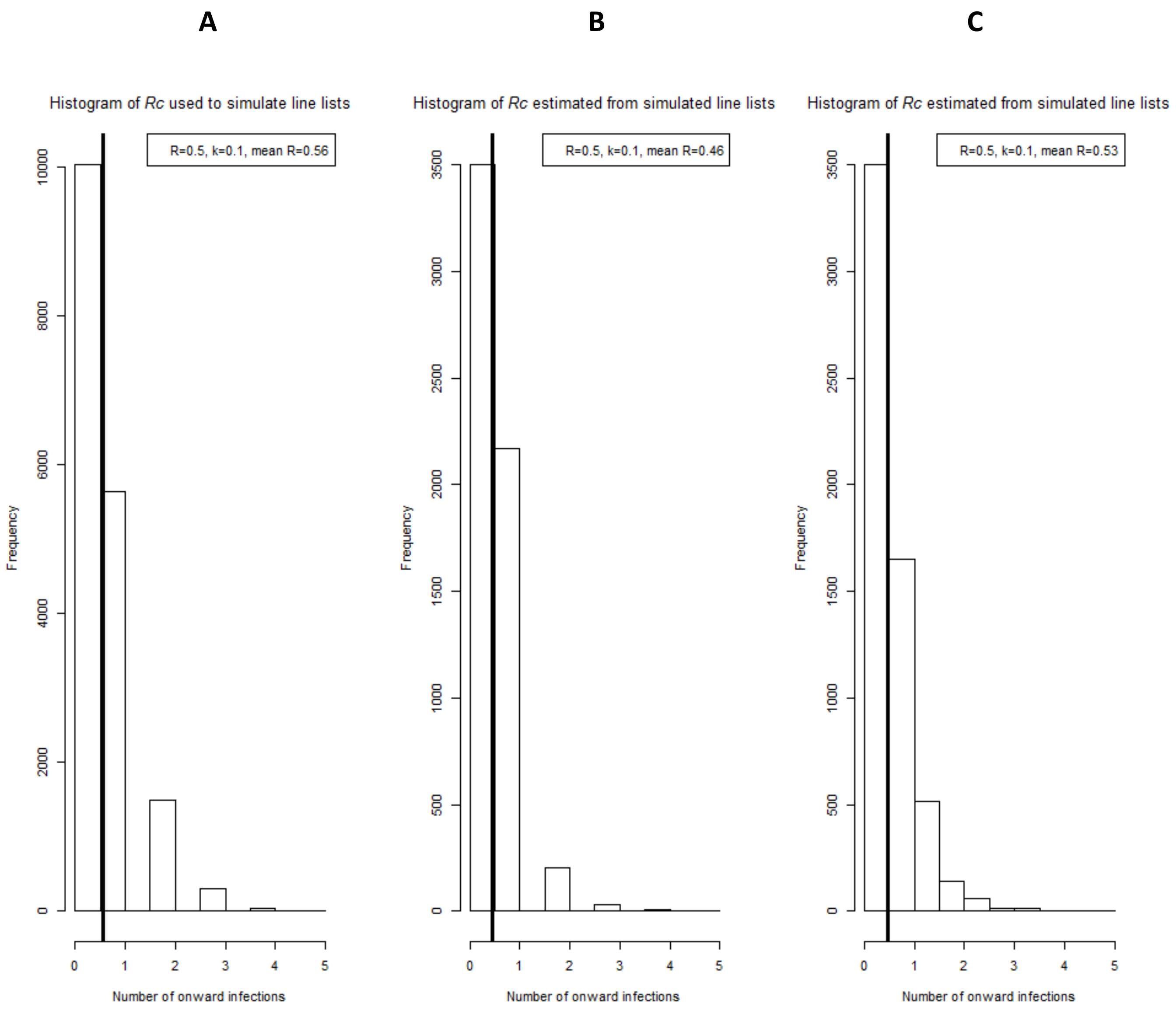


Figure 2.7: Histograms of simulated Rc and model-estimated Rc when P(observation) is 0.7 and K is 0.1

A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used

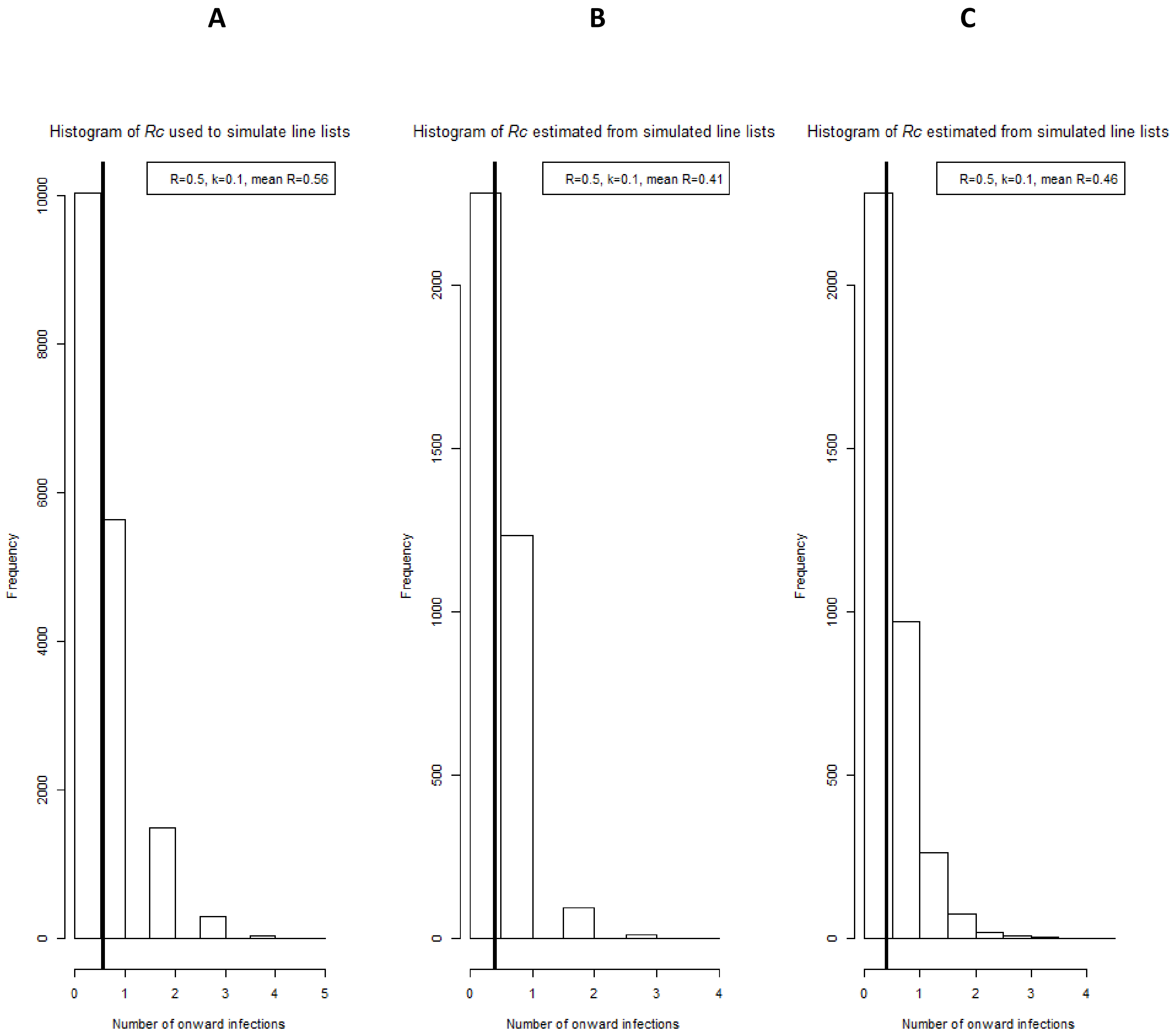


Figure 2.8: Histograms of simulated Rc and model-estimated Rc when P(observation) is 0.3 and K is 0.1

A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used.

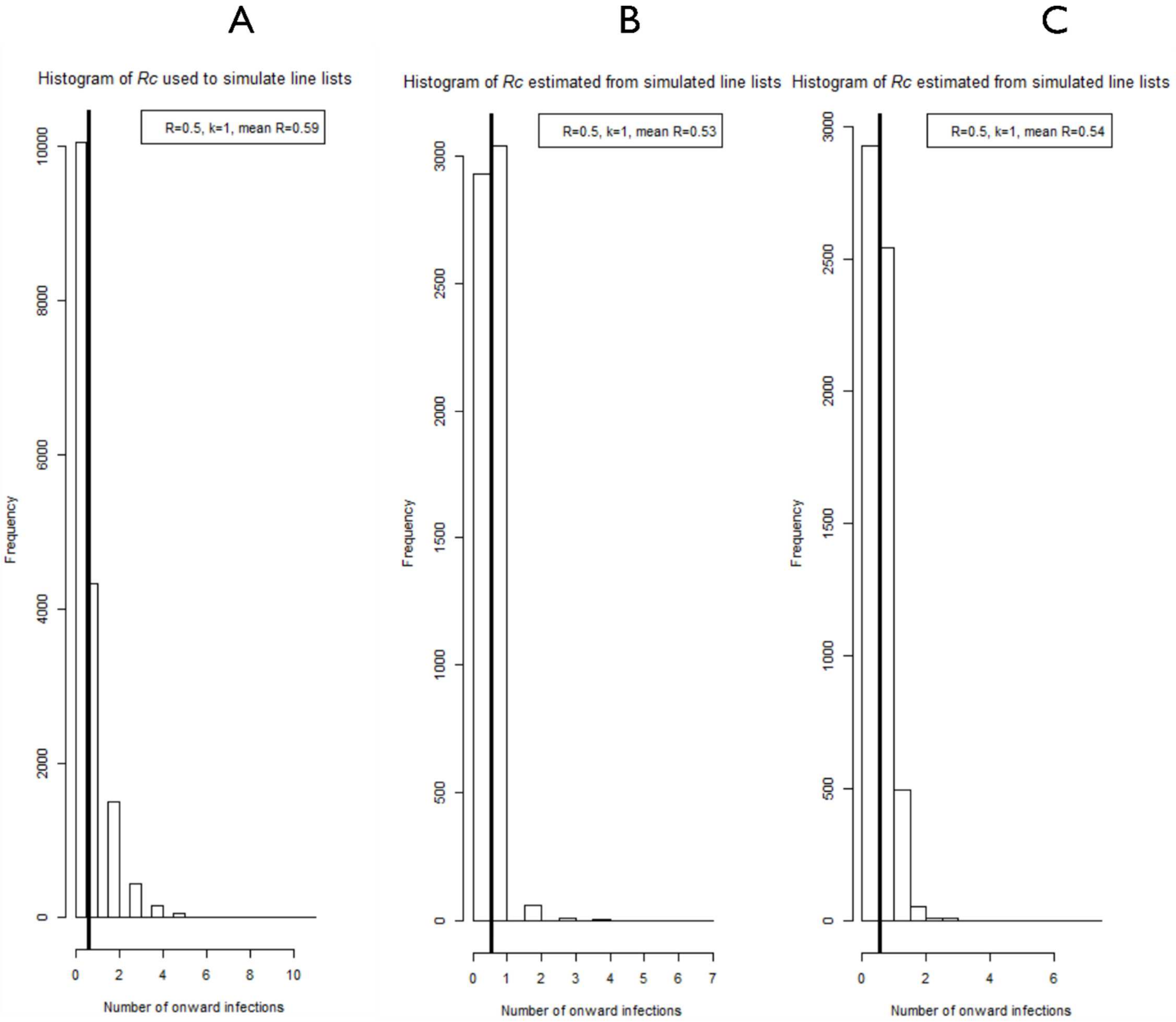


Figure 2.9: Histograms of simulated Rc and model-estimated Rc when P(observation) is 1 and K is 1

When P(case observed) = 1.0 and an uninformative prior used for ε (Truncated Normal(mean=0.0001,standard deviation=1). A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used.

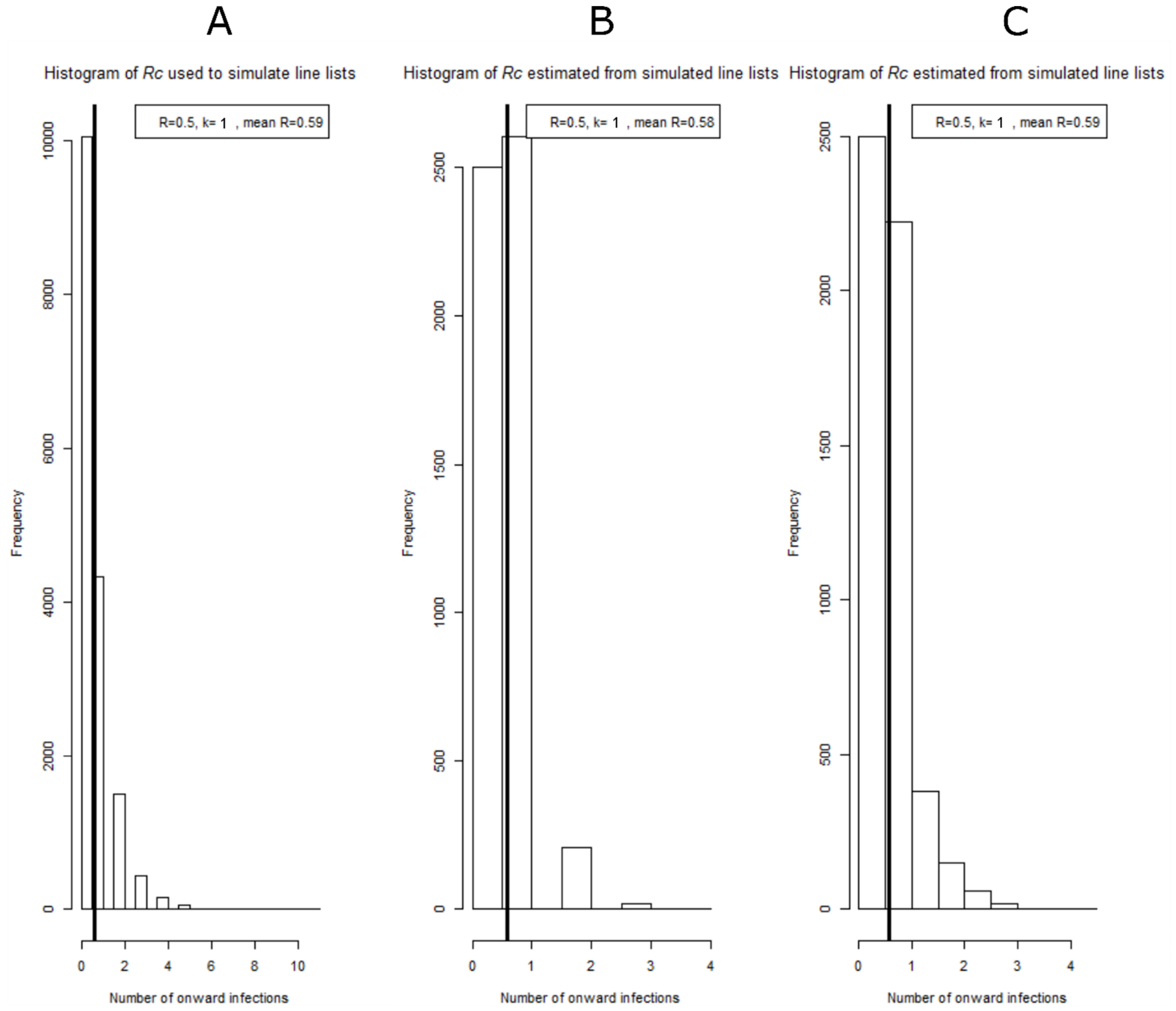


***Figure 2.10:** Histograms of simulated Rc and model-estimated Rc when P(observation) is 0.9 and K is 1*

When informative and accurate prior used for ε, Truncated Normal(mean=0.1, standard deviation=0.00001 A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used.

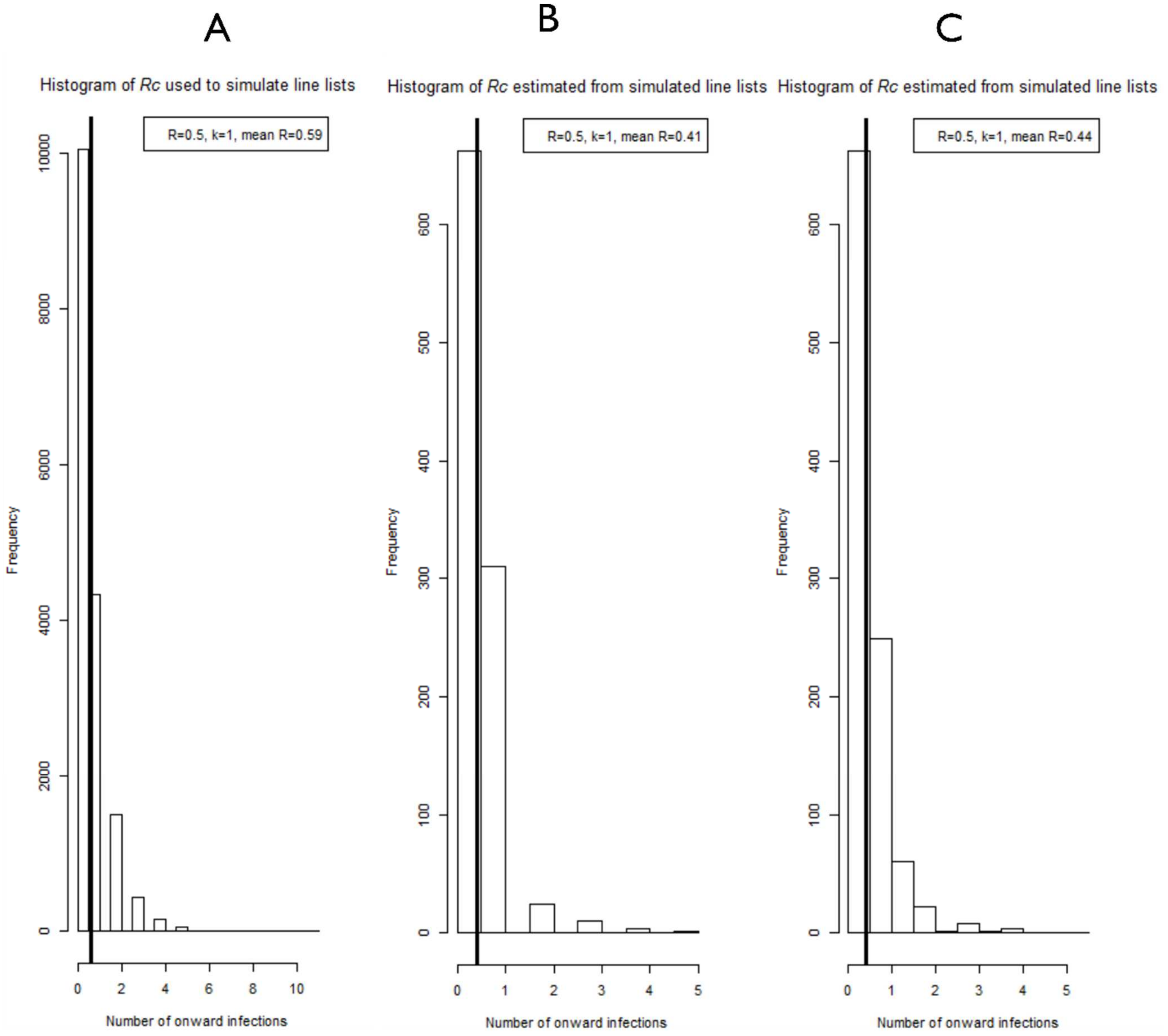


***Figure 2.11:** Histograms of simulated Rc and model-estimated Rc when P(observation) is 0.9 and K is 1*

When P(case observed) = 0.9 and an uninformative prior used for ε (Truncated Normal(mean=0.0001,standard deviation=1). A) Histogram of individual reproduction numbers from simulated data, B) model estimated results when estimates rounded to the nearest integer and C) when decimal estimates are used.

2.6 Discussion

This chapter introduced the key algorithms utilised in this thesis and tested the approach on several forms of simulated data. Firstly, the algorithm utilised in Chapter 3 on data from El Salvador was derived and described. This approach defines a range of the shaping parameters defining serial interval distributions for symptomatic, treated *P. vivax* malaria and samples from this to define the most likely route of transmission between cases, based on the time of infection and the likelihood of any case having an unobserved source of infection. This approach uses a greedy algorithm which uses pairwise likelihoods of transmission to build a transmission tree. Then the consensus or average connectivity between cases, as well as uncertainty around this estimate can be calculated. This approach is demonstrably suited to smaller observations of cases (Rodriguez and Schölkopf, 2012), but is heuristic, and harder to incorporate other sources of information within one statistically rigorous framework. In addition, it was not suited to use with large datasets due to computational running time. Therefore, this approach was built upon by the second algorithm introduced, which was a fully Bayesian framework implemented within TensorFlow for efficient inference from larger datasets. Then this algorithm was extended to incorporate features other than time of symptom onset, primarily Euclidian distance between cases, but as will be discussed in detail in Chapter 5, offers flexibility to incorporate other metrics such as accessibility matrices, travel times, or genetic distance.

Simulations were carried out to test Algorithm 2 (the time-only network rate inference approach). It was found that the model was robust to missing data when up to 30% of data were missing. However, these simulations have not extensively explored whether this is robust to different parameterisations for the serial interval or for epsilon edges, representing missing sources of infection.

There are several limitations to the simulations and findings. An important limitation of the simulation was that all infections occur at the end of the designated "infectiousness period". Whilst this is a draw from a distribution and varies for each case, it still makes the temporal signal of transmission potentially more identifiable than in a real epidemic.

There is also a limitation to how missing data was simulated in this chapter. Missing cases here are not biased in any way. Sampling was carried out in a random and proportional way and so at each time point, of all cases a proportion of those cases will be missing. This potentially could have less of an impact on inferred results compared to biased missingness (not at random), and in reality it may be that individuals with persistent malaria infections (e.g. due to asymptomatic infection, lack of access to healthcare) are less likely to be detected by surveillance systems. This could be explored in further simulations which sample the fully observed dataset in non-random ways.

It is important to note that the probability distribution of the serial interval used to simulate line lists is different to the assumptions made in our approaches. Namely, the simulation uses an exponential distribution whereas the algorithms developed in this thesis use a shifted Rayleigh distribution. Given the inherent uncertainty and variability in the serial interval of malaria, it is reassuring that this approach can approximately recover reproduction number distributions despite different assumptions about the serial interval. Future work to use a Rayleigh distribution would be helpful to compare like for like and ensure that any divergence between actual and estimated R_c values is error in the approach, rather than the result of slightly different assumptions.

It is also important to note that outbreaks were not simulated over space and therefore Algorithm 3 was not evaluated here. However, in order to address this, Chapter 5 does include a detailed sensitivity analysis, exploring the interaction between parameters and the impact of priors used.

3

Estimating spatiotemporally varying malaria reproduction numbers in El Salvador, a near elimination setting

3.1 Introduction

As introduced in Chapter 1, great strides have been made since 2000 in reducing the burden and mortality of malaria. The World Health Organisation (WHO) estimates that 57 out of the 106 countries with endemic malaria transmission in 2000 reduced their incidence of malaria by more than 75% between 2000 and 2015 (Cibulskis *et al.*, 2016). As a result, malaria elimination at the national level, defined as the absence of local transmission within a country (Cohen *et al.*, 2010), is now one of the targets in the WHO Global Strategy for Malaria 2016-2030 (Griffin *et al.*, 2016). In 2016 the WHO identified 21 countries for which it would be realistic to eliminate malaria within the next five years (WHO, 2016).

As countries attempt to move towards malaria elimination, tracking progress through quantifying changes in transmission over space and time is key. This information is necessary to effectively target resources to remaining 'hotspots' and 'hotpops' (Sturrock *et al.*, 2013) where transmission remains, decide if and when it is appropriate to scale back interventions, and to evaluate the success of existing interventions. However, as countries approach zero cases, increasing focality in transmission and the impact of imported cases pose challenges to both reaching elimination (Cotter *et al.*, 2013) and measuring progress towards that goal. Increased spatial and temporal heterogeneity in malaria cases (Carter, Mendis and Roberts, 2000; Bousema *et al.*, 2012; Sturrock *et al.*, 2016) in low transmission settings reduces the usefulness of national or regional level trends in incidence or prevalence, which can mask small areas of high transmission intensity. Furthermore, end-game surveillance and control measures are increasingly expensive per case. Therefore, while interventions must be targeted efficiently to be cost-effective (Carter, Mendis and Roberts, 2000; Bousema *et al.*, 2012), the identity of areas driving remaining transmission and their stability over time are poorly understood.

As touched on in Chapter 1, a wide variety of contextually varying factors have been hypothesised to drive transmission in low transmission settings, including increased risk in concentrated populations due to factors such as occupation (e.g. agricultural workers) (Cotter *et al.*, 2013), asymptomatic individuals acting as reservoirs of infection (Sturrock *et al.*, 2013; Bousema *et al.*, 2014), changes in vector behaviour (Moiroux *et al.*, 2012) and resistance to antimalarial (Dondorp *et al.*, 2009) and insecticidal interventions (Sokhna, Ndiath and Rogier, 2013). Importation of malaria cases from neighbouring countries poses an additional challenge in many elimination settings. If many cases of malaria are imported, control measures may appear less effective due to small numbers of locally-acquired cases arising from imported cases (Blumberg *et al.*, 2013; Churcher *et al.*, 2014). If there is sufficient importation, local cases can continue to occur even when the reproduction number of malaria under control, $\mathcal{R}_c$, is below 1. Conversely areas with a high underlying $\mathcal{R}_c$ but little importation may see sudden outbreaks of cases following a rare importation event due to their receptivity to malaria infection (Patel *et al.*, 2014). Challenges arise in measuring the sustainability of elimination (Cotter *et al.*, 2013; Churcher *et al.*, 2014), both in terms of quantifying the impact of control measures on transmission in the lead up to elimination, and in determining the risk of resurgence once elimination is achieved (Cohen *et al.*, 2012; Chiyaka *et al.*, 2013; Smith *et al.*, 2013). This information is also important when deciding if, when, and how to scale back intervention and surveillance methods (Chiyaka *et al.*, 2013).

Meeting these challenges requires measuring changes in transmission, often at fine spatial scales. However, existing methods used to quantify malaria transmission are poorly suited to elimination settings (Sturrock *et al.*, 2016). Parasite prevalence rates (PR) are not accurate below a PR of 1-5% (Yekutil, 1980; Hay, Smith and Snow, 2008) due to the large sample sizes necessary for precise estimates at low prevalence. The entomological inoculation rate (EIR), often seen as the "gold standard" in measures of transmission intensity, is not reliable when transmission is highly focal and potentially unstable since EIR is very sensitive to heterogeneities in vector populations (Hay *et al.*, 2000; Mbogo *et al.*, 2003). Use of serological data, whilst promising (Corran *et al.*, 2007; Dewasurendra *et al.*, 2017; Yalew *et al.*, 2017), is not currently feasible for use in many very low transmission contexts, as suitable cross-sectional survey data and/or appropriate markers to determine changes in malaria transmission are not available in all contexts.

A possible alternative, or complementary, measure of malaria transmission is the incidence of malaria cases, obtained through routine surveillance by Ministries of Health. Surveillance data are widely collected and sensitive to short term changes in transmission. Whilst utilising these data can pose challenges, particularly in low-resource settings due to limitations in surveillance infrastructure and difficulty in establishing completeness of reporting, they can provide a wealth of information when such challenges are overcome. Individual level incidence data can be used to reconstruct the most likely pathways of transmission and estimate individual reproduction numbers, providing fine-scale insights into spatiotemporal transmission characteristics. Whilst individual level surveillance data is often used in outbreak analysis of epidemic infections (Wallinga and Teunis, 2004; Jombart *et al.*, 2014), robust methods are rarely applied to vector-borne diseases such as malaria, with a few notable exceptions (Churcher *et al.*, 2014; Reiner *et al.*, 2015; Salje *et al.*, 2016).

In this chapter I aim to estimate individual reproduction numbers over time and space by adapting methods from the study of information diffusion processes described fully in Chapter 2 and reviewed in section 3.2. of this chapter. These methods address the general problem of reconstructing information transmission using known or inferred times of infection by a 'contagion' (Kempe, Kleinberg and Tardos, 2003; Gomez-Rodriguez, Leskovec and Krause, 2010; Rodriguez and Schölkopf, 2012; Gomez Rodriguez *et al.*, 2014). They provide an adaptable framework to integrate multiple data types (Wang, Ermon and Hopcroft, 2012), identify likely unobserved cases/external infection sources, and have been evaluated using real and simulated transmission processes at multiple scales and network structures (Gomez Rodriguez *et al.*, 2014).

3.1.1 Malaria elimination in Central America

Mesoamerica has made large strides towards malaria elimination over the past twenty years. Cases in Mesoamerica declined from roughly 123 000 in the year 2000 to roughly 10 000 cases in 2015 (Herrera *et al.*, 2015) despite population growth, and strengthened surveillance and case detection systems which likely increased the proportion of cases which were reported. However the need for continued effort has been highlighted by recent halts in progress, with over 16 000 cases reported in the region in 2017(WHO, 2018a). The potential for elimination in the region led to the formation of a regional eradication programme, Elimination of Malaria in Mesoamerica and Hispaniola (EMMIE: Eliminación de Malaria en Mesoamerica y la Isla Española) in 2014, which aims to achieve zero cases of locally transmitted malaria in Mesoamerica by 2020 (Herrera *et al.*, 2015). Half of the 8 countries which form this area (Belize, Costa Rica, El Salvador and Mexico) have been designated by the WHO as likely to eliminate malaria by 2020 (WHO, 2016). Nonetheless countries including Panama, Nicaragua, Honduras and Guatemala still are in the control phase, with substantial levels of transmission still occurring, particularly in north eastern coastal areas of Nicaragua, south eastern coastal areas of Honduras and some western and eastern coastal areas of Guatemala (Carter *et al.*, 2015; WHO, 2018a). No country in Central America has yet been certified malaria-free.

3.1.2 Malaria elimination in El Salvador

In 1980, El Salvador had the highest incidence of malaria amongst all Mesoamerican countries – with 95,835 cases and a 38% share of all cases in Mesoamerica. However, by 1995, the country contributed just 2%, maintaining low incidence until the present day (Figures 3.1 -3.3). The country is now in the elimination phase and reported seven malaria cases in 2015 (0.1% of cases in Mesoamerica) (Schneider *et al.*, 2016). In 2017 the country reported zero locally acquired cases for the first time (WHO, 2018a). Epidemiologists in El Salvador have kept records at a high spatial and temporal resolution throughout their malaria control and elimination efforts. In addition there has been a long history of reactive and active case detection, testing and treating all patients with fever with antimalarials and an extensive network of community malaria workers has been in place since the 1950s (Schneider *et al.*, 2016), evidence suggesting that case detection and treatment is strong. A full understanding of elimination in El Salvador could therefore provide useful insights for other countries as they aim to achieve and sustain elimination.

Using the epidemiological line-list maintained by the Ministry of Health, I applied methods described in Chapter 2 (Algorithm 1, submodular inference using a greedy algorithm) to these data to estimate how transmission varied over space and time in El Salvador between 2010 and 2016. The subsequent results illustrate the role of importation in driving transmission dynamics in this country and provide independent estimates of the likelihood that El Salvador can eliminate malaria by 2020.

3.2 Methods

3.2.1 Data

The data, obtained from the Salvadorian Ministry of Health (MINSAL), consisted of all confirmed cases of malaria between 2010 and the first two months of 2016 (N= 91 cases, of which 30 imported, 6 *P. falciparum*, 85 *P. vivax*). All but two cases had an address listed. For these cases the location was available at the municipio, or municipality level, and the coordinates of the centroid of the municipality (which for both were cities) were used as the geo-location. Two cases had addresses listed outside of El Salvador, both of which were in Guatemala. All cases within El Salvador with full addresses (N=85) were georeferenced by latitude and longitude to *caserío/ lotificación* level, which is approximately neighbourhood or hamlet level. Name searches of streets, *caseríos*, and landmarks were carried out using Nominatim on Open Street Map[1]. Google and Bing maps[2] were also used to cross check and in the absence of information available on open street map. I also used several locality listing websites[3] to obtain and cross check georeferences for *caseríos*.

Municipality (*municipio*) and district (*distrito*) were also provided, allowing cross checking for duplicate neighbourhood names and ensure continuity. In addition, searches were made online for local schools, churches, news stories and community groups to cross check locations. Many addresses listed geographic features such as landmarks or road names. Where possible, Google satellite imagery were examined for these features and/or evidence of dwellings.

Data were captured through El Salvador's national epidemiological surveillance system (VIGEPES). These include cases reported by 30 public hospitals, 746 health facilities and thousands of community health workers stationed throughout the country (approximately 3,246 in 2010)(El Salvador Ministerio

[1] https://nominatim.openstreetmap.org/
[2] https://www.bing.com/maps; https://www.google.co.uk/maps/
[3] http://www.mapmonde.org/central-america/el-salvador/; http://www.maplandia.com/el-salvador/; https://es.wikipedia.org/wiki/Categoría:Cantones,_caseríos_y_comunidades_de_El_Salvador; https://geographic.org/geographic_names/el_salvador/index.html#F;

de Salud, 2011; Schneider *et al.*, 2016). During this period, the number of blood slides tested per year remained similar (Table 3.1). The line-list featured a unique patient identifier, address, age, sex, symptom onset date, and treatment seeking date, as well as details about treatment and diagnostic testing. All confirmed cases were treated. For cases recorded in 2010, time of both symptom onset and treatment were available, providing an opportunity to estimate the delay between symptom onset and treatment for that year (Figure 3.4).

Detailed case investigation was carried out by MINSAL and cases were identified as imported or locally acquired based on travel history, as well as primary, secondary, tertiary or orphan cases without clear sources, based on relationship with and proximity to previous cases. I obtained the latitude and longitude of the address, accurate to caserío (hamlet) level, using Open Street Map (OpenStreetMap contributors, 2017). El Salvador carries out reactive case detection following presentation at health facilities. However, in 2011, of 4,500 slides examined through reactive case detection (representing 4.5% of all slides examined), just one additional case was detected. Both passive and active screening of migrants at key border crossings and in agricultural areas near borders also takes place. In these targeted areas, individuals are monitored for fever in the past 30 days, tested, and a single dose of chloro-primaquine prophylaxis is provided. In 2011, the Ministry of Health reported that 33,000 migrants were reached through active and passive case detection and an additional four cases of malaria were found (El Salvador Ministerio de Salud, 2011). Most cases were detected through passive surveillance in health facilities, at borders and by community health workers in rural areas.

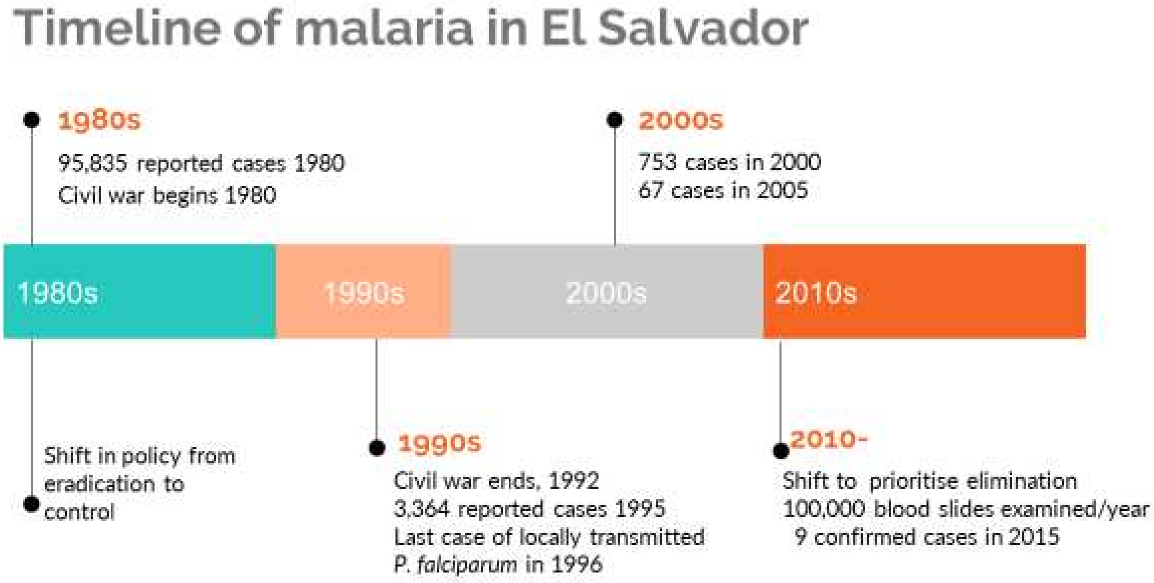

Timeline of malaria in El Salvador
1980s
95,835 reported cases 1980
Civil war begins 1980
2000s
753 cases in 2000
67 cases in 2005
1980s
1990s
2000s
2010s
Shift in policy from eradication to control
1990s
Civil war ends, 1992
3,364 reported cases 1995
Last case of locally transmitted P. falciparum in 1996
2010-
Shift to prioritise elimination
100,000 blood slides examined/year
9 confirmed cases in 2015

Figure 3.1 Timeline of malaria in El Salvador

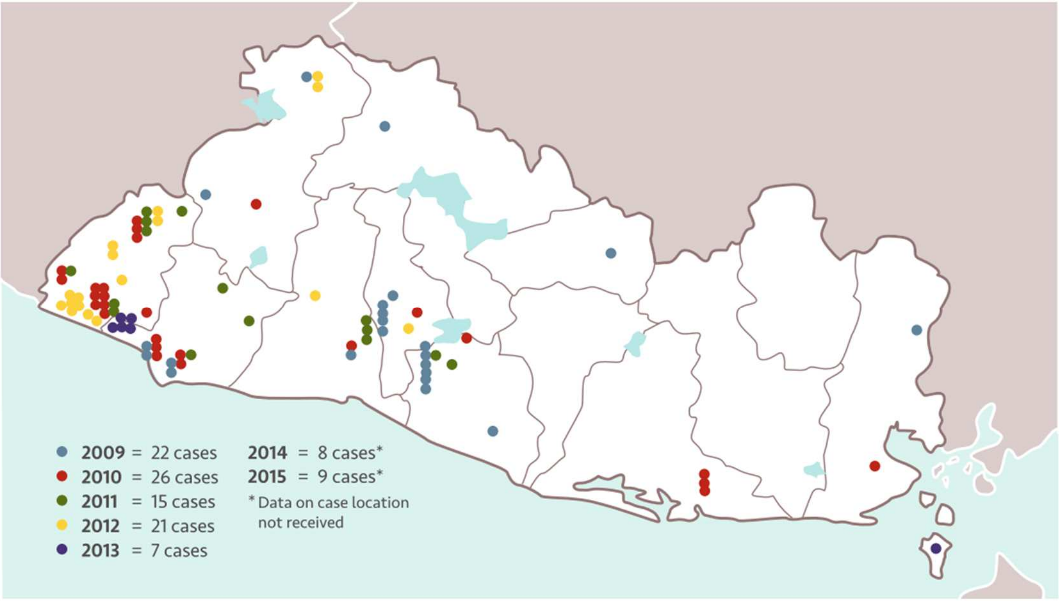


Figure 3.2 Distribution of cases of malaria in El Salvador 2009-2015*, reproduced from* (Schneider *et al.*, 2016)

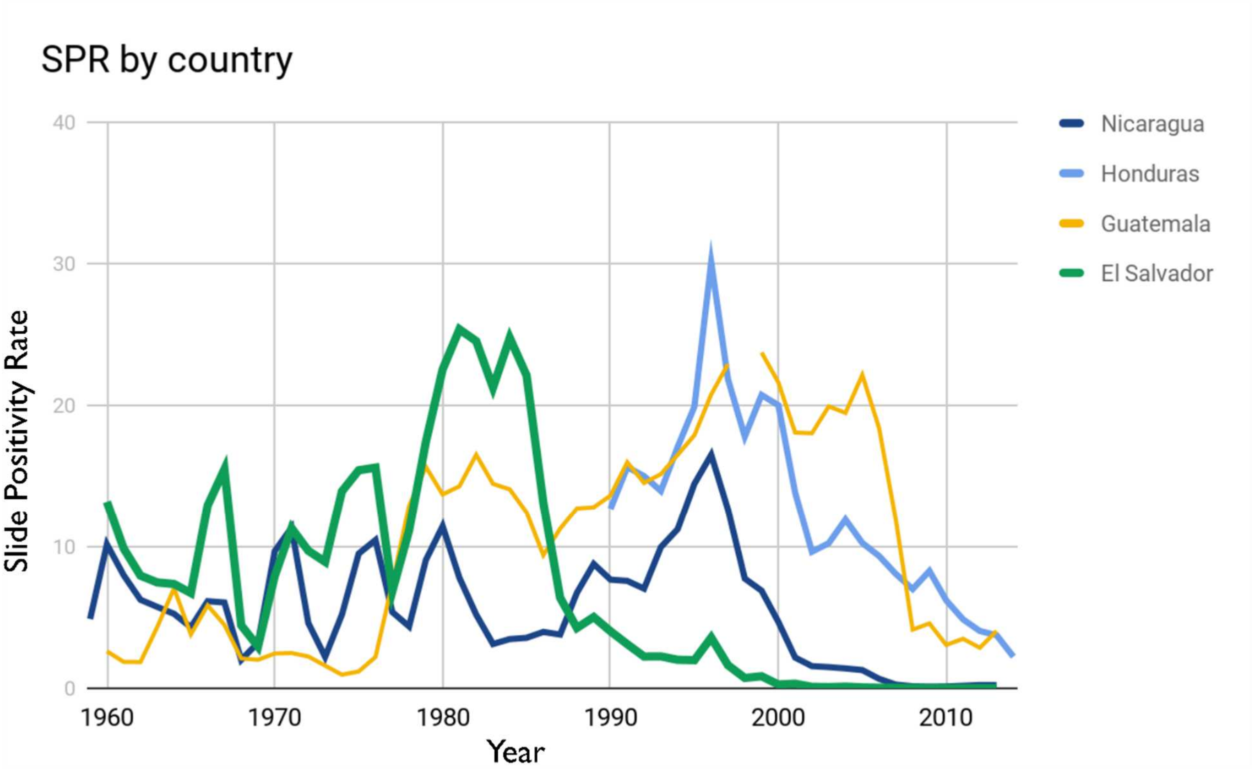


Figure 3.3 Slide Positivity Rate (SPR) by country. *Plot showing SPR over time for Central American Countries. Note El Salvador's rapid decline in Malaria, which was mirrored at later dates by other countries.*

Table 3.1 Slides examined per year (Schneider** et al.**, 2016)

Year	Slides examined
2010	115 000
2011	100 883
2012	124 885
2013	103 748

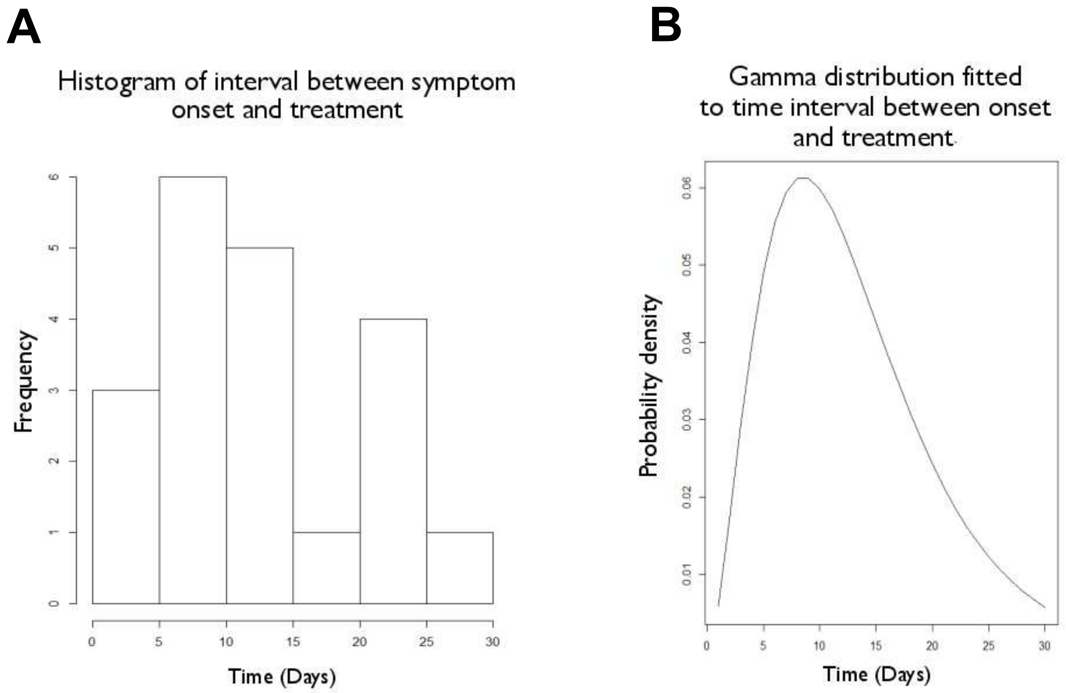


Figure 3.4: Distribution of time from symptoms to treatment, based on available data from 2010. *A) Raw data as histogram and B) gamma distribution fitted to data*

3.2.2 Serial interval distribution

The serial interval is defined as the time between a given case showing symptoms and the subsequent cases they infect showing symptoms (Fine, 2003). For a given individual j at time t_j, showing symptoms before individual i at time t_i, the serial interval distribution specifies the normalised likelihood or probability density of case i infecting case j based on the time between symptom onsets, $t_i - t_j$. The serial interval summarises a number of distributions including the distribution of a) the times between symptom onset and infectiousness onset, b) the time for humans to transmit malaria parasites to mosquito vectors, c) the period of mosquito infectiousness, and d) the human incubation period.

I defined a prior range of possible serial interval distributions for malaria. The serial interval distribution of treated, symptomatic *P. falciparum* malaria, previously characterised using empirical and model based evidence(Huber *et al.*, 2016) was adapted to inform the prior distribution for the relationship between time and likelihood of transmission between cases in El Salvador. Two cases imported from West Africa were *P. falciparum*, however the remainder of cases were *P. vivax*. As a result, the prior distribution was altered to better reflect the biology of *P. vivax* and the dominant vector species in El Salvador, *Anopheles albimanus,* but was uninformative enough to allow for possible variation in transmission dynamics, for example due to imported infections with *P. falciparum*. In addition, there is a possibility of a small number of asymptomatic or undetected and therefore untreated infections contributing to ongoing transmission, which will take on a longer serial interval. By defining a prior distribution for the serial interval distribution one can account for some of this uncertainty.

A shifted Rayleigh distribution was used to describe the serial interval of malaria, which can be varied by changing two parameters: α and γ. The parameter α governs the overall shape of the distribution, and γ is the shifting parameter accounting for the incubation period between receiving an infectious bite and the onset of symptoms (Figure 3.7A). The γ shifting parameter was defined as ranging between 10 and 15 days to account for the minimum extrinsic incubation period within the mosquito and the minimum time between infection and suitable numbers of gametocytes in the blood to lead to symptom onset (Warrell and Gilles, 2002). The prior for the α parameter determining the shape of the distribution was given a Uniform distribution and bounded, giving an expected time between symptom onset of one case and symptom onset of the case it infects of 29 days (95%CI = 16 – 300 days, sd = +/- 7 days), with the lower bound having an expected serial interval of 25 days (95%CI =16 – 299 days, sd = +/- 4 days) and the upper bound 47 days (95%CI = 16-300 days sd= +/- 18 days). By comparison the expected values for treated *P. falciparum* from existing literature range between 33 (Churcher *et al.*, 2014) and 49.1 days (95%CI = 33- 69)(Huber *et al.*, 2016).

3.2.3 Determining the transmission likelihood

I assume cases were classified correctly from case investigation as imported or locally acquired based on recent travel history. Following this assumption, locally acquired cases could have both infected others and been infected themselves. However imported cases could only infect others, as it is assumed that their infection was acquired outside of the country. There were no confirmed relapse cases in the dataset, and all cases were treated with primaquine and chloroquine (radical cure) after being detected. Treatment is initiated before cases are confirmed by microscopy (Ministerio de Salud El Salvador (MINSAL), 2015). Active case detection is initiated locally following a confirmed case and in active foci in which local surveillance is believed to be weak. In these scenarios blood slides are examined within 24 hours of being taken (Ministerio de Salud El Salvador (MINSAL), 2015). Given this, my approach assumes that an individual can only be infected once by a case that has shown symptoms earlier in time.

The data input consisted of a time series of symptom (fever) onset $t \in \{t_1, \dots, t_n\}$, time ordered such that $t_1 < t_2, \dots, < t_n$. While the times of symptom onset are known, the data do not indicate who infected whom and the underlying transmission chain, $\mathcal{T}$. As described in Chapter 2, the goal of the model is to infer the most probable network structure, $\mathcal{G}$, connecting these n infections. One can view cases as nodes in a network $\mathcal{G}$, and possible transmission events as the edges linking nodes. $\mathcal{G}$ is inferred solely from the symptom onset times t, a serial interval distribution, and prior probability distributions for the serial interval distribution parameters.

$\mathcal{G}$ contains all possible spanning transmission chains over which an infection could spread given the observed times. $\mathcal{G}$ therefore, includes the most likely transmission tree, but also includes, other possible trees supported by the data. One therefore can view a transmission tree $\mathcal{T}$ as a realisation of a stochastic diffusion process generated over an underlying network $\mathcal{G}$. Crucially, $\mathcal{G}$, accounts for competing edges and is sparse (only includes plausible edges).

For a given transmission tree $\mathcal{T}$ describing infection events linking cases and assuming the independent cascade model (Kempe, Kleinberg and Tardos, 2003), the (upper triangular) likelihood of observing the times of symptom onset is simply the product of all permissible pairwise transmission likelihoods in the tree(Rodriguez and Schölkopf, 2012). This description until this point is the same as that introduced by Wallinga and Teunis (Wallinga and Teunis, 2004) and extended to a wide variety of contexts by others (Ypma *et al.*; Walker *et al.*, 2010; Morelli *et al.*, 2012; Jombart *et al.*, 2014; Reiner *et al.*, 2015; Salje *et al.*, 2016). However, in contrast to previous methods based on Wallinga and Teunis this approach maximises the likelihood $f(t|G)$ conditional on an underlying $\mathcal{G}$, a problem that is NP-hard (Khuller, Moss and Naor, 1999). Previous approaches have either allowed all possible connections in $\mathcal{G}$ (Wallinga and Teunis, 2004), sampled from the likelihood (Ferguson, Donnelly and Anderson, 2001) or explored a limited number of pathways (Salje, Cummings and Lessler, 2016). Here, by following the approach introduced by Gomez-Rodriguez and Schölkopf (Rodriguez and Schölkopf, 2012), I find the most likely underlying transmission network given the timing of symptom onset for a set of k transmission events linking cases. The computational hardness of maximising $f(t|G)$ meant that an optimal solution could only be found by exploring every possible transmission tree on G. However, due to the submodularity of the independent cascade model (Kempe, Kleinberg and Tardos, 2003) a near optimal solution could be found using a greedy algorithm. Briefly, the greedy algorithm used starts with an empty graph, and then add edges sequentially such that the *marginal gain* in the likelihood of the transmission tree for each iteration is maximised. The marginal gain measures of importance for each edge of the network through the gain that each edge provides to the total solution over competing edges, and therefore applies shrinkage to the raw pairwise likelihood with the likelihood of competing edges. This process stops when have k edges are reached. Stopping at k edges ensures that the resulting network is sparse which not only ensures a parsimony but removes unnecessary edges that could influence R_c calculations. An appropriate value of k is defined by adding edges until the marginal gain in likelihood of adding additional edges is below a given threshold (0.0005). Sensitivity analysis revealed that these results are robust to changes in this threshold between 0.001 and 1e-10 (Appendix 1).

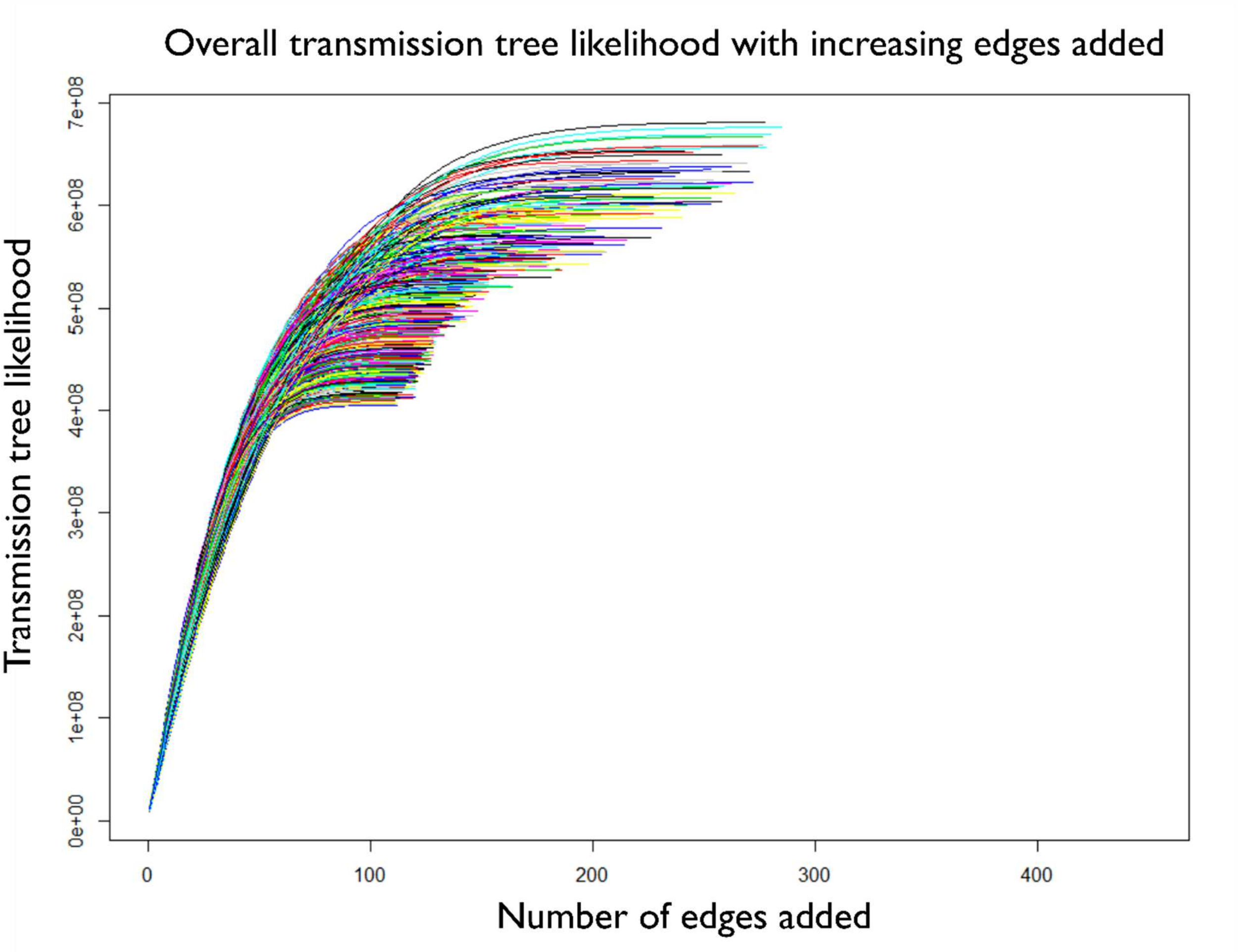


Figure 3.5: Plot showing the marginal gain in likelihood by adding edges to network using greedy algorithm. *Each coloured line represents different draws of alpha, and shows the marginal gain in likelihood of adding edges to the network. The cut off-point for marginal gain in likelihood used here is 0.0005.*

3.2.4 Estimating R_c

Individual reproduction numbers for each case were established by creating a matrix where each column represents a potential infector and the rows represent a potential infectee, describing which infector edges are connected to infectees and the normalised marginal gain of that edge. Intuitively then, by taking the row sums of $\mathcal{R}$ we get the (fractional) number of secondary infections and therefore a point estimate of the time varying reproduction number $R_c(t_j)$. This reflects for an individual, how many people they are likely to have gone onto infect. When multiple individuals have been infected at a given time and/or place, one can take the mean individual R_c and uncertainty in this value as an indicator of reproduction numbers for a given time and/or location.

In contrast of traditional methods based on Wallinga and Teunis (Wallinga and Teunis, 2004) using the marginal gain in this way encapsulates not only the pairwise likelihood of transmission between two cases, but conditions this likelihood on the impact of competing edges in the inferred network. Given the provable near optimal solution of the greedy algorithm and the use of marginal gains in calculating $\mathcal{R}$, my estimates of $\mathcal{R}$ provide more rigorous estimates of reproduction numbers than just using standard Wallinga and Teunis (Wallinga and Teunis, 2004) approaches, which do not consider the overall transmission tree in optimisation and do not account for missing cases.

I assume cases were classified correctly from case investigation as imported or locally acquired based on recent travel history. Following this assumption, locally acquired cases could have both infected others and been infected themselves. Imported cases could only infect others, as I assume their infection was acquired outside of the country. I also assume a case showing symptoms at time *t* has been infected by a case which began showing symptoms earlier in time, due to the short time between symptom onset, presentation at health facilities and the beginning of treatment.

3.2.5 Accounting for missing cases

Assuming all cases reaching community health workers or health facilities are recorded, missing cases may be generated by two processes. Symptomatic cases may be missed by not seeking care or being found through active case detection. On the other hand, cases may be asymptomatic and therefore unlikely to seek care or be detected. They may have densities of parasites in their blood which are too low to be detectable by microscopy if active case detection occurs. These reasons for missed detection apply to both imported cases and locally acquired cases. We assume the pool of asymptomatic cases in the country is low and has a small contribution to ongoing transmission. To explore the amount of cases which may be going undetected within the independent cascade framework, we consider additional edges π, that represent unobserved individuals who can infect any observed individual, i, in a transmission chain. Every observed individual i can get infected by unobserved individuals, π, with an arbitrarily small probability ε. This so called ϵ-edge is connected to every node in the network and do not, by design, participate in the diffusion propagation. The ϵ-edges prevents breaks in the network diffusion cascade where the likelihood of transmission between observed cases is sufficiently low, the case is linked to an external source. Additionally, ϵ-edges ensure the likelihood is monotonic, that is, converting an ϵ-edge to a network edge in $\mathcal{G}$ only increases the likelihood. The addition of ϵ-edges was achieved by augmenting the pairwise transmission likelihood as follows:

$$f(t_i|t_j; \alpha, \gamma) = \epsilon^{-1}\alpha(t_i - t_j - \gamma)e^{-\alpha(t_i - t_j - \gamma)}$$

The specific value of ε was set at 1e-5 to balance between false positives and false negatives when linking cases by infection events. The higher the value of ε, the larger the number of nodes that are assumed to be infected by an external source.

3.2.6 Covariate assembly

The environmental covariates (i.e., independent variables) used in the spatial mapping of R_c >1 risk consisted of raster layers that spanned El Salvador 2.5 arc-minute (~5 km x 5 km) spatial resolution. Covariate choice was based on key variables used within past malaria mapping endeavours (Bhatt *et al*., 2015). Raster datasets were then acquired or produced, and wherever possible dynamic versions (i.e., temporally varying products) were utilized to support the temporal aspect of the analysis. The majority of the raster covariates were derived from high temporal resolution satellite images and then aggregated to create dynamic covariates for every month throughout the study period (2010-2016). The covariates used are listed below in **Table 3.2**.

Table 3.2: Covariates used in risk mapping R_c >1

Variable Class	Variable(s)	Source	Type
temperature	land surface temperature (day, night and diurnal flux)	MODIS product	dynamic monthly
precipitation	mean annual precipitation	WorldClim	synoptic
elevation	digital elevation model	SRTM	static
infrastructural development	accessibility to urban centres and night-time lights	modelled product and VIIRS	static
moisture metrics	aridity and potential evapotranspiration	modelled products	synoptic

3.2.7 Spatial methodology

The underlying spatial statistical model was fitted to binomial data of $R_c > 1 = 1; R_c < 1 = 0$, using the logit link function:

$$R^{+}_{>1,i} \sim Binomial(p_i, N_i)$$

$$log(p_i/(1-p_i)) \sim GP(\mu, Q)$$

$$\mu = \alpha + X_i\beta$$

$$Q = \boldsymbol{K}_{space}^{-1}$$

$$\boldsymbol{K}_{space}^{-1} = \ solve\ (k^2 - \Delta)^{\frac{\alpha}{2}}\big(\tau x(s)\big) = W(s)$$

where $R_{>1,i}$ are the number binary data points for $R_c > 1 = 1; R_c < 1 = 0,\ p_i$ is the estimated $R_{>1}$, expressed as a logit transformed probability and modelled as a Gaussian process with μ and precision Q. The GP mean μ is a linear function of a global intercept α and space-time indexed covariate values $\boldsymbol{X_i}$. Q is a sparse precision matrix and $\boldsymbol{K}_{space} = Q^{-1}$ is the covariance matrix. Q is the sparse finite element solution to the stochastic partial differential equation $(k^2 - \Delta)^{\frac{\alpha}{2}}\big(\tau x(s)\big) = W(s)$, where Δ is the Laplacian, k is the spatial scale/range parameter, τ controls the variance, α is the spatial smoothness parameter (fixed at $\alpha = 2$), and $W(s)$ is the spatial white noise process. To account for the curvature of the earth the distance metric *s* is defined on a spherical manifold in Cartesian $\mathbb{R}^3$.

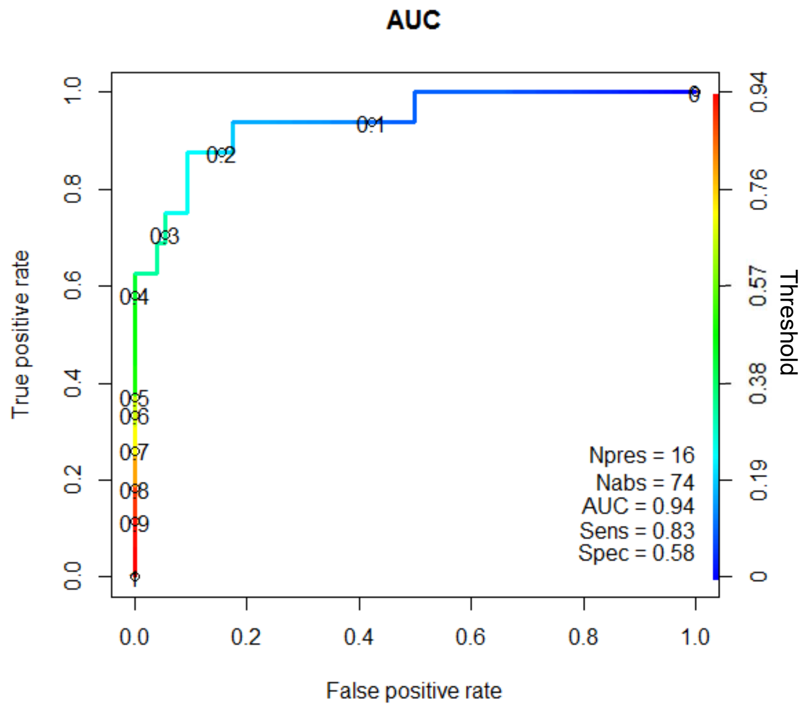


Figure 3.6: Area Under the ROC Curve (AUC) from cross validation of geostatistical model *used to create riskmaps of P(Rc>1) AUC = 0.94, Sensitivity = 0.83, Specificity=0.58. The colours and labels (illustrated in the scale bar on the right side of the x axis) represent the threshold for classification as 1 (Rc>0) or 0 (Rc=0). When the threshold is decreased, more positive values are returned, thus sensitivity (the true positive rate) increases and specificity (1- false positive rate) decreases.*

3.2.8 Estimating timelines towards elimination

To explore trends in R_c over time, we fitted a generalised additive (GAM) model to the estimated $R_c(t)$ values and extended this line beyond the period of observation to 2030. We then also fitted Gamma, Power law and Exponential distributions to the estimated $R_c(t)$ values, and found they were best represented by Gamma distribution according to AIC scores (Akaike, 1974). To explore the likelihood of elimination by a given time point, we randomly drew 10,000 R_c values from Gamma distributions with increasingly small mean reproduction numbers, keeping the fitted shape parameter constant. We then found the threshold mean R_c below which the probability of an individual $R_c$exceeding one is less than 5%. By extending the current fitted trendline for $R_c$values to 2030, we identified the expected timepoint for R_c to reach this threshold value, given the observed decline in R_c observed over the study period.

3.2.9 Mapping R_c

To map estimates of transmission risk, individual reproduction numbers were divided into those above and below one. The latitude and longitude of the reproduction numbers were included in a geospatial hurdle model implemented in rINLA (Rue, Martino and Chopin, 2009) where demographic and environmental covariates were used to estimate the likelihood of a case having a reproduction number above 1 if imported into the area in 2010. This is a measure of malaria "receptivity" or underlying transmission potential rather than overall malaria risk, as importation likelihood is not quantified in this analysis. Area under the ROC curve scores from leave one out cross validation were used to assess model fit (Figure 3.6).

3.3 Results

Between 2010 and the first two months of 2016, a total of 91 cases of malaria were confirmed by microscopy in El Salvador, of which 30 were classified as imported. There was a total of six cases of *P. falciparum*, all of which were imported. The resulting estimated transmission network is shown in Figure 3.7. Overall, the temporal dimension dominates the identification of infector-infectee pairs (Figure 3.7B), informed by the prior distribution for the serial interval (Figure 3.7B). We identified two locally acquired cases which could not be plausibly linked to other cases within the dataset (Figure 3.7C). These were estimated in periods in which a clear gap in the data was apparent, and may therefore represent unidentified importations, relapse cases or unreported locally acquired sources of infection.

We estimated the mean individual reproduction number over 2010-2016 to be 0.61 (95% CI = 0.56,0.65). This is consistent with the ratio of locally acquired to total cases (61:91 = 0.66), which has been proposed elsewhere as an approximate estimate of R_c (Cohen *et al.*, 2010). When fitting a generalized additive model to the data, the overall trend was a decline from a fitted $\mathcal{R}_c$ of 0.73 at the start of the observation to 0.47 by the end of the period (Figure 3.8). Individual reproduction numbers showed seasonal fluctuations through time, with regular peaks observed in December, which coincides with the end of harvest season for many crops in El Salvador and Guatemala, and August, which coincides with a period of national holiday and the end of the rainy season.

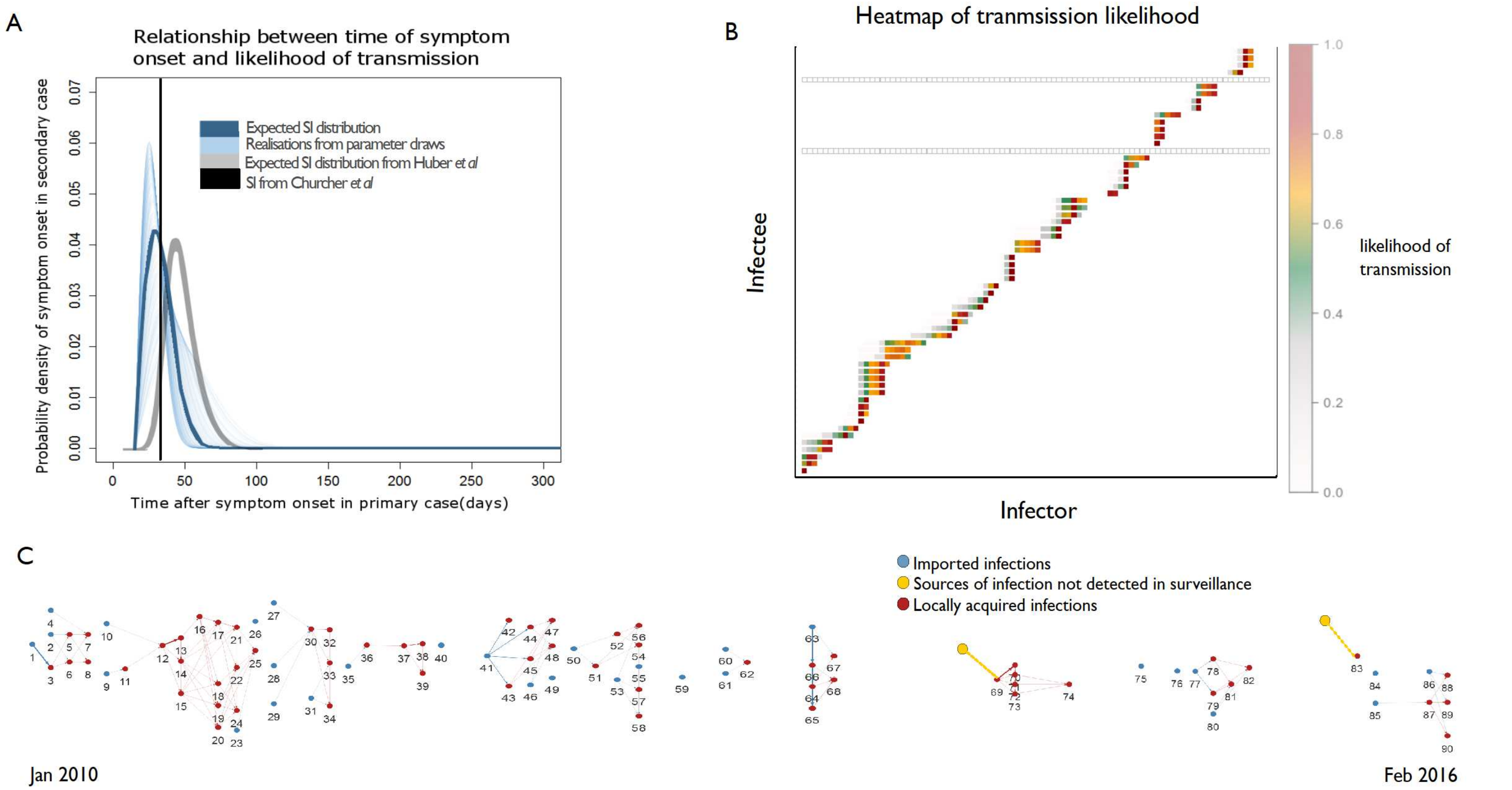


Figure 3.7 ***Plots showing stages of transmission network estimation.*** *A) Serial interval (SI) distribution used in the analysis. Thin blue lines represent 300 realisations of the SI distribution resulting from draws from the distributions of the parameters determining the shape of the SI and incubation period. The thicker blue line represents the expected SI distribution. For comparison, the grey line represents the SI distribution estimated for symptomatic, treated P. falciparum infection from (Huber et al., 2016) and the black line shows the expected SI for P. falciparum from (Churcher et al., 2014). B) Heatmap showing likelihood of transmission occurring between infector and infectee pairs. The x axis represents all possible infectors (all reported cases) of the observed cases, organised by symptom onset date. The Y axis represents all possible infectees (all locally acquired cases, as by definition it is assumed imported cases were infected outside of the country). Each square represents a potential infector/ infectee pair. The colours of the heatmap represent the normalised likelihood of infector j having been the infector of infectee I, where red is 1 and grey is 0. Grey squares show where cases were not likely to be infected by to any observed case, and therefore presumably infected by an individual who was not detected by surveillance. These could be asymptomatic or unreported clinical cases C) Reconstructed network, where numbers represent the ID of cases in temporal order. The strength of likelihood of connection represented by weight of edges linking cases. The two locally acquired cases identified to be infected by unobserved sources of infection are highlighted.*

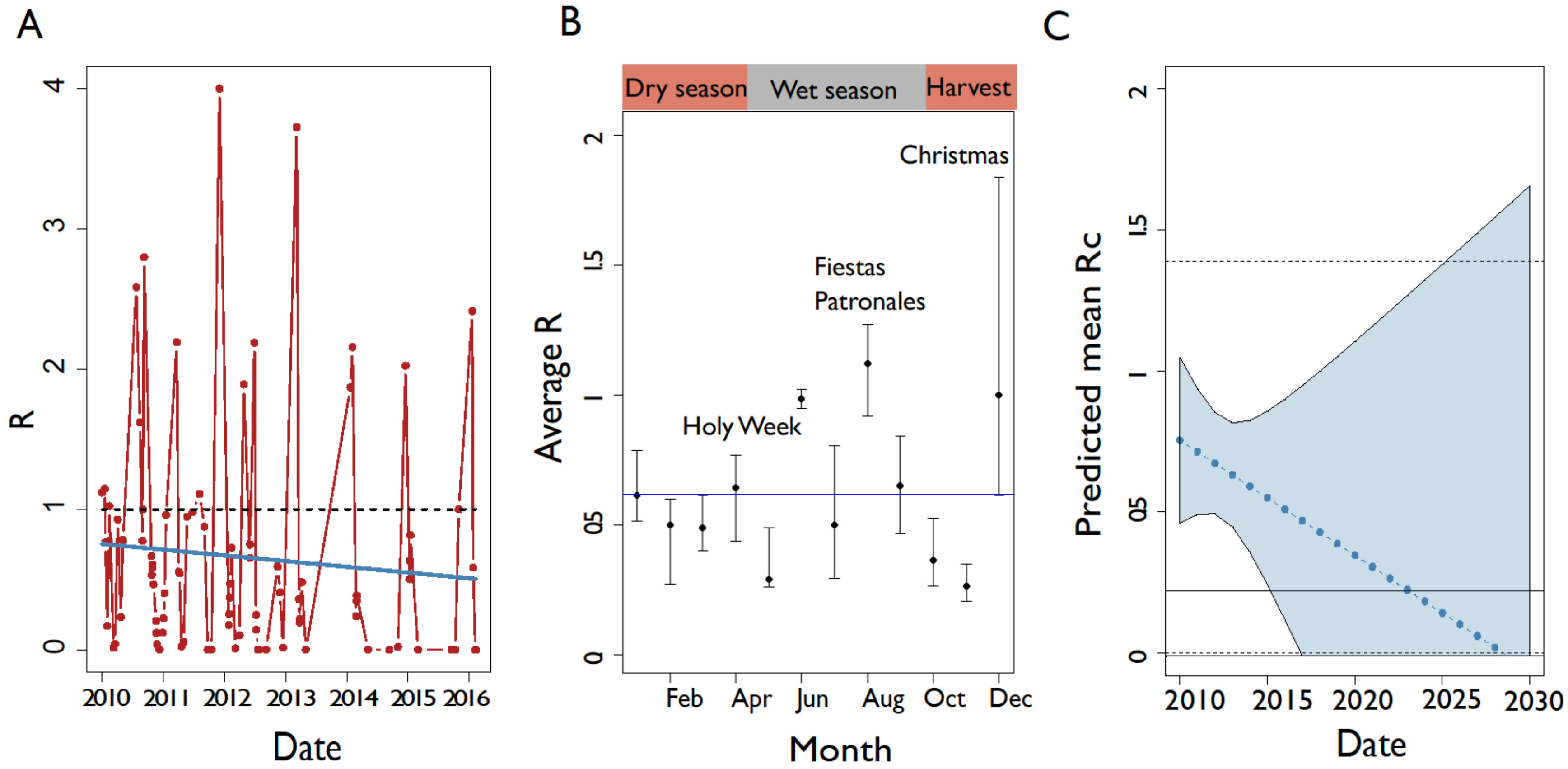


Figure 3.8 Temporal trends and forecasting using reproduction number estimates *A) Individual reproduction numbers plotted over time. Dashed line shows $\mathcal{R}_c = 1$, blue line shows fitted Generalised Additive Model. B) Posterior estimates of $\mathcal{R}_c$ by month of year. Bars show 95% credible interval. Blue line shows the mean estimated $\mathcal{R}_c$ for the observation period. Key holidays, seasons and agricultural patterns are labelled. C) Extended trendline to 2030 showing predicted $\mathcal{R}_c$. Shaded area shows 95% credible interval from prediction and solid line shows mean threshold of P =0.05 of cases occurring with an $\mathcal{R}_c$ above one. Dashed lines show 97.5 and 2.5th quartiles for this threshold from 10000 simulations.*

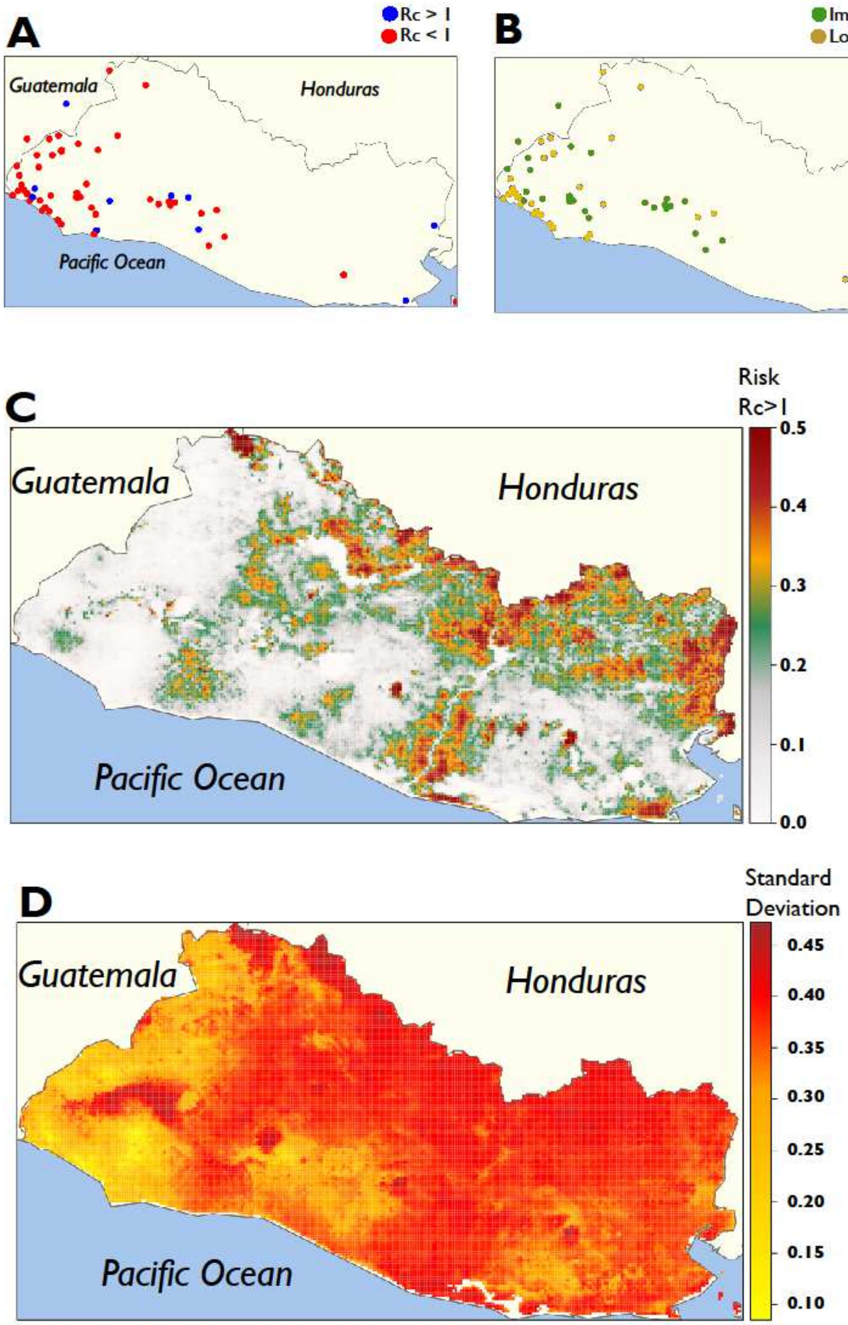


Figure 3.9 Maps of risk of R_c exceeding one. *A) Distribution of R_c values by location of residential address. Red points represent an R_c value below one, blue points represent an R_c value above 1. B) Distribution of imported and locally acquired cases by location of residential address. Yellow points represent locally acquired cases; green points represent imported cases. C) Map of risk of R_c exceeding 1 if a case were to occur in an area. Note this estimate does not consider risk of importation but estimates receptivity to transmission if importation were to occur. D) Standard deviation in risk estimates from C.*

3.3.1 Spatial distribution of cases and R_c

Data were highly focal, with 70% of cases originating from two adjacent administrative departments neighbouring Guatemala, and 32% of cases originating from just two municipalities within these regions (Jujutla and Acajutla) (Figure 3.9A-B). This pattern was also reflected in the spatial distribution of $\mathcal{R}_c$. While most areas of the country are predicted to have a low risk of R_c reaching above one over the time observed, several regions have a much higher predicted risk of R_c >1 (Figure 3.9C). In these regions, the majority of cases imported into the region could be expected to result in at least one onward transmission event. However, it is important to note that uncertainty in these predictions is high in areas where cases have not been seen. The area with the least uncertainty in the estimate, around the borders of Guatemala, suggest that the majority of cases occurring there did not contribute to onward transmission.

3.3.2 Impact of imported cases on transmission

The mean marginal gain to the likelihood of including infections from imported cases into the constructed transmission networks was much higher than including locally acquired cases (0.081 compared to $3.44\ e^{-7}$), suggesting that imported cases are a major driver of transmission. Visual inspection of the most likely chains of transmission (Figure 3.7) also are suggestive of this, where the index case in a cluster of linked cases was almost always an imported case.

3.3.3 Endgame predictions based on $\mathcal{R}_c$ and stochasticity

To investigate potential timelines to elimination (i.e. the absence of local transmission) I characterised heterogeneity in the reproduction number using a Gamma distribution which, when fitted to the data, suggests a threshold mean $R_c$of 0.22, below which there would a less than 5% chance of any individual reproduction number exceeding one. Using the fitted trend in the mean $R_c,$ one would expect this level to be reached by 2023, assuming no change in the rate of importation (Figure 3.8C).

3.4 Discussion

Understanding how transmission varies over time and space is critical to efforts to achieve and maintain elimination of infectious diseases such as malaria. Reconstructing transmission chains and estimating individual reproduction numbers has been used widely within epidemiological analysis (Ghani *et al.*, 2009; Walker *et al.*, 2012; Jombart *et al.*, 2014), but rarely used to study vector-borne or endemic diseases, albeit with a few notable exceptions (Reiner *et al.*, 2015; Salje *et al.*, 2016). Separately, similar problems have been approached within human social network analysis, through a family of approaches known as independent cascade models (Kempe, Kleinberg and Tardos, 2003; Gomez-Rodriguez, Leskovec and Krause, 2010; Rodriguez, Balduzzi and Schölkopf, 2011; Rodriguez and Schölkopf, 2012). Here I have adapted these methods to routine data from an eliminating Central American context, El Salvador, in order to inform progress towards their malaria elimination goals.

My results suggest that the time-averaged R_c has been below 1 in El Salvador since 2010, suggesting that sustained endemic transmission at the country level has already been interrupted. However, I estimated individual reproduction numbers exceeding one, resulting in ongoing outbreaks of transmission. Assuming the downward trend observed in R_c between 2010 and 2016 continues, one would expect the probability of such outbreaks to be less than 5% by 2023 if current levels of malaria importation and control continue. However, because imported cases were found to have higher reproduction numbers and their inclusion in the transmission tree increased the overall likelihood of the tree much more than locally acquired cases, it is important to note that the rate of importation is likely to affect the distribution of R_c. With increased importation this timeline to elimination could lengthen. Conversely, if importation was reduced, the timeline would be shortened. Thus, the levels of malaria importation from neighbouring countries would likely need to be decreased in order to achieve elimination by 2020, following current WHO certification policy of three years of zero locally acquired cases.

Given the extensive surveillance of migrants already carried out by El Salvador, as well as the free-movement and trade agreements which exist between El Salvador, Guatemala, Honduras and Nicaragua, the most efficient way of achieving this is likely to be through reducing the prevalence of malaria throughout Central America. However, given the seasonal peaks in R_c estimated to occur in August and December, which coincide with national holidays and the end of harvest season, there could additionally be an opportunity to increase surveillance activities and interventions during these key periods of high human mobility.

The Elimination of Malaria in Mesoamerica and Hispaniola (EMMIE) initiative aims to eliminate local malaria transmission from the entire Mesoamerican region by 2020 (Herrera *et al.*, 2015). My results support the need for a regional approach to elimination. The clear impact of importation in driving residual transmission in El Salvador highlights the need for cross-border collaboration. In order to drive transmission down, areas of the highest "receptivity" to intervention and "vulnerability" to importation of cases must be identified. Approaches such as this, which map transmission risk, could be combined with information about human movement to identify foci for increased surveillance, vector control and other interventions. This approach using El Salvador as a case study could be adapted and used in other Central American countries or other contexts aiming for elimination.

The analysis identified two cases with no clear source. When raising the threshold likelihood for linking observed cases as part of the sensitivity analysis and reducing the number of possible edges in the network, I find 7 missing cases. There is evidence in some low transmission contexts, especially where rapid declines of malaria have been seen recently, of significant asymptomatic and/or sub microscopic reservoirs of infection which may transmit to onwards transmission (Okell *et al.*, 2012). These could be sources of the missing infections identified in this study. However, El Salvador is unlikely to have a large amount of asymptomatic cases due to a long history of low numbers of cases. If the missing source of infections was mainly indigenous asymptomatic infections, it would signify that there is an asymptomatic reservoir contributing to onward transmission and that must be controlled to reach elimination. This could be achieved through PCR-based screening and treatment or increased vector control in focal areas. An alternative explanation is that there may be a small number of unreported symptomatic cases or relapse cases which were not reported or detected, which could be indigenous or imported. If due to importation this would further support the need for strong regional cooperation via initiatives such as EMMIE to reduce burden in neighbouring countries, and to maintain vigilance over extended periods in a very low transmission stage.

There are several limitations to this work. Firstly, whilst this approach uses epsilon edges to identify potential external sources of infection, this approach is only appropriate for smaller numbers of missing cases. Given the long history of small numbers of cases and testing and treating ~100,000 febrile patients per year (of which only 6 were positive for malaria in 2015), and the programme of active case detection, as well as cross-sectional surveys of school age children in historic foci finding 0% prevalence by PCR (Sorto *et al.*, 2015), this is a reasonable assumption. However, in other contexts, this may be a larger concern and methods such as reversible jump MCMC methods (Green, 1995; Chis Ster, Singh and Ferguson, 2009) for data augmentation and inference may be appropriate.

Secondly, by the nature of a near elimination context, the sample size is very small. The methods used for estimating $\mathcal{R}_c$ are well suited to small, well observed infection cascades, however this small sample size does provide a limitation for mapping, meaning the resulting maps have relatively high levels of uncertainty outside of the areas of El Salvador where cases are seen principally around the pacific coast, Guatemalan border and in San Salvador. There is scope to incorporate expert knowledge to refine the map in areas where data are lacking. It is important to reiterate the uncertainty in risk map estimates for most of the country, where the standard deviation in risk estimates neared one in many areas where no cases were observed. If this uncertainty is not clearly communicated to decision-makers this may lead to misleading conclusions, or reduce trust in the estimates in areas where there is high uncertainty.

Finally, there is a large amount of uncertainty and simplifying assumptions inherent in the forward projections illustrated in figure 3.8C. Here a logistic regression was extended, assuming the observed decline between 2010 and 2016 would continue – i.e. there would be no change to importation, interventions or environmental and social factors which may shape the decline, or other epidemiological processes which could come into play as zero cases are reached. This is highly simplified, and whilst the uncertainty associated with this estimate is illustrated in the figure, and the figure was designed as a tool to show the feasibility of elimination at or around the 2020 target, but also illustrating the large amount of uncertainty around this, and the potential for much higher reproduction numbers, highlighting the need for sustained control and surveillance efforts.

It is important to consider whether methods presented here can be used in low resource settings that are earlier in the elimination process. In these contexts, the number of cases is likely to be higher and there may be less complete reporting data and potentially a higher reservoir of asymptomatic infection. In order to address these challenges several adaptations to the methods presented here may be required. Firstly there may be a need to incorporate more sources of information, e.g. demographic, spatial and possibly genetic data (Wang, Ermon and Hopcroft, 2012; Jombart *et al.*, 2014). Secondly, Bayesian data augmentation techniques (Walker *et al.*, 2010)may be required to explore the implications of large amounts of missing infection and potential reporting biases. In the case of more asymptomatic or untreated malaria there may be more uncertainty in the serial interval of malaria, however using my current approach can propagate this uncertainty through the model. Generalisations to full likelihood based or Bayesian hierarchical inference (Gomez Rodriguez *et al.*, 2014) can be beneficial by providing flexibility through parametric forms by allowing for the incorporation of additional factors (e.g. genetic distance) specific to the disease and context.

This work provides a novel framework for making use of routine surveillance data and allows quantification of malaria transmission and its variation over space and time in contexts where traditional methods such as parasite prevalence are unsuitable. This is key in designing optimal strategies to accelerate, achieve and maintain elimination. To apply to other contexts several adaptations and extensions may be required. Firstly, in this dataset there were no confirmed relapse cases, however in many contexts we may see *P. vivax* relapse, in which case the algorithm could be adapted to allow for a likelihood for "reinfection" or a looped network edge. Secondly, in settings where transmission links are less clearly identifiable or different data sources are available, this approach can be adapted to incorporate additional features such as spatial or genetic distance weightings into the likelihood function (Wang, Ermon and Hopcroft, 2012), following on from work based on Wallinga and Teunis approaches (Walker *et al.*, 2010; Morelli *et al.*, 2012; Jombart *et al.*, 2014). Finally, asymptomatic reservoirs and causes of missing cases as well as their impact on transmission dynamics could be explored in more detail to consider surveillance system design and evaluation of its strength.

In conclusion, this work adapts concepts from network theory to build and apply novel methods to map transmission over space and time in a very low transmission setting, using only routine malaria surveillance data. Such approaches offer opportunities to better understand transmission dynamics and their heterogeneities in near elimination settings to better target interventions for elimination. I estimated timescales for reaching elimination and clarified the effect of importation on the speed and feasibility of achieving and maintaining zero cases. In the context of El Salvador, these results highlight the impact of importation on sustained transmission and highlights the need for cross-border collaboration. This approach could be useful in a wide range of contexts where good quality routine surveillance data are collected, such as outbreaks and endemic diseases nearing elimination.

4

Estimating spatiotemporally varying malaria reproduction numbers in Yunnan province, China

4.1 Introduction

In 2017 China reported no indigenous malaria cases for the first time since malaria became a notifiable disease in 1956 (Feng *et al.*, 2018; WHO, 2018a). The country has experienced a major decline in the burden of malaria, from an annual incidence of 24 million cases (2961 cases per 100,000) in 1970 (Zhou, 1981). This reduction has been attributed to a combination of socioeconomic improvements and the scale-up of interventions to control malaria (Yin *et al.*, 2014). In 2010, China set out an ambitious plan for the national elimination of malaria by 2020 (the National Malaria Elimination Programme, NMEP). Elements of the plan included improved surveillance, timely response, more effective and sensitive risk assessment tools and improved diagnostics (Feng *et al.*, 2014). A key policy change implemented in 2010 as part of the NMEP was the introduction of the 1-3-7 system: aiming for case reporting in one day, which is then investigated within three days, with a focused investigation and action taken in under seven days (Cao *et al.*, 2014).

Although China is making rapid progress towards this goal, 2,675 imported cases were reported in 2017, highlighting the risk of re-introduction (Feng *et al.*, 2018). Large numbers of people move between China and malaria endemic countries, both from sub-Saharan Africa and from South East Asia (Zhou *et al.*, 2016; Lai *et al.*, 2019), driven by tourism and Chinese overseas investment (Lai *et al.*, 2016). Concerns remain about re-emergence of malaria, which has occurred several times in the early 2000s as a result of importation and favourable climatic conditions for competent vectors (Lu *et al.*, 2014). Therefore, in order to achieve three consecutive years of zero indigenous cases (the requirement for WHO certification of elimination), a sustained and targeted investment in surveillance together with efficient treatment is necessary.

Yunnan province has recorded malaria outbreaks and remains an identified foci of residual transmission as other areas in the country have reached elimination (Xia *et al.*, 2014; Feng *et al.*, 2015; Hu *et al.*, 2016; Lai *et al.*, 2017; Shi *et al.*, 2017). The province shares borders with Myanmar, Vietnam and Laos and has a strong agricultural focus. Previous studies suggest that seasonal agricultural workers and farmers are at highest risk of contracting malaria in Yunnan, with rice yield and the proportion of rural employees being spatial factors positively associated with malaria incidence (Yang *et al.*, 2017). The border region of Myanmar and Yunnan is generally ecologically suitable for malaria transmission, has a large mobile population, with few natural geographic borders separating the two countries, as well as being a site of socio-political conflict and instability (Zhang *et al.*, 2016)**.** In this context, it can be unclear whether there is any sustained local transmission or if all the observed cases are the result of short, stuttering transmission chains following importation into suitable areas. As the area of highest concern for re-emergence in China and the last to reach zero cases, I therefore sought to characterise the transmission dynamics of both *P. vivax* and *P. falciparum* in the region as China approaches elimination certification.

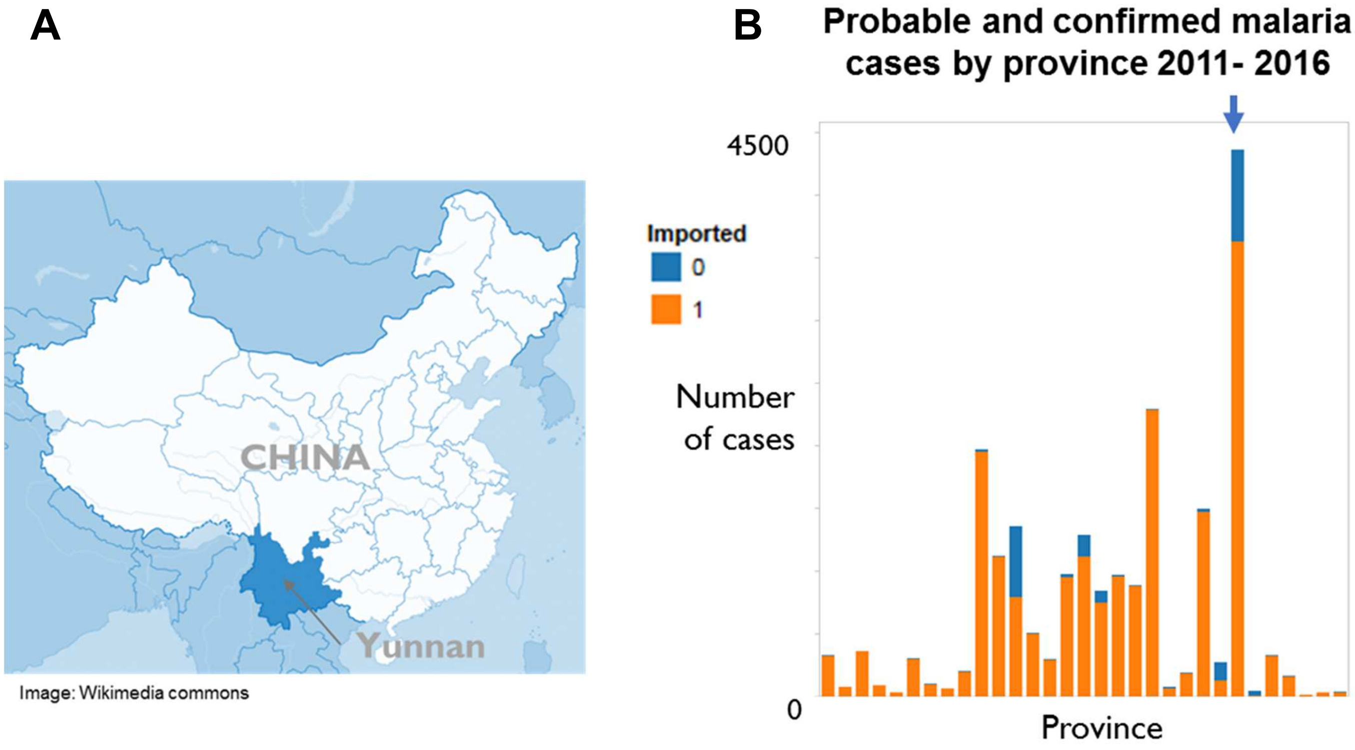


Image: Wikimedia commons

Figure 4.1 : Characteristics of Yunnan province, China. *A) Map showing location of Yunnan Province. B) Case counts of confirmed and probable P. vivax malaria 2011-2016, blue arrow shows Yunnan province, demonstrating both highest number of cases but also significant proportion of local cases, unlike most other provinces, (with exception of Hainan province).*

As illustrated in previous chapters, methods from outbreak analysis and network research have recently been developed and applied to quantify the transmission of malaria and other infectious diseases in very low transmission and epidemic settings (Reiner *et al.*, 2015; Routledge *et al.*, 2018; Wesolowski *et al.*, 2018). In China, as in other eliminating contexts, traditional metrics of malaria such as parasite prevalence are not appropriate due to small numbers and extremely sparse and spatiotemporally heterogeneous distributions of infections. However due to the strength of the surveillance system in China, detailed information is available about each individual case (including the time of symptom onset and location of residence), and case reporting is believed to be very high.

By adapting and applying a continuous diffusion network approach (Gomez Rodriguez *et al.*, 2014) within a Bayesian framework introduced in Chapter 2 I quantify case reproduction numbers, R_c, and uncertainty in these estimates for all *P. vivax* and *P. falciparum* cases of malaria recorded in Yunnan province between 2011 and 2016. I incorporate these estimates into Bayesian geostatistical models and time series approaches to estimate how R_c varied over space and time which I use to estimate timelines to elimination and likelihood of resurgence.

4.2 Methods

4.2.1 Data

Anonymised case data for all confirmed (N=4078) and probable (N=285) malaria cases reported between 2011 and 2016 in Yunnan Province (N =4390) were obtained from the Chinese Centre for Disease Control (CCDC). For each case, data included date of symptom onset, GPS coordinates of symptom onset address, health facility address, travel history, and in some cases, the GPS coordinates of presumed location of infection.

Of these cases, the majority were *P. vivax* (N = 3469, of which 2858 were classified as imported). Of all recorded *P. falciparum* cases (N=791), 91% (N=720) were imported. Small numbers of *P. malariae* (N=8) and *P. ovale* (N=1) were excluded from the analysis. Cases defined as "untyped" (N=67) were also excluded. A small number (N=27) of cases classified as mixed infection were included in the separate analyses of each species. A full breakdown of the cases and species composition across Yunnan province between 2011 and 2016 is included in Tables 4.1 and 4.2.

Table 4.1: Cases by diagnosis type (probable and confirmed) and species across Yunnan Province

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Confirmed	27	770	8	1	3269	3
Probable	0	21	0	0	200	64

Table 4.2 Cases by imported/local status and species across Yunnan province

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Local	4	71	0	0	611	51
Imported	23	720	8	1	2658	16

4.2.2 Surveillance system in China

The PRC has a sophisticated malaria surveillance system, described in detail elsewhere (Yang *et al.*, 2012; Cao *et al.*, 2014; Feng *et al.*, 2014; Zhou *et al.*, 2015; Hu *et al.*, 2016) and summarised here. Surveillance is carried out in both a passive and reactive manner, organised and administered at the national, provincial and county level. The centralised China Information System for Disease Control and Prevention (CISDCP) receives daily updates on case reports from health facilities

Passive detection occurs according to a protocol at the local level, such that cases are tested by microscopy or Rapid Diagnostic Test (RDT) and reported to the central information system within 24 hours. Case investigation is then pursued, where cases are confirmed via double readings of microscopy slides and in some cases polymerase chain reaction (PCR) confirmation at provincial laboratories. At this point it is also determined whether the case is locally acquired or imported by taking patient travel history – if a patient has travelled to a malaria endemic country within a month of symptom onset the case is then classified as imported (Cao *et al.*, 2014). Case investigation should be completed within three days.

Foci investigation occurs once a case is detected to determine whether the foci is inactive, active or a pseudo-focus based upon the absence or presence of suitable vectors (inactive), and presence or absence of malaria in the resident area of the case if imported (pseudo-focus). Reactive Case Detection (RACD) of case contacts and populations with demographic similarities (for example individuals working in the same industry and vicinity as the case) is carried out. In active foci more intensive RACD screening of a larger pool of neighbours and contacts is carried out using Rapid Diagnostic Tests (RDTs) for immediate detection, followed by PCR testing of blood spots to detect low-density infections. IRS (Indoor Residual Spraying) is also carried out(Cao *et al.*, 2014; Feng *et al.*, 2014; Zhou *et al.*, 2015).

The Ministry of Health (MoH) in China has also been measuring the timeliness of the recommended protocol and follow-on ability to meet these targets. It was found that the one-day target for case reporting was almost always met because this is required by law**.** In the years following the introduction of the 1-3-7 policy, the proportion of cases investigated within three days increased from roughly 55% in 2011 to almost 100% by 2013. However the programme took longer to achieve the seven day focal point investigation goals, with just over 50% of foci investigated and treated within seven days by the end of 2013 (Cao *et al.*, 2014). Nevertheless, by 2015, adherence to the 1-3-7 strategy improved and this figure increased to an estimated 96% (Zhou *et al.*, 2015)**.** Whilst some cases could still be missed, the thoroughness of the approach means numbers of missing cases are likely to be small.

4.2.3. Defining the serial interval distribution

The serial interval is defined as the time between a given case showing symptoms and the subsequent cases they infect showing symptoms (Fine, 2003). For a given individual j at time t_j, showing symptoms before individual i at time t_i, the serial interval distribution specifies the normalised likelihood or probability density of case i infecting case j based on the time between symptom onsets,

$t_i - t_j$. The serial interval summarises several distributions including the distribution of a) the times between symptom onset and infectiousness onset, b) the time for humans to transmit malaria parasites to mosquito vectors, c) the period of mosquito infectiousness, and d) the human incubation period.

Taking a similar approach to my work (Routledge *et al.*, 2018) described in Chapters 2 and 3, I defined a prior distribution of possible serial interval distributions for malaria. The serial interval distribution of treated, symptomatic *P. falciparum* malaria, previously characterised using empirical and model based evidence (Thomas S. Churcher *et al.*, 2014; Huber *et al.*, 2016a) was adapted to inform the prior distribution for the relationship between time and likelihood of transmission between cases in China. I analysed *P. vivax* cases and *P. falciparum* cases separately. The prior distribution was defined to be flexible enough to reflect both the biology of *P. vivax* and *P. falciparum* as well as the dominant vector species in Yunnan (recent surveys in Yunnan province have found *Anopheles sinensis* to be the dominant vector species in mid-elevation areas and rice paddies and *Anopheles minimus* s.l. the dominant species in low elevation areas (Shi *et al.*, 2017; Zhang *et al.*, 2018)) and to allow for possible variation in transmission dynamics, for example due to untyped infections or delays in seeking treatment. In addition, there is a possibility of a small number of asymptomatic or undetected and therefore untreated infections contributing to ongoing transmission, which will typically have a longer serial interval. I use a shifted Rayleigh distribution to describe the serial interval of both species, which can be varied by changing two parameters: α and γ. The parameter α governs the overall shape of the distribution, and γ is the shifting parameter accounting for the incubation period between receiving an infectious bite and the onset of symptoms. The γ shifting parameter was fixed at 15 days to account for the extrinsic incubation period within the mosquito and the minimum time between infection and suitable numbers of gametocytes in the blood to lead to symptom onset (Warrell and Gilles, 2002). The prior for the α parameter determining the shape of the distribution was given a Normal distribution with mean 0.003 and standard deviation 0.1 (illustrated in **Figure 4.2**), giving an expected time between symptom onset of one case and symptom onset of the case it infects of 36 days, with the parameter value in the 2.5 percentile of prior having an expected serial interval of 21 days and the equivalent parameter from the 97.5 percentile having an expected serial interval of 60 days. By comparison the expected values for treated *P. falciparum* from existing literature range between 33 and 49.1 days (95%CI = 33- 69) (Churcher *et al.*, 2014; Huber *et al.*, 2016). Depending on how much uncertainty there is in the serial interval of malaria, the prior for α, the shaping parameter for the SI of malaria, may be varied. I explored the effects of different priors on the likelihood and posterior estimates. I used the same mean value for α (0.003) but set the α prior to standard deviation between 1 and 0.01. The results of considering different priors for α, the parameter shaping SI distribution on estimated R_c values over time is shown in Figure 4.4.

4.2.3 Defining the transmission likelihood

I assume cases were classified correctly from case investigation as imported or locally acquired based on recent travel history. Following this assumption, locally acquired cases could have both infected others and been infected themselves. However imported cases could only infect others, as I assume their infection was acquired outside of the country. Given the evidence (Cao *et al.*, 2014; Zhou *et al.*, 2015; Hu *et al.*, 2016) of strong adherence to the 1-3-7 policy for reporting and response to case detection, and no evidence of relapse within the dataset (as each patient is given a unique identifier), I assume that an individual can only be infected once by a case that has shown symptoms earlier in time.

4.2.4 Transmission model specifics

To estimate the underlying pathways of transmission and likelihood of cases being linked by infection, I adapt and extend the NetRate algorithm (Gomez Rodriguez *et al.*, 2014) as described in Chapter 2.

The adapted model introduces the ability to model serial interval functions, account for imported versus local infections and provides provision for missing or unobserved sources of infection, called epsilon edges (Rodriguez and Schölkopf, 2012; Routledge *et al.*, 2018)). I also extended the *NetRate* algorithm from a frequentist to a Bayesian framework to incorporate prior knowledge about the serial interval of malaria. This analysis was carried out via TensorFlow, via the *TensorFlow* and *reticulate* packages in R (version 3.6.0).

The data analysed consider of a set of n infections/nodes $\boldsymbol{I} \in (I_1, \dots, I_n)$ with associated times $\boldsymbol{t} = \{t_1, \dots t_n\} \in \mathbb{R}^+$ and binary yes/no importation status $\boldsymbol{\pi} = \{\pi_1, .., \pi_n\} \in \{1,0\}$ .The serial interval distribution of malaria, defining the probability individual I_j infected individual I_i at times $t_i > t_j$ is defined through a shifted Rayleigh distribution $f(t_i|t_j; \alpha, \gamma) = \alpha(t_i - t_j - \gamma)e^{-\alpha(t_i - t_j - \gamma)}$ for shaping parameters α and γ (Routledge et al., 2018). For this analysis I fix $\gamma = 15$ days, fixed at 15 days to account for the extrinsic incubation period within the mosquito and the minimum time between infection and suitable numbers of gametocytes in the blood to lead to symptom onset (Kitchen and Boyd, 1937; Warrell and Gilles, 2002).

If one assumes that infections are conditionally independent given the parents of infected nodes, then the likelihood of a given transmission chain can be defined as

$$f(\boldsymbol{t}; \boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} f(t_i | t_1, \dots, t_n \backslash t_i; \boldsymbol{\alpha}) \ (1)$$

Where $\boldsymbol{\alpha}$ is a parameter matrix. Computing the likelihood of a given transmission chain thus involves computing the conditional likelihood of the infection time of each infection (t_i) given all other infections, leaving out t_i ($t_1, \dots, t_n \backslash t_i$). If I make the assumption that a node gets infected once the first parent infects it (Kempe, Kleinberg and Tardos, 2003) and define a survival function

$$S(t_i|t_j; \alpha_{j,i}) = 1 - \int_0^{t_i - t_j} f(t_i|t_j; \alpha_{j,i})\, dt \ (2)$$

as the probability that infection I_i is *not* infected by infection I_j by time t_i then I can simplify the transmission likelihood as

$$f(\boldsymbol{t};\boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} \sum_{I_j : t_j < t_i} f(t_i | t_j; \alpha_{j,i}) \prod_{I_k : t_k < t_i, I_k \neq I_j} S(t_i | t_k; \alpha_{k,i}) \quad (3)$$

In this conditional likelihood the first term computes the probability the I_j infected I_i and the second term computes the probability that I_i was not infected by any *other* previous infections excluding I_j. This likelihood therefore accounts for competing infectors and finds the infector most likely to have infected I_i. To remove the $k \neq j$ condition makes the product independent of j and results in the likelihood

$$f(\boldsymbol{t};\boldsymbol{\alpha}) = \prod_{t_i \in \boldsymbol{t}} \prod_{I_k : t_k < t_i} S(t_i | t_k; \alpha_{k,i}) \sum_{I_j : t_j < t_i} \frac{f(t_i | t_j; \alpha_{j,i})}{S(t_i | t_j; \alpha_{j,i})} \quad (4)$$

In equation 4, $f(\cdot) / S(\cdot) = H$ is the hazard function or rate and represents the instantaneous infection rate between individuals I_i and I_j.

Assuming all cases reaching health workers or health facilities are recorded, missing cases may be generated by two processes. Symptomatic cases may be missed by not seeking care or not being found through active case detection, or cases may be asymptomatic and therefore unlikely to seek care or be detected. The latter may have densities of parasites in their blood which are too low to be detectable by microscopy if active case detection occurs. These processes apply to both imported cases or locally acquired cases. I assume the pool of asymptomatic cases in the country is low and has a small contribution to ongoing transmission. To account for unobserved infectors within this framework I include a time-independent edge that can infect any individual. The survival and hazard functions for this edge are defined as $S_0(\epsilon_i) = e^{-\epsilon_i}$ and $H_0 = \epsilon_i$. The introduction if this edge also makes the likelihood stable and never singular because the probability will not collapse to zero. As we will see below, because of this optimisation problem these edges are encouraged to be sparse and only invoked if no other infectors can continue the transmission chain.

In addition to unobserved edges, I assume that observed imported infectors can infect other cases but cannot be infected themselves. The final likelihood incorporating these two modifications becomes

$$f(\boldsymbol{t};\boldsymbol{\alpha},\boldsymbol{\epsilon}) = \prod_{t_i \in \boldsymbol{t}} S_0(\epsilon_i) \prod_{I_k : t_k < t_i} S\left(t_i \middle| t_k; \alpha_{k,i}\right) \left(H_0(\epsilon_i) + \sum_{I_j : t_j < t_i} H\left(t_i \middle| t_k; \alpha_{k,i}\right)\right) \quad (5)$$

In order to find the optimal parameters for $\boldsymbol{\alpha}, \boldsymbol{\epsilon}$ I minimize the following log likelihood subject to positive constraints on the parameters:

$$minimize_{\boldsymbol{\alpha},\boldsymbol{\epsilon}} - \log f(\boldsymbol{t};\boldsymbol{\alpha},\boldsymbol{\epsilon}) \qquad subject\ to\ \boldsymbol{\alpha}, \boldsymbol{\epsilon} > 0 \text{ for all values of } i, j \quad (6)$$

This optimisation problem is convex and guarantees a consistent maximum likelihood estimate (Gomez Rodriguez *et al.*, 2014). To prevent biologically implausible serial interval distributions, I impose a truncated normal prior probability distribution on $\boldsymbol{\alpha}$ ~Normal(0.003,0.1) [0,0.01]. When optimising the likelihood, I include this prior probability and therefore evaluate the Bayesian Maximum-a-Posteriori estimate.

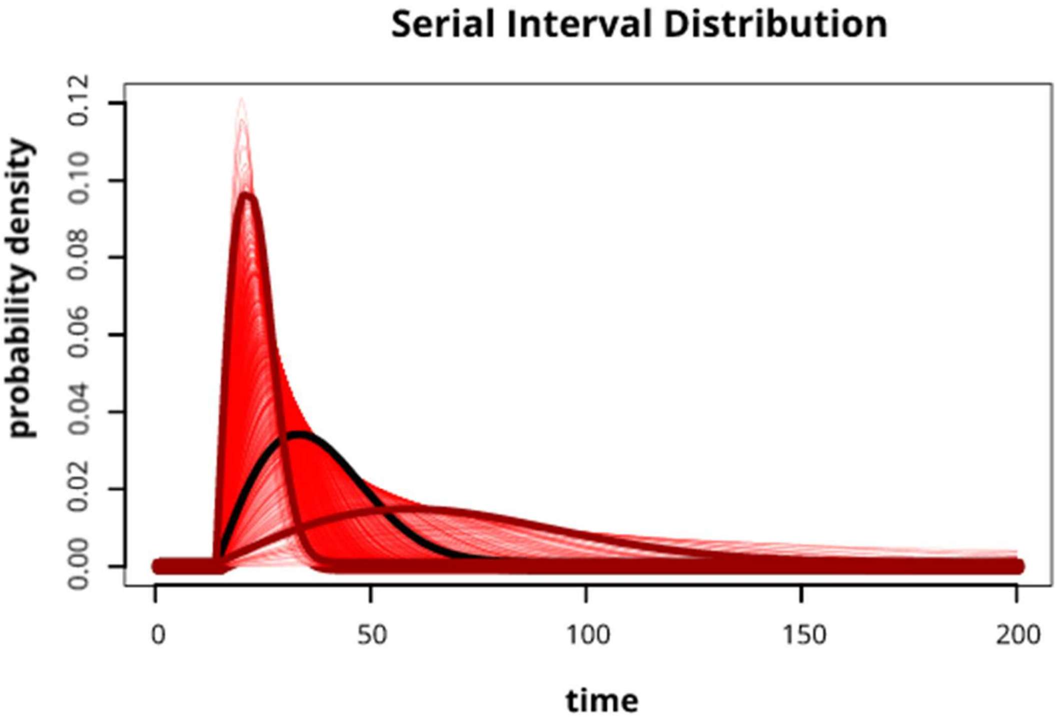


Figure 4.2: Plot illustrating the serial interval distributions used in the analysis. *Red lines show 300 draws from the prior distribution used in the analysis for the Serial Interval distribution. The black line represents the expected function and the maroon lines represent the 2.5 and 97.5 quantile values of the prior distribution for the shaping parameter, a.*

4.2.5 Estimating R_c

Individual reproduction numbers were estimated for each case by creating a matrix where each column represents a potential infector and the rows represent a potential infectee, describing which infector edges are connected to infectees and the normalised likelihood of the cases being connected by a transmission event. Intuitively then, taking the row sums gives the (fractional) number of secondary infections and therefore a point estimate of the time varying reproduction number $R_c(t_j)$ This reflects for an individual, how many people they subsequently infect. When multiple individuals have been infected at a given time and/or place, one can take the mean individual R_c and uncertainty in this value as an indicator of reproduction numbers for a given time and/or location.

In contrast to traditional methods based on Wallinga and Teunis (Wallinga and Teunis, 2004) the algorithm presented here encapsulates not only the pairwise likelihood of transmission between two cases, but conditions this likelihood on the impact of competing edges in the inferred network (the survival of an edge). The resulting estimates of R_c therefore consider the overall transmission tree in optimisation and allow for missing cases within the tree.

4.2.6 Estimating timelines towards elimination and risks of resurgence

It is important for national malaria control programmes to have information about likely timelines to elimination, chances of resurgence and uncertainty in these estimates. Using the distribution of $\mathcal{R}_c$ values and their seasonal and general trends, I analysed time series using the *Prophet* tool and R package (Taylor and Letham, 2017) to explore general and seasonal trends as well as the impact of holidays on results.

This approach applies an additive regression model

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (6)$$

which is composed of trend, seasonal and holiday functions , where $y(t)$ is the observations at time t, $g(t)$ is the general trend, modelled by a logistic growth model, $s(t)$ is the seasonal effect, modelled by Fourier coefficients, $h(t)$ is the effect of specific holiday dates and ϵ_t is the error term. I explored the overall trend as well as seasonal trends, in addition tothe predicted R_c between 2011 and the beginning of 2020. I also explored the impact of the national holiday periods, some of which involve large scale movement, such as the *Chunyun* period around the spring festival. I cross-validated predictions and calculated root mean squared error (RMSE) and mean absolute error (MAE).

4.2.7. Mapping R_c

Transmission risk map estimates were constructed by separating individual reproduction numbers into those above and below $R_c = 1$ The latitude and longitude of the reproduction numbers were included in a binomial Gaussian random field model implemented in rINLA (Rue, Martino and Chopin, 2009), in which demographic and environmental covariates were used (Table 4.3) to estimate the likelihood of a case having $R_c > 0$ in the area each year from 2011 to 2016. This is a measure of malaria "receptivity" or underlying transmission potential rather than overall malaria risk, as importation likelihood is not quantified in this analysis. Area under the curve (AUC) scores from leave-one-out cross validation were used to assess model fit (**Figure 4.3**)

The underlying spatial statistical model was fitted to binomial data, where when $R_c$was above zero, it was assigned a value of one, and when Rc was equal to zero it was assigned a value of 0 ($R_c > 0 = 1; R_c = 0$) using the logit link function:

$$R^{+}_{>0,i} \sim Binomial(p_i, N_i)$$

$$log(p_i/(1-p_i)) \sim GP(\mu, Q)$$

$$\mu = \alpha + X_i\beta$$

$$Q = \boldsymbol{K}^{-1}_{space}$$

$$\boldsymbol{K}^{-1}_{space} = \ solve\ (k^2 - \Delta)^{\frac{\alpha}{2}}\big(\tau x(s)\big) = W(s)$$

where $R^{+}_{>0,i}$ are the number of binary data points where $R_c > 0 = 1$, N_i is the number of trials, p_i is the estimated $R_{>0}$, expressed as a logit transformed probability and modelled as a Gaussian process with μ and precision Q. The GP mean μ is a linear function of a global intercept α and a vector of

β coefficients from space-time indexed covariate values $\boldsymbol{X}_i$. Q is a sparse precision matrix constructed as the sparse finite element solution to the stochastic partial differential equation $(k^2 - \Delta)^{\frac{\alpha}{2}}(\tau x(s)) = W(s)$, where Δ is the Laplacian, k is the spatial scale/range parameter, τ controls the variance, α is the spatial smoothness parameter (fixed at $\alpha = 2$), and $W(s)$ is the spatial white noise process. To account for the curvature of the earth the distance metric *s* is defined on a spherical manifold in Cartesian $\mathbb{R}^3$.

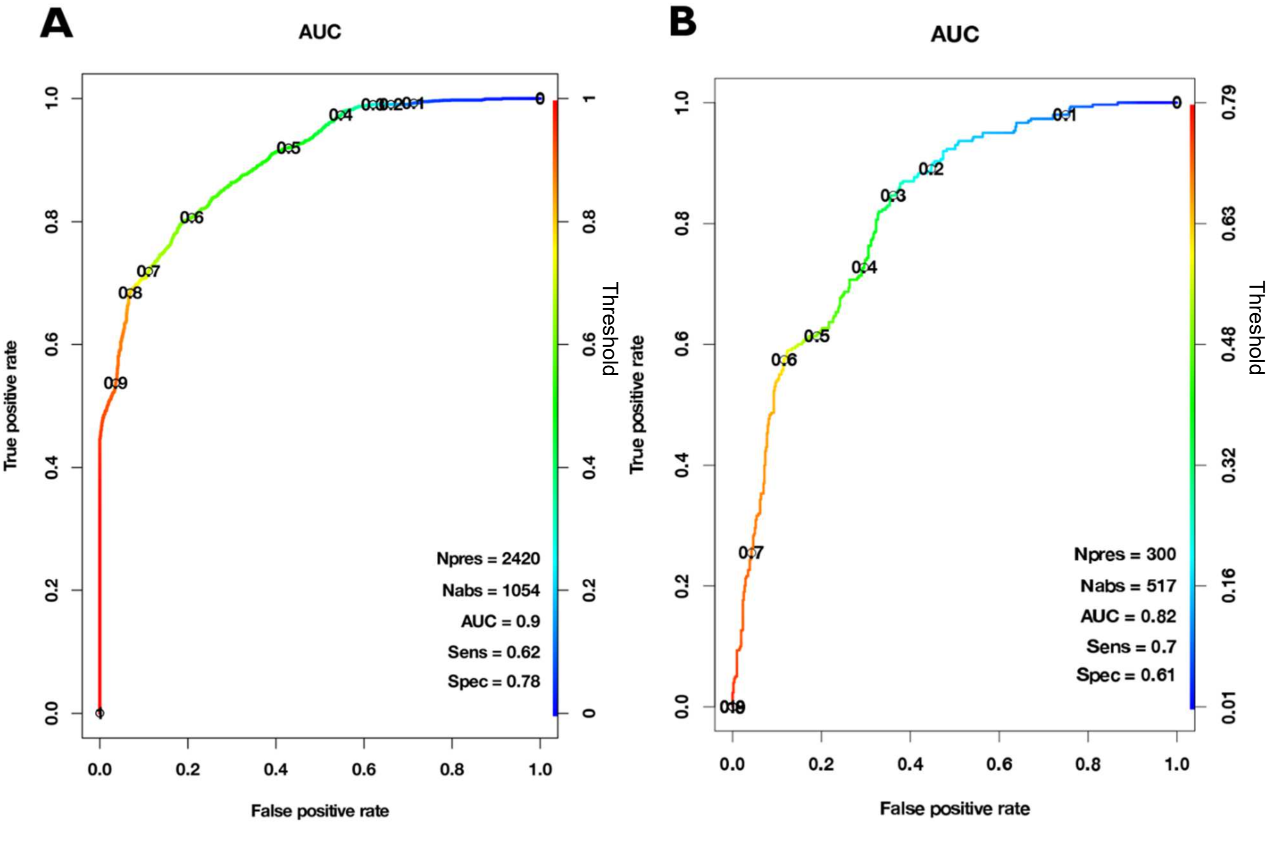


Figure 4.3: Area Under the ROC Curve (AUC) from cross validation of geostatistical model used to create riskmaps of P(Rc>0) *for A) P. vivax and B) P. falciparum. The colours and labels (illustrated in the scale bar on the right side of the x axis) represent the threshold for classification as 1 (Rc>0) or 0 (Rc=0). When the threshold is decreased, more positive values are returned, thus sensitivity (the true positive rate) increases and specificity (1- false positive rate) decreases.*

Table 4.3: Table summarising covariates used in geostatistical model

Variable Class	Variable(s)	Source	Type
temperature	land surface temperature (day, night and diurnal flux)	MODIS product	dynamic monthly
precipitation	mean annual precipitation	WorldClim	synoptic
elevation	digital elevation model	SRTM	static
infrastructural development	accessibility to urban centres and night-time lights	modelled product and VIIRS	static
moisture metrics	aridity and potential evapotranspiration	modelled products	synoptic

Table 4.4: Posterior covariate parameter estimates for P. vivax R_c risk map

Covariate	*Mean*	*SD*	*0.025 Quantile*	*0.5 Quantile*	*0.975 Quantile*	*Mode*
Elevation	-0.00065	0.000369	-0.00137	-0.00065	7.82E-05	-0.00065
Day temperature (monthly)	0.040258	19.13436	-37.5269	0.03972	37.57611	0.040258
Night temperature (monthly)	-0.11265	19.13447	-37.6801	-0.11319	37.42342	-0.11265
Difference between day and night-time temperature (monthly)	-0.07346	19.13443	-37.6408	-0.074	37.46253	-0.07346
Precipitation	-0.00041	0.000248	-0.00089	-0.00041	7.96E-05	-0.00041
Urban	-0.06908	0.301495	-0.66102	-0.06909	0.522361	-0.06908
Intercept	4.065973	1.985619	0.167532	4.065917	7.961159	4.065973

Table 4.5: Posterior covariate parameter estimates for P. falciparum $\mathbf{R_c}$ risk map

Covariate	*Mean*	*SD*	*0.025 Quantile*	*0.5 Quantile*	*0.975 Quantile*	*Mode*
Elevation	0.000112	0.000502	-0.00087	0.000112	0.001097	0.000112
Day temperature (monthly)	-0.01005	19.12776	-37.5643	-0.01059	37.51285	-0.01005
Night temperature (monthly)	-0.03118	19.12771	-37.5853	-0.03172	37.49163	-0.03118
Difference between day and night-time temperature (monthly)	-0.00245	19.12769	-37.5566	-0.00299	37.52031	-0.00245
Precipitation	0.00015	0.00029	-0.00042	0.00015	0.000718	0.00015
Urban	0.361755	0.452532	-0.52672	0.361743	1.249487	0.361755
Intercept	-1.86989	3.045358	-7.84895	-1.86998	4.104185	-1.86989

4.3 Results

4.3.1 R_c estimates over time

Between 2011 and 2016, 3496 cases of probable and confirmed *P. vivax* infection including mixed infections were observed in Yunnan province (2881 imported, 615 locally acquired). Including mixed infections, 818 *P. falciparum* infections were observed, of which 75 were locally acquired. The mean R_c value estimated for *P. vivax* during this period was 0.171 (95% CI = 0.165, 0.178) and 0.089 (95% CI = 0.076, 0.103) for *P. falciparum* case. A decline in R_c over time was estimated for both *P. vivax* (Figure 4.6) and *P. falciparum* (Figure 4.6), with the most rapid declines occurring between 2012 and 2014 (Figure 4.5, Figure 4.6). No R_c values above one were observed after 2014 for either species. These findings are consistent with varying levels of uncertainty about the serial interval distribution (Figure 4.4).

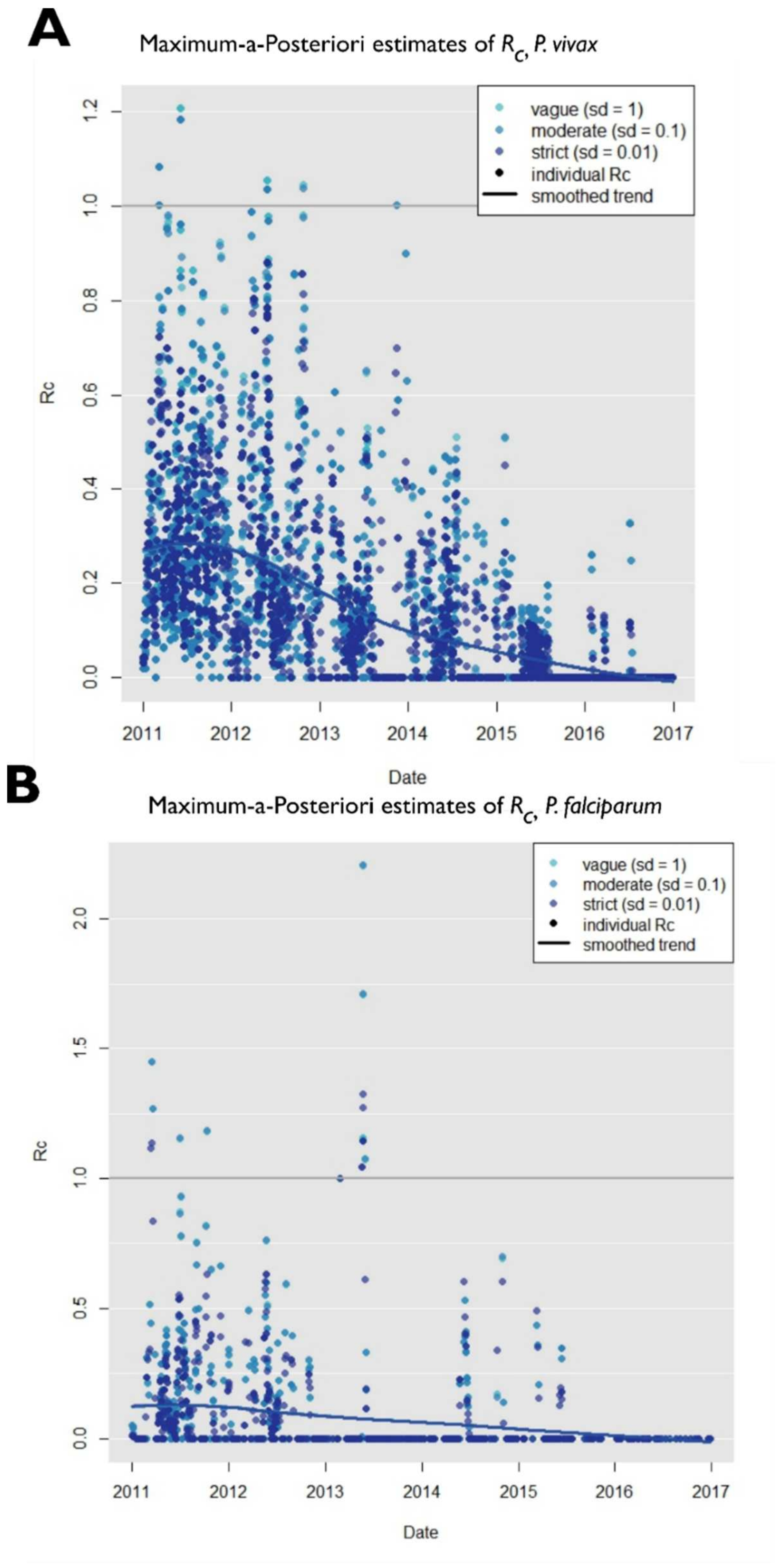


Figure 4.4: Plot showing the impact of varying the prior distribution for alpha on results *for A) Plasmodium vivax and B) Plasmodium falciparum. Figures A and B show the estimated maximum a posteriori Rc estimates for a normally distributed prior with a mean of xx, but where the standard deviation was varied between 1 (light blue) and 0.01 (dark blue). The grey horizontal line represents an Rc of one. The smoothed loess curve for each prior is shown as a solid line and was consistent across all priors.*

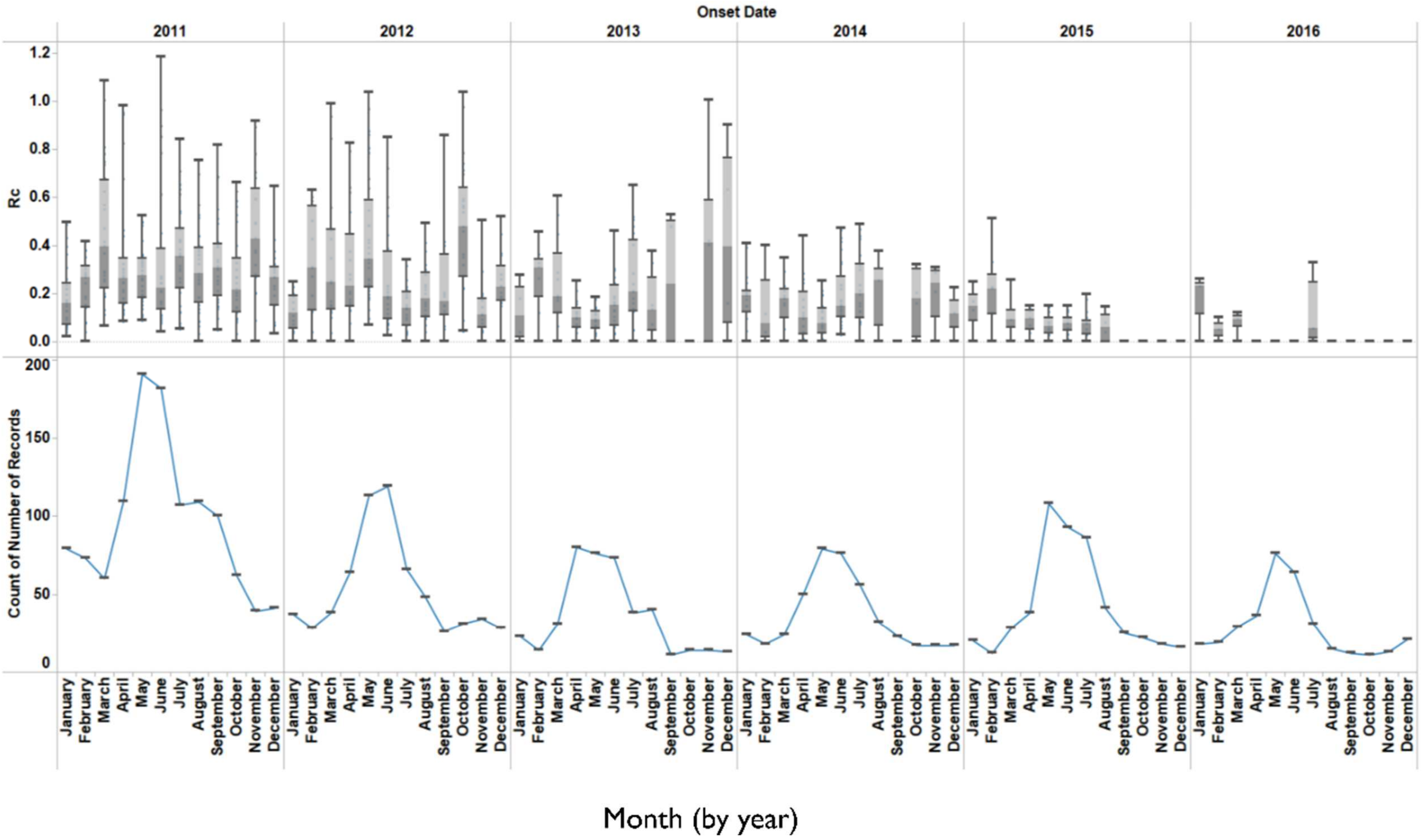


Figure 4.5: ***Rc and case counts by month and year.*** *The upper panel shows* R_c *estimates by month of symptom onset date, stratified by year . The lower panel shows case counts by month, stratified by year.*

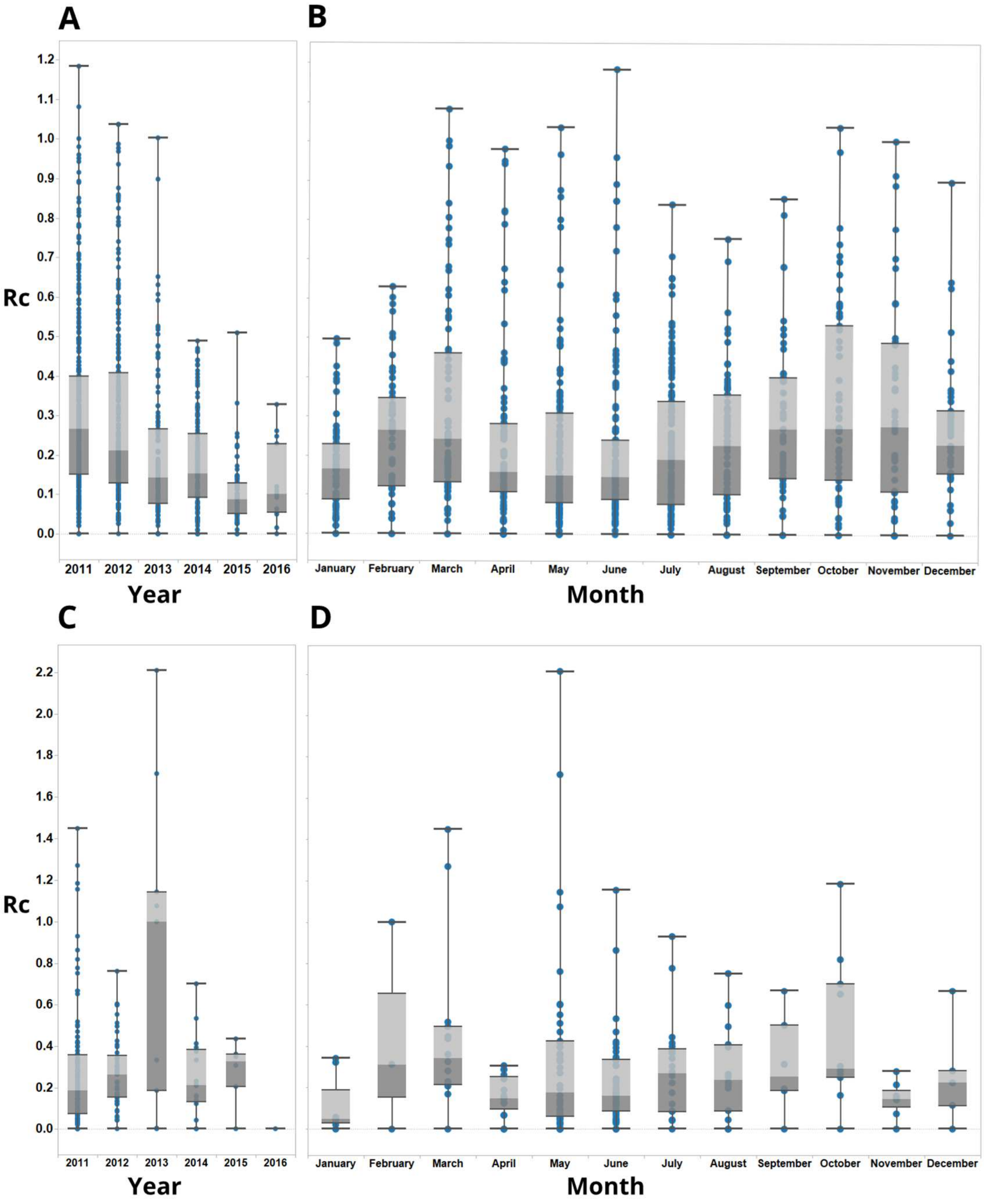


Figure 4.6: R_c estimates by year and month. *Boxplots showing R_c estimates for P. vivax (A and B) and P. falciparum (C and D), aggregated by year (A and C) and month (B and D) of symptom onset. Points represent individual R_c estimates. Boxplots show median, upper and lower quartiles for R_c each.*

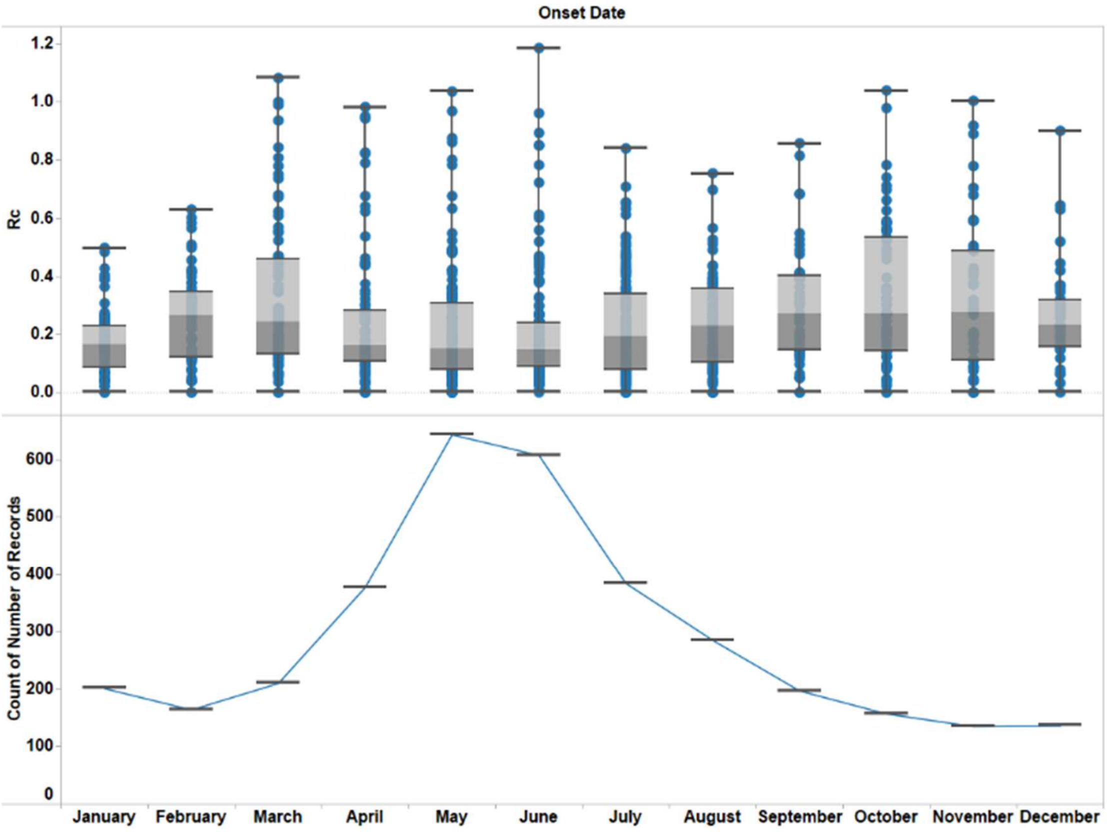


Figure 4.7:* R_c *by month compared to incidence. *Figure showing reproduction numbers over time compared to incidence – showing patterns of incidence are different to reproduction numbers – likely importation driving increase in records rather than* R_c

4.3.2 *Unobserved sources of infection*

For *P. vivax*, 19 out of 615 locally acquired cases were estimated to have a moderate chance of having an unobserved source of infection (estimated $0.8 \geq \varepsilon \geq 0.5$) and 2 cases were estimated to have a high chance of an unobserved source of infection (estimated $\varepsilon \geq 0.8$) . Together, this represents 3% of locally acquired cases with a moderate to high chance of external infection sources. For *P. falciparum,* 2 out of 75 local cases were estimated to have a high chance of having an unobserved source of infection (estimated $\varepsilon \geq 0.8$) and no other cases were estimated to have a moderate change of having an unobserved source of infection (Figure 4.8)

4.3.3 Spatial patterns of R_c

As transmission declined between 2011 and 2016, I observed a reduction in the incidence of locally-acquired cases which is reflected in a reduction in the estimates of the reproduction number of each locally-acquired case for both species (Figure 4.9). I estimate a decline in the probability of a reproduction number for a *P. vivax* case being above zero over this period (Figure 4.10), with the central parts of the province being the first to reach lower risks of non-zero R_c. The border area neighbouring Myanmar, where most cases were observed, had the lowest amount of uncertainty in the estimates. *P. falciparum* shows a decline in risk of $R_c > 0$ across the province, with the more isolated areas in the north of the province showing both the highest predicted risk (Figure 4.10) but also the most uncertainty, due to a lack of cases observed there (Figure 4.11). By 2016 all areas have reached a low risk, although there is more uncertainty in these estimates compared to *P. vivax*, almost certainly due to the smaller sample size.

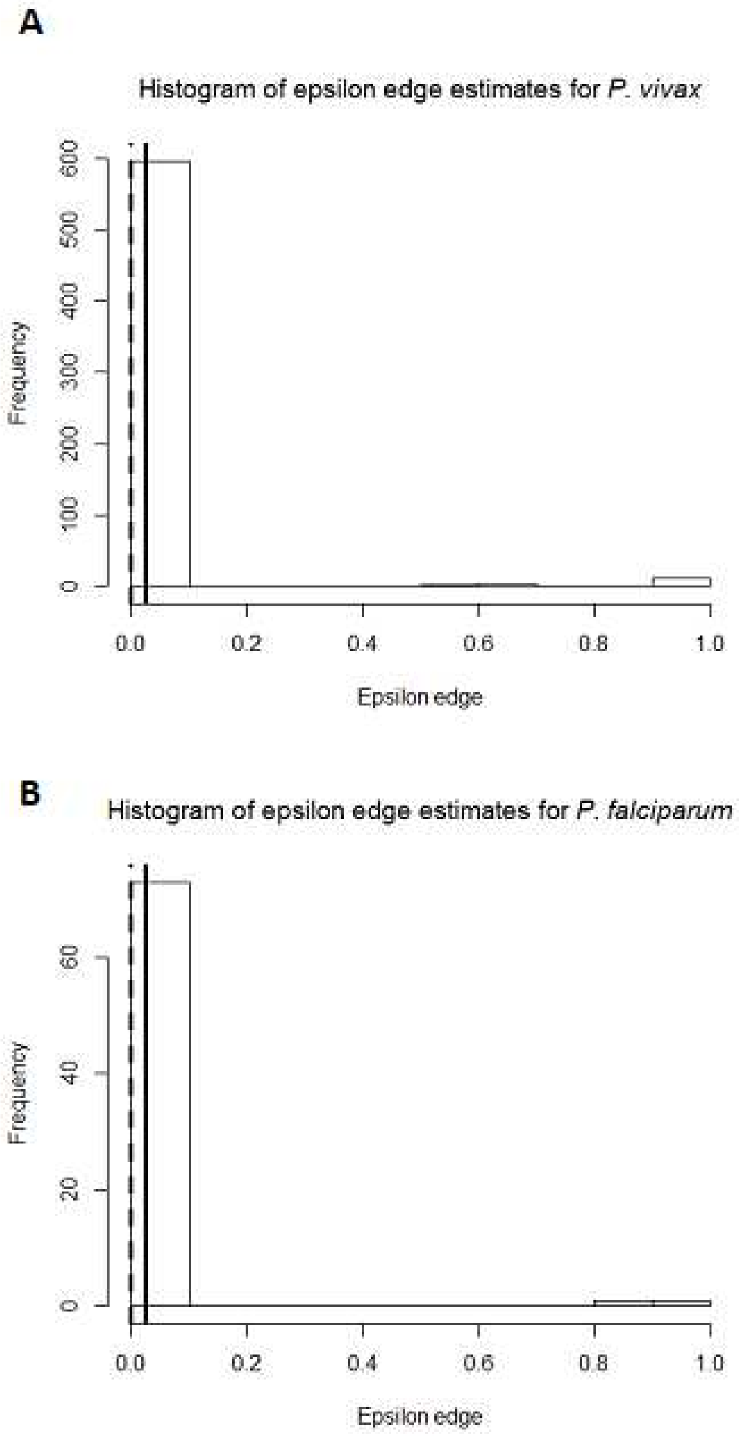


Figure 4.8: Histogram of epsilon edges estimated by model.

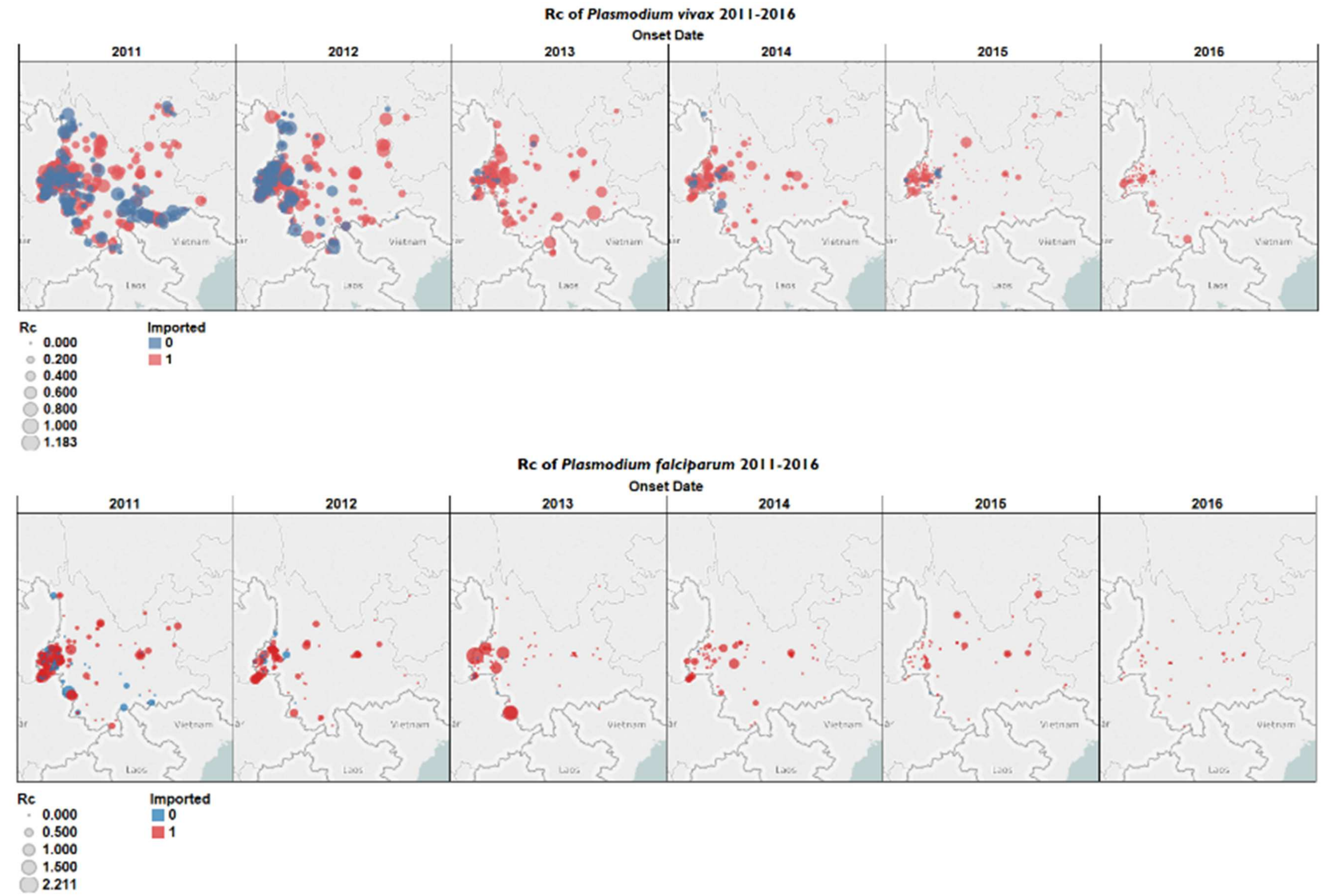


Figure 4.9: Map of R_c estimates by year *for A) P. vivax and B) P. falciparum. Blue points represent locally acquired cases; red points represent imported cases. The diameter of the point represents the size of the R_c estimate*

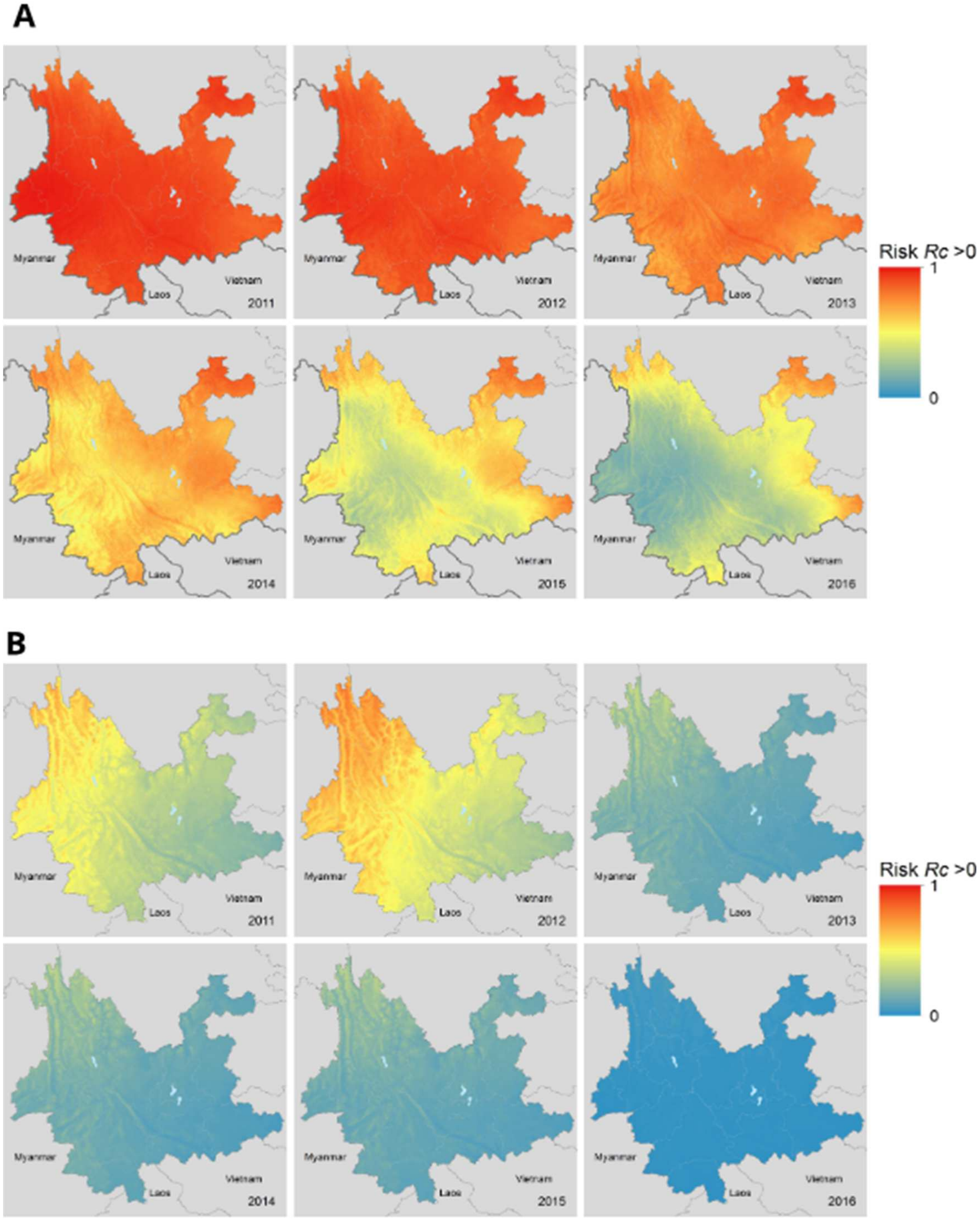


***Figure 4.10: Map of risk of Rc > 0 and uncertainty in this estimate from application of a Gaussian Process geostatistical model with a logit link function to times and locations of observed cases** for A) P. vivax and B) P. falciparum malaria across Yunnan province in each year 2011-2016. This represents the risk of a case having an Rc>0 if observed, stratified by year.*

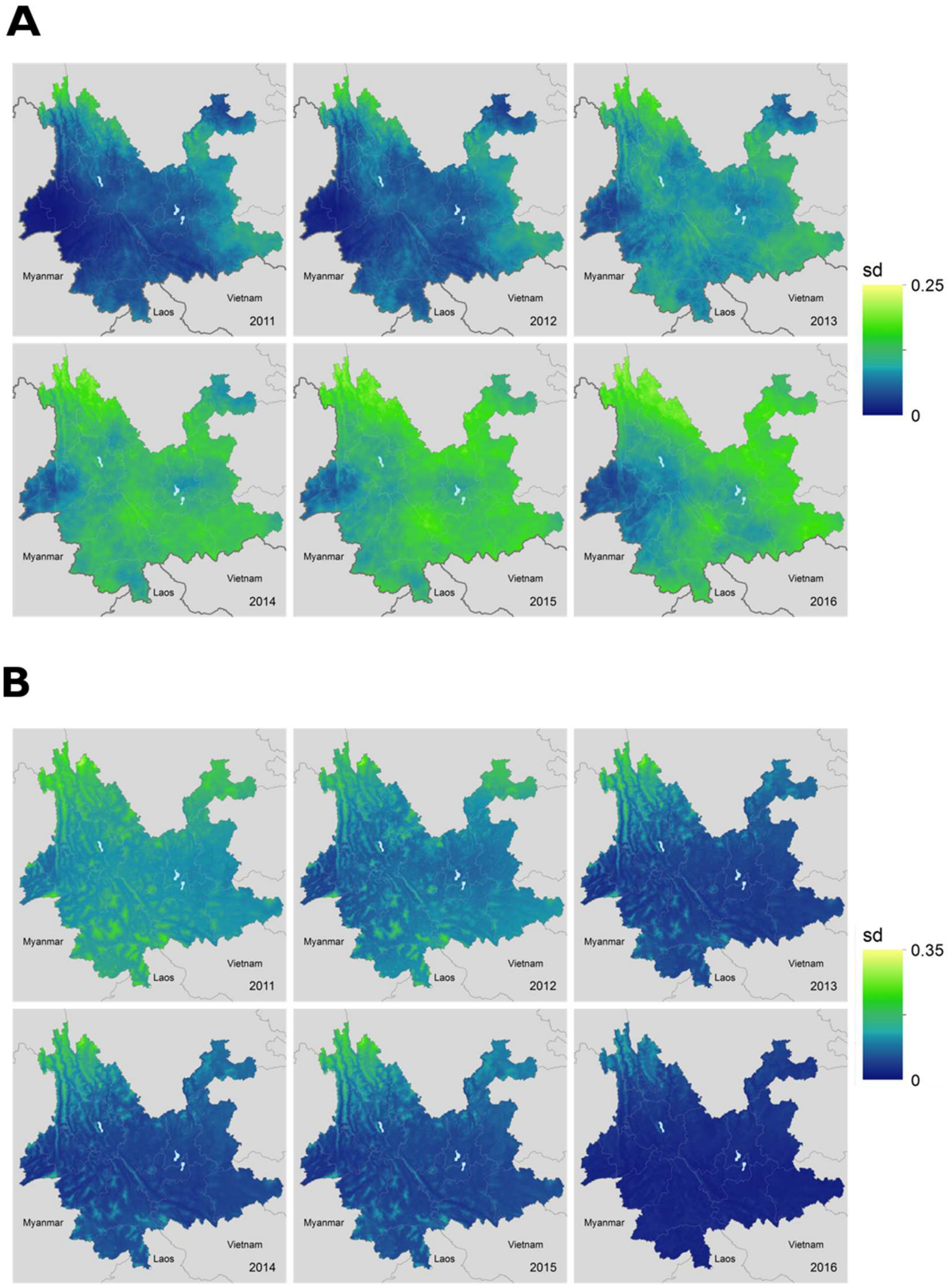


Figure 4.11: Standard deviations in estimate of risk mapped in Figure 4.10 from binomial INLA model. For A) P. vivax and B) P. falciparum

4.3.4 Short – term predictions and temporal patterns in timeseries of Plasmodium vivax cases

Using the *Prophet* additive regression model to make short-term predictions, a posterior mean R_c of 0.005 (95% CI = 0 - 0.34) was estimated for *Plasmodium vivax* cases up to 2020 (Figure 4.12A). A declining trend was observed, with the fitted trend for R_c, which estimates the general trend, separate to the influence of seasonal and holiday effects, declining from 0.31 (95% CI = 0.31, 0.34) at the start of 2011 to 0.004 (95% CI =0.002-0.006) by the end of 2019 (Figure 4.12B). I estimate a small effect of holiday periods to differences in R_c observed, with Chinese New Year and National Day associated with small increase risk in R_c of 16% (95% CI = -112%, 152%) and 39% (95% CI = -43%, 118%) (Figure 4.12B) which in this very low transmission context could increase the probability of small outbreaks of local transmission in areas in which high rates of importation occur, although very wide credible intervals were associated with these estimates. I did not identify a clear seasonal trend, however two peaks were identified, with up to 20% (95% CI = 14%, 26%) increases and 28% decreases (95% CI -35%, -22%) in risk of R_c associated with April/October and the beginning of January respectively (Figure 4.12B).

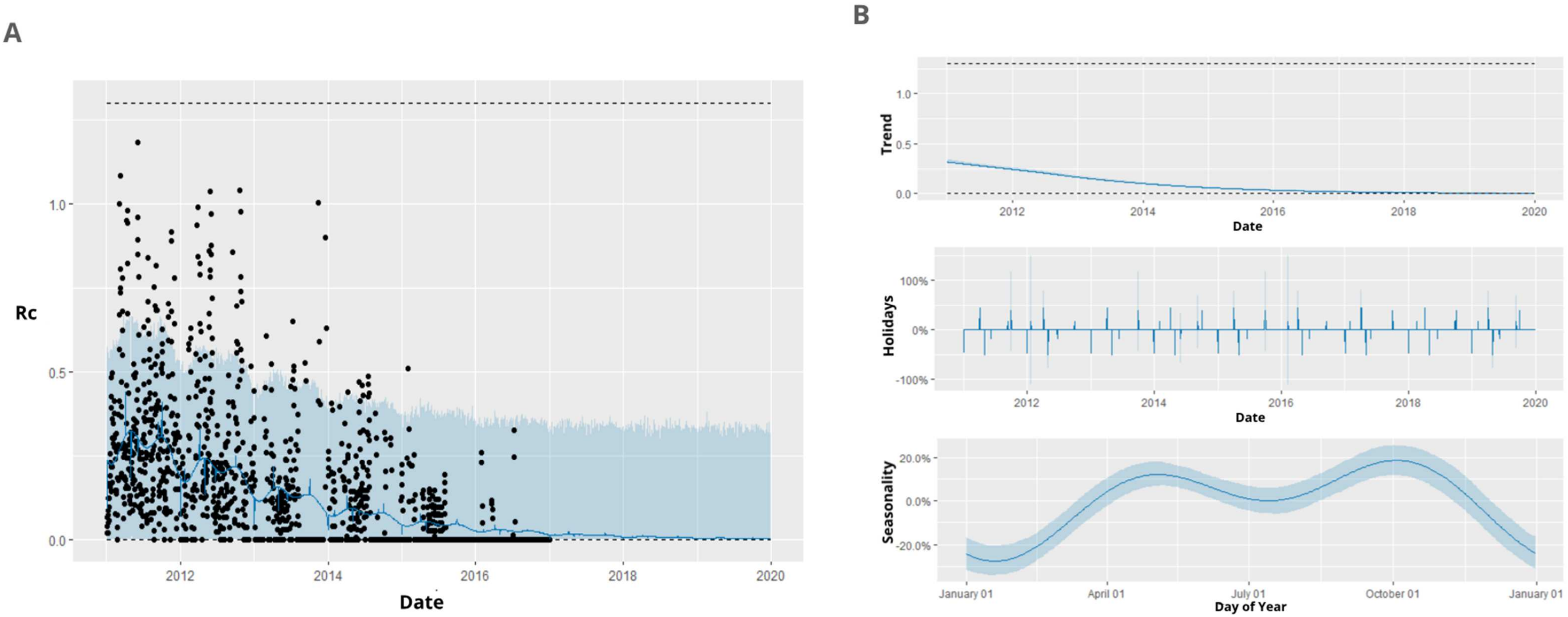


Figure 4.12: Results of time series analysis using an additive regression approach *A) Black points show estimated individual R_c values, blue line represents prophet model predictions for mean R_c on that day, shaded blue area shows 95% credible interval of prediction. B) Decomposed time series model, showing the general trend, fitted holiday effect and seasonal effect. For seasonal and holiday effects the y axis shows the percentage increase or decrease in R_c predicted which is attributable to a seasonal or holiday effect*

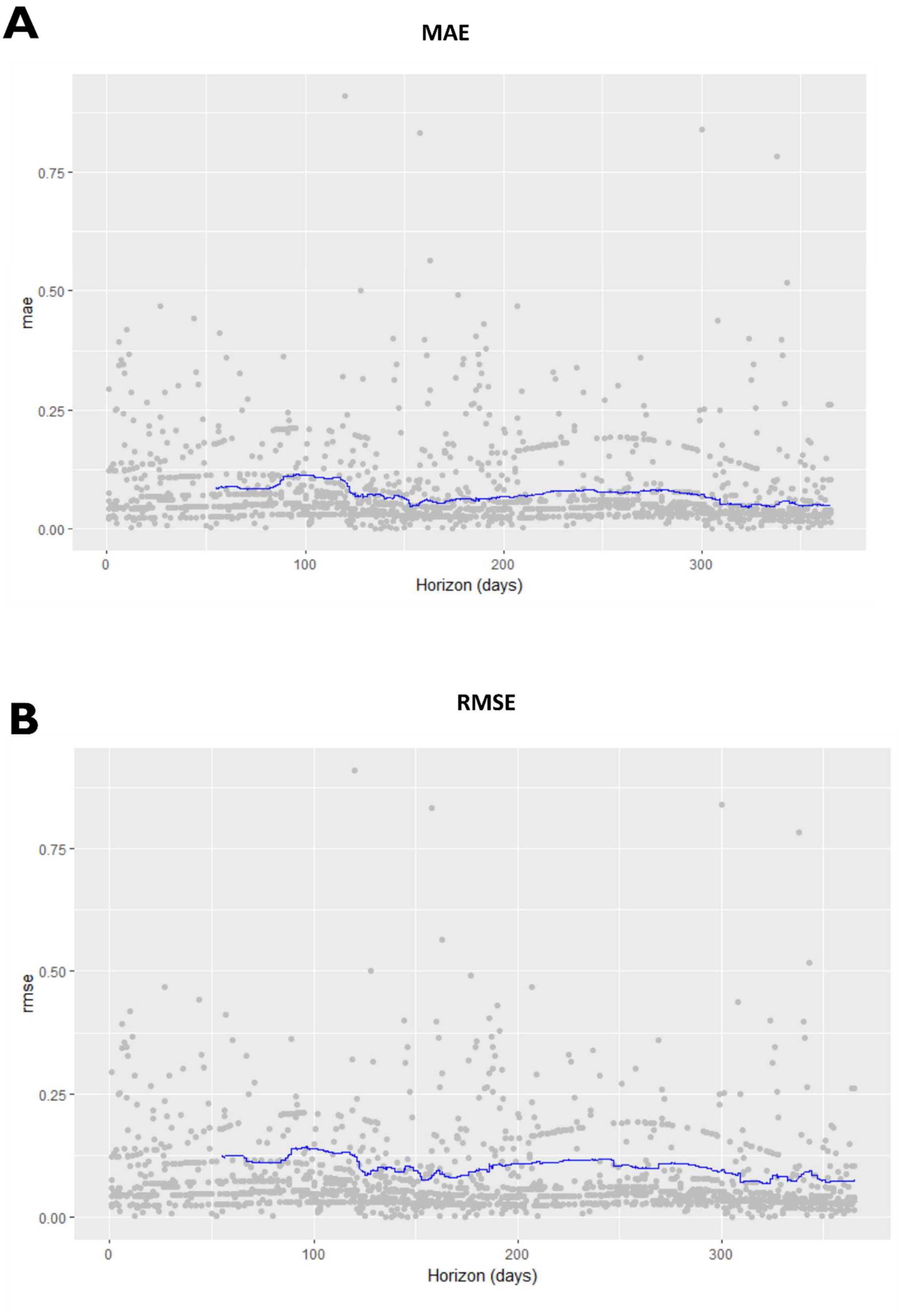


Figure 4.13 Results from cross validation of additive regression forecasting model *showing A) Mean Absolute Error (MAE) and B) Root Mean Squared Error (RMSE) using a horizon window of 365 days, training dataset of first 730 days.*

4.4 Discussion

Quantifying reproduction numbers and their spatio-temporal variation can provide useful information to inform strategies to achieve and maintain elimination in contexts where traditional measures of transmission intensity are not appropriate. I used individual level surveillance data to infer reproduction numbers by estimating the likelihood of cases being linked by transmission and applied this to a dataset of all confirmed and probable cases of *P. vivax* and *P. falciparum* occurring in Yunnan province between 2011 and 2016, which is a focus of concern for re-emergence. My results suggest that transmission in this province decreased rapidly between 2011 and 2016 as shown by a declining risk of R_c exceeding zero across the province. This decline is relatively robust to assumptions about the serial interval distribution. Extrapolating this trend using time-series methods, I expect this trend to continue, predicting a mean R_c of 0.005 up to 2020.

Given the consistently very low R_c values estimated by 2014 onwards, and the future projections based on observed reproduction numbers over time, the results suggest that re-emergence or outbreaks of sustained transmissions are unlikely, provided interventions are continued. However, as all data analysed was collected whilst the NMEP was in place, I cannot draw conclusions about the impact of scaling back interventions or consider other counterfactuals. There is also some uncertainty in the estimates of current and future R_c, although the 95% credible intervals of these estimates remain below 1. It is important to note that even with low R_c values it is still possible for locally acquired cases to occur following importation, however the probability of sustained chains of transmission decreases as R_c decreases. There also is more uncertainty in the estimates of risk in areas that have not observed many cases. It is difficult to determine whether an absence of cases is due to a lack of detection, a lack of importation events occurring or a low underlying receptivity to transmission.

However, it is worth noting that the greatest uncertainty in spatiotemporal risk estimates of $R_c > 0$ tends to be in areas of high elevation (elevation > 3000m), where there is unlikely to be transmission. Given the large numbers of imported cases, it is important to highlight these uncertainties and ensure control measures are maintained. Nonetheless, my findings are promising for China to meet their 2020 elimination goal. The results presented here highlight the success the country has had in malaria control and highlights the difficulty of elimination certification in contexts where both distant and local cross border importation is common.

The work presented in this chapter attempts to quantify receptivity, or the potential for local transmission to occur following the introduction of a case. It is important to note that while competent vectors are present in a place we would not expect a receptivity of zero (although if case detection and management is fast and effective we could expect a receptivity which is near zero). There are many areas where there have been no importation events and therefore there has been no opportunity to observe resulting local cases, therefore in some areas the geostatistical model predicts a risk of zero or near zero, but the uncertainty associated with this prediction is high. Conversely areas where there has been more importation allow more certainty in risk estimation.

Whilst there is a clear peak in incidence of cases occurring in May (Figure 4.7) , the seasonality of R_c estimates were less clear, although there seemed to be two peaks in seasonal increases in R_c, one occurring in March/April, and one in October. This pattern could be an artefact of human movement, with both periods associated with seasonal movement and holiday periods – the *Chunyun* period occurs in China for Chinese New Year and the holiday week of the National Day in October and is associated with intranational travel to visit family. During this time, there is often movement from cities to rural areas, and so in these contexts there may be more opportunities for infection to occur as more people are exposed to bites from suitable vectors. This is supplemented by the finding that these specific holidays are associated with small to moderate increases in R_c, however it is worth noting the very wide credible intervals and the great deal of uncertainty associated with these estimates, and therefore caution is required when interpreting this finding.

There are several limitations to this study. Firstly, there is a limitation in the classification of local and imported cases used in this study. For instance, the definition of importation used in case classification is defined by travel to any malaria-endemic areas outside China in the month prior to illness onset. This definition might include people who travelled abroad within the week prior to illness onset, but biologically their infection could not have been obtained during that time given the incubation period. However, in the absence of alternative information, travel history may provide a better indication of the likely importation status of a case than attempting to infer importation without this information, however there could be scope in future work to allow for incorrect travel history. As certification of elimination is now tolerant of introduced (first generation imported-to-local transmission) but not indigenous (second generation local-to-local transmission) cases, being able to differentiate between the two, and understanding how much transmission is indigenous versus imported or introduced is an important area of focus for future work.

It is important to consider unobserved cases and their potential contribution to transmission dynamics. I do account for unobserved cases via epsilon edges; however, this method is still more suited to scenarios where most cases are observed. In contexts with a high level of asymptomatic infection contributing to transmission or with poor case detection and/or reporting, these approaches would not be suitable.

For the *P. vivax* data, it is important to note that the approach used in this chapter does not explicitly model relapse or recrudescence, but does allow for relapse as an unobserved source of infection. In addition to finding no duplicated patient identifiers (suggestive of repeated infection or relapse) within in the individual- level electronic database, there are several other features of malaria control in China that also provide evidence that it is unlikely that a large proportion of the observed cases are due to relapse. According to China's National Malaria Elimination Technology Program (2011), the epidemiological history of each case has been investigated to check the source of the infection and the history of previous infection and relapse malaria. In addition, all malaria cases received free antimalarial treatment, and each case of *P. vivax* malaria was treated with radical cure.

Nonetheless, there is a chance of some relapse malaria which was missed during case investigation, treatment and surveillance. In our approach we jointly estimate unobserved sources of infection but are agnostic as to the specific cause of the unobserved source. As a result, relapses are considered as one of the potential unobserved sources of infection. Although large amounts of relapse are unlikely for the reasons outlined above, it is true that if there were very large amounts of relapse, the estimated reproduction numbers could be over estimated. However, given that we find such low reproduction numbers, even if this unlikely situation were the case, this does not impact our key findings and in fact would be stronger evidence of China achieving strong reductions in malaria transmission.

A second limitation is the type of data available for inference. Although not available for this study, there are several data sources that increasingly are being collected and could enhance similar analyses in the future in eliminating and pre-eliminating contexts. Firstly, methods to make use of contact tracing data have been developed for emerging outbreaks (Nagraj *et al.*, 2018) but have not to my knowledge been applied to endemic disease in the elimination. Although contact tracing for indirectly transmitted diseases is more difficult, identifying if the likely source of infection is a breeding site near the home or a place of work is carried out through active case detection schemes, but often the resulting data are not made available alongside line list data. This information could be used to weight certain connections. Genetic data are also increasingly available, and provide useful information about movement of parasites (Chang *et al.*, 2019; Tessema *et al.*, 2019), the likelihood of two cases being linked by transmission, and can provide useful information to help distinguish imported from local cases and chains of transmission resulting from importation from on-going local transmission (Wesolowski *et al.*, 2018). Such data were not available in this context; however, a similar methodological framework or approach could incorporate information such as genetic distance. Historical data on incidence at fine scale (e.g. village level) could also be used to inform likelihood of asymptomatic infection.

The effect of holiday periods had a large amount of uncertainty associated with the estimated effects they had on transmission, and therefore their impact should be interpreted with caution. However, the behavioural changes associated with Chinese New Year could lead to behavioural changes which may impact reporting and treatment seeking rates and therefore bias reporting during these times – due to individuals not being in work, travelling to other parts of the country. Changes in importation during this time due to travel to see family may also bias estimates.

I introduced a new framework for analysing individual level surveillance data and found that in Yunnan province, R_c has seen a notable downward trend since 2011 and is expected to remain low with maintained interventions into 2020. This decline coincides with 1-3-7 strategy in improved adherence to guidelines. I predict a mean R_c of 0.005 for 2019, however even with such low R_c values estimated, there may still be a need to continue to invest in detecting and rapidly responding to imported cases in order to achieve three consecutive years of zero cases and prevent resurgence. Nevertheless, China's elimination strategy and investment in surveillance provides a useful roadmap for other countries planning for malaria elimination by illustrating how coordinated and timely surveillance and response can be implemented, as well as sustained investment in surveillance, and region-focused international collaboration.

5
Incorporation of distance features into the inference framework

5.1 Background and motivation

As discussed in earlier chapters, individual-level disease surveillance data, collected routinely and as part of outbreak response, capture a wealth of information which could improve measurements of transmission and its spatiotemporal variation, in turn informing the design of epidemiological interventions. In many cases, this includes additional forms of information to the primary data inputs used in previous chapters, namely the time of symptom onset and classification of cases as imported or locally acquired. For example, there may be geo-located health facility or residence data, demographic data about the patients such as sex, age and occupation. In some cases, molecular data such as parasite or viral genetic sequences or markers are also available. Robust methods to utilise these different forms of information are required in order best support decision making. However, challenges exist in making use of these diverse data sources and leveraging the information they contain within a single inference framework. Geographic information, in the form of GPS coordinates or address of residence or health facility, is often collected but could be more effectively utilised, especially in combination with other information such as symptom onset time and genetic distance. Furthermore, the relative importance of location in determining observed patterns of infection and transmission risk compared to other factors remains poorly understood for many diseases. It is unclear whether simple models of distance can explain the variation observed and inform the design of effective interventions or whether more complex information and data are required, for example models incorporating realistic models of human movement. In this chapter I extend the algorithm introduced in Chapter 4 to incorporate spatial or similar distance-based information to estimate reproduction numbers and their spatiotemporal variation. This approach is then applied to four malaria line list datasets from national malaria elimination programmes: all confirmed cases recorded between 2010 and early 2016 in El Salvador (used in Chapter 3), all suspected and confirmed cases of *Plasmodium vivax* and *Plasmodium falciparum* malaria recorded between 2011 and 2016 in China (the subset of this from Yunnan province was analysed in Chapter 4), and all confirmed cases between 2010 and 2016 in the Kingdom of Eswatini, formally Swaziland (previously analysed in (Reiner *et al.*, 2015) to explore various assumptions about the relationship between locations of cases and likelihood of transmission occurring between them, as well as the impact of unobserved cases. The approach introduced in this chapter is flexible and provides the potential to incorporate other sources of information which can be converted into distance or adjacency matrices such as travel times or molecular markers.

5.1.1. The importance of location in malaria transmission

The importance of spatial location has long been identified as important in infectious disease transmission, as illustrated by the often-cited example of John Snow's 19th century map of cholera cases in London which identified the Broad Street water pump as the likely source of infection (Snow, 1855; Cameron and Jones, 1983). Diseases often are distributed non-randomly in space, and this distribution is often determined by co-variates which also vary over space such as temperature, land use, vector or human population distributions. Analyses of this variation and associated co-variates can identify disease risk factors, and importantly make predictions about risk of disease occurrence in unobserved localities. Over the past 15 years, there has been increased interest in using geostatistical methods to map malaria due to the development of statistical techniques, suitable computational power, and necessary data to carry out rigorous statistical analysis. Great strides have been made in mapping many aspects of malaria epidemiology including burden and distribution of different Plasmodium species (Dalrymple, Mappin and Gething, 2015), vector distributions (Hay *et al.*, 2010), clinical incidence (Bhatt *et al.*, 2015) and climate/habitat suitability (Gething *et al.*, 2011). Whilst these methods have been powerful in demonstrating changes in malaria transmission over time and the impact of control measures (Bhatt *et al.*, 2015) they require large numbers of cases to estimate values of interest such as prevalence with a reasonable amount of uncertainty. As a result, they cannot be easily applied to elimination settings where case counts are low. Furthermore, in elimination settings malaria transmission is thought to take on more epidemic dynamics (Cotter *et al.*, 2013) ,meaning the importance of time and other highly dynamic factors such as human movement patterns becomes more relevant, therefore space becomes more related to time and how mobile and connected infected individuals are and how far they travel.

Malaria transmission requires a human infected with blood stage parasites to be bitten by a female mosquito, for that mosquito to ingest gametocytes and then for that mosquito to bite a susceptible human and inject sporozoites during the blood meal. Therefore, several spatially relevant processes must be considered, which occur on different scales (Figure 5.1). In the absence of human movement, the flight range of the mosquito vector limits transmission distance. Historical mark-release-recapture studies of *Anopheles albimanus* in El Salvador found the mean dispersal distance of vectors based on mark-release-recapture to be 548m in the dry season (Lowe, 1974) and 942m in the wet season (Lowe *et al.*, 1975), with a maximum dispersal distance of 3km (Lowe, 1974), whilst a more recent study in Belize found recapture of *Anopheles albimanus* at 0m from the release point only. *Anopheles sinensis*, now thought to be the dominant vector species in South-western China (Huang *et al.*, 2015; Zhang *et al.*, 2017), was found to have a range of up to 12km, with 90% of captures occurring within 6km in a study in Korea. Studies within a Chinese city found that 90% of mosquitoes were recaptured within 100m, with a maximum range of 400m (Liu *et al.*, 2012). Blood fed Anopheles gambiae has been found to have a maximum flight distance of 10km, however as this was carried out within a flight chamber it is not clear what wild dispersal distance may be. Based on the information available, it is reasonable to assume that in most contexts the maximum range is 10km and most dispersal likely to be within 1km.

Nonetheless, due to the period of time in which malaria parasites can reside within a human body, human commuting and migratory patterns can allow for the movement of parasites across longer distances (Lynch and Roper, 2011; Wesolowski *et al.*, 2012; Wangdi *et al.*, 2015), and for transmission to occur far from the point of infection if suitable vectors are present. Daily or weekly commutes over shorter ranges introduce the potential for frequent opportunities for parasites to travel between a residence location and a place of work. Seasonal or one-off migration events, driven by economic, environmental, cultural or socio-political forces such as pilgrimages, fleeing violence or instability, or seeking seasonal employment opportunities, e.g. as a logger or agricultural worker also can lead to infections occurring over long distances (Cruz Marques, 1987; Wangdi *et al.*, 2015; Surjadjaja, Surya and Baird, 2016).

5.1.2. Modelling the relationship between Euclidian distance and transmission likelihood

A variety of models and functions have been used to describe the spatial component of infectious disease transmission. The most basic approach discussed in this thesis is to define binary near/far threshold, where all cases within a certain area or distance from each other are considered equally likely to occur (weighted 1) and cases outside this threshold are deemed unlikely (weighted as 0). This may be appropriate when there is an epidemiological reason for distance to be important at a certain threshold. An example of this may be a highly immobile, isolated population where only the pathogen moves, whereby the movement range of pathogen or vector determines the relevant threshold. In a quarantine scenario a binary threshold may also be a valid model by differentiating between cases which occur within a quarantine area versus outside of it. Such approaches are simple however are likely to make oversimplifying assumptions in contexts where there is human movement, which often can occur to and from population centres, resources such as bodies of water or places of work such as agricultural land or mining sites.

Another approach, adopted in this chapter, is to utilise a spatial kernel, also known in ecological studies as a dispersal kernel which is a probability distribution which describes the likelihood of either an infection event or dispersal of an organism or contagion as a function of distance (Lindström *et al.,* 2010). In many contexts empirically estimating the kernel is not possible and so a well-studied probability distribution is used to model the relationship between distance and the likelihood of dispersal/transmission occurring across that distance. The Gaussian kernel has been the traditional kernel of choice to model population spread and has been used to model a wide variety of diffusion processes, including the spread of vector borne livestock diseases (Szmaragd *et al.*, 2009, Gerbier *et al.*, 2008), and is the resulting assumption of a random-walk movement pattern (Turchin, 1998). However, there is evidence that in many contexts, more leptokurtic, or "fat-tailed" distributions are found in outbreak data (Ferguson *et al.,* 2001) and also human mobility patterns (Brockmann *et al*, 2006), which are indicative of higher frequencies of both short-distance and long-distance movements. One such example of a leptokurtic distribution is the exponential distribution.

There are also several potential extensions to the spatial aspect of the model not explored in this chapter due to a lack of required population data to parameterise them, however which could be incorporated into the analytical framework in the same way the current exponential and Gaussian kernels are used. One such approach is to use a gravity model. Gravity models assume human movement follows gravitational "pulls" to population centres, whereby distance to local centres of dense populations are considered as well as Euclidian distance between two points or cases. The most basic gravity model is $T_{ij} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})}$, which describes the number or probability rate of individuals *T*$_{ij}$ moving between locations *i* and *j* per unit time. The Gravity model assumes this is proportional to a power (determined by *a* and *β*) of population sizes of both locations *i* (m_i^{α}) and *j* (n_j^{β}) and decreases with the distance between the populations, *r*$_{ij}$, following a function *f*(*r*$_{ij}$) which can be adapted according to the context and fit to empirical data. However, the gravity model is dependent on these parameters which can be difficult to estimate. In recent years, radiation models have been proposed as an alternative to gravity models to model population flows (Simini *et al.*, 2012). Another model adapted from physics to model human movement, the radiation model, instead models population flow as $T_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}$, where *s*$_{ij}$ the total population in the circle of radius *r*$_{ij}$ centred at *i* (excluding the source and destination population), T_i is the total number of individuals moving from location I and n_j is m_i are the respective populations in location j and i respectively.

Recent work (Marshall *et al.*, 2018) fitted travel data collected from Mali, Burkina Faso, Zambia and Tanzania specifically to elucidate travel patterns relevant to malaria transmission (Marshall *et al.*, 2016) to both Gravity and Radiation models. They found that the radiation model was a better fit for travel to nearby populations, whilst the gravity model was a better fit to the overall data and for travel to large population centres. However other work has suggested that, based on mobile phone data, human movement can be described by a Truncated Power Law (Brockmann, Hufnagel and Geisel, 2006; González, Hidalgo and Barabási, 2008; Meyer and Held, 2014).

5.1.3. Alternative distance measures

Due to the different spatial scales at which malaria transmission and the processes driving it operate, Euclidian distance may not be the most appropriate metric of distance between cases due to human travel patterns and travel times not always being proportional to Euclidian distance, and other measures of distance could be incorporated into the framework introduced in this chapter. Indeed similar approaches to the one presented in this chapter (Wang, Ermon and Hopcroft, 2012) using Rayleigh hazard and survival functions have been developed to incorporate features of tweets posted on twitter such as the language of tweet and the similarity in the wording of a text, in combination with temporal information. In the same way, features such as occupation, sex or other factors which may affect travel patterns to higher risk areas may be incorporated into the framework. This would be particularly important in contexts where *P. knowlesi* is a concern, and proximity to forest/occupation which takes one into a forest could indicate zoonotic transmission, whereas complete absence of time spent near or in a forest would indicate human to human transmission.

One approach is to replace Euclidian distance with a measure of travel time or accessibility between places. Accessibility indices consider movement by looking at transport networks such as roads and calculate a "friction surface" which estimates the difficulty and time required to go from point A to point B (Weiss *et al.*, 2018). If available, travel or mobility data could also be used either to parameterise a spatial kernel or radiation/gravity model or used on an individual basis to weight likely transmission events.

Increasingly genetic and molecular data are being collected as part of disease surveillance and outbreak response. Increasing interest in using for endemic diseases nearing elimination such as malaria (Wesolowski *et al.*, 2018). There is evidence that genetic data can provide signals of movement of parasites between populations (Chang *et al.*, 2019; Tessema *et al.*, 2019a). There could be scope to use measures of genetic distance, particularly in contexts where the population genetics of malaria is not complicated by cases being infected with multiple clones, and therefore identity by descent (Taylor, 2015) could be used as a metric of distance.

5.1.4. Aims and approach

In this chapter, I introduce a flexible framework which could extend the approaches introduced in chapter 4 to incorporate a range of approaches to modelling distance and transmission likelihood, as well as non-Euclidian distance metrics, whilst continuing to allow for unobserved sources of infection and incorporate estimates of uncertainty through prior distribution definitions. Then as proof of concept, I apply versions of this framework using Gaussian and Exponential kernels to four datasets from malaria elimination and near elimination contexts, as well as carrying out a detailed sensitivity analysis to explore the impact of varying assumptions about both the relationship between Euclidian distance and transmission as well as the likelihood of a case having been infected by an unobserved source of infection.

5.2 Methods

5.2.1 Data

5.2.1.1 The Kingdom of Eswatini

This dataset, analysed in by Reiner and colleagues (Reiner *et al.*, 2015) captures malaria cases recorded by the national malaria elimination programme between January 2010 and June 2014. For each case detected during this time (N= 1373), case investigation was carried out. For each case the following were collected: GPS coordinates of household location, demographic information (age, occupation and sex), use of malaria prevention interventions such as long-lasting insecticide treated bednets (LLINs), and date of symptom onset, diagnosis and treatment, as well as travel history. Based on travel history cases were defined as locally acquired, imported. For a small number of cases (N=58) the local/imported status was determined “unknown”. For the purposes of this analysis, these cases were treated the same as local cases, i.e. they were assumed to have potentially been infected by other cases in the dataset and/or been infectors themselves.

5.2.1.2 China

This dataset consists of individual-level case data for all confirmed and probable cases reported in China between 2011 and 2016 (Table 5.1 and Table 5.2). The data consist of an individual identifier, date of symptom onset, date of diagnosis and date of treatment, as well as the geolocated address of residence and health facility. If the suspected location of infection was in China and not in the same district, then the presumed location of infection was also included in the dataset. Demographic information such as age and sex were also collected. A subset of this dataset, focussing on Yunnan province, is analysed in Chapter 4. For the analysis the data were separated into *P. falciparum* and *P. vivax*. *P. malariae* (N=252) and *P. ovale* (N=822) were reported but excluded from the analysis due to the lower public health concern of these species. Untyped cases (N= 398)

Table 5.1: Cases by diagnosis type (probable and confirmed) and species across China

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Confirmed	260	11830	252	822	6631	87
Probable	0	176	0	0	693	311

Table 5.2: Cases by imported/local status and species across China

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Local	5	92	4	1	1711	95
Imported	255	11914	248	821	5613	303

5.2.1.3 El Salvador

This dataset is analysed and described in Chapter 3. Briefly, the data consist of all confirmed cases of malaria between 2010 and the first two months of 2016 (N= 91 cases, of which 30 imported, 6 *P. falciparum*, 85 *P. vivax*). For each case, the date of symptom onset was recorded. Residential address was available for all but two cases. For these cases the location was available at the *municipio,* or municipality level, and the coordinates of the centroid of the municipality (which for both were cities) were used as the geo-location. Two cases had addresses listed outside of El Salvador, both of which were located in Guatemala. All cases within El Salvador with full addresses (N=85) were georeferenced by latitude and longitude to *caserío/lotificación* level, which is approximately neighbourhood or hamlet level.

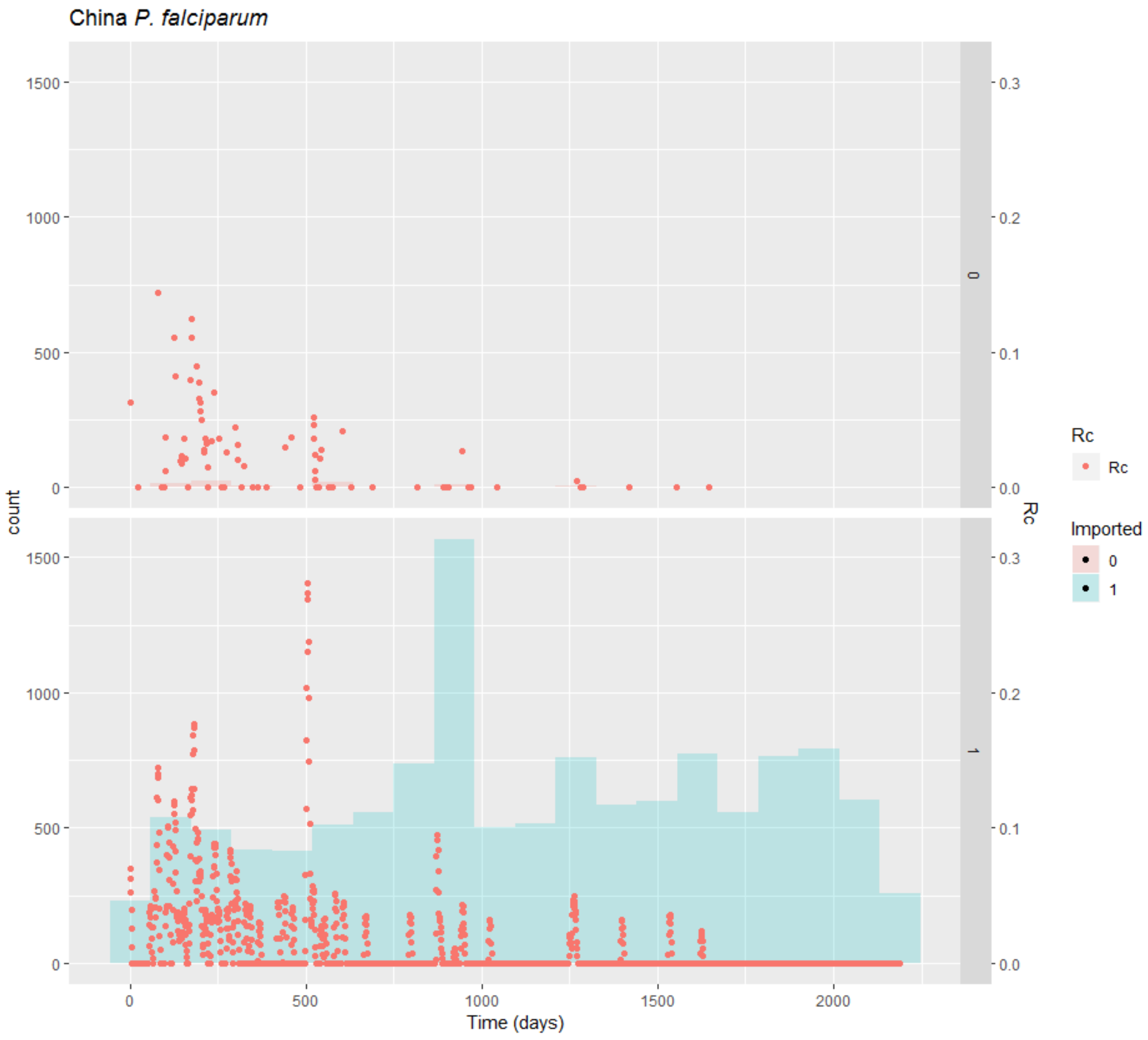


Figure 5.1: Temporal patterns of incidence and reproduction number estimates for P. falciparum in China

Plot showing the relationship between estimated R_c (red points) and incidence (shaded histogram) over time for both imported (lower row, blue, imported =1) and local cases (upper row, red, Imported =0) for P. falciparum in China

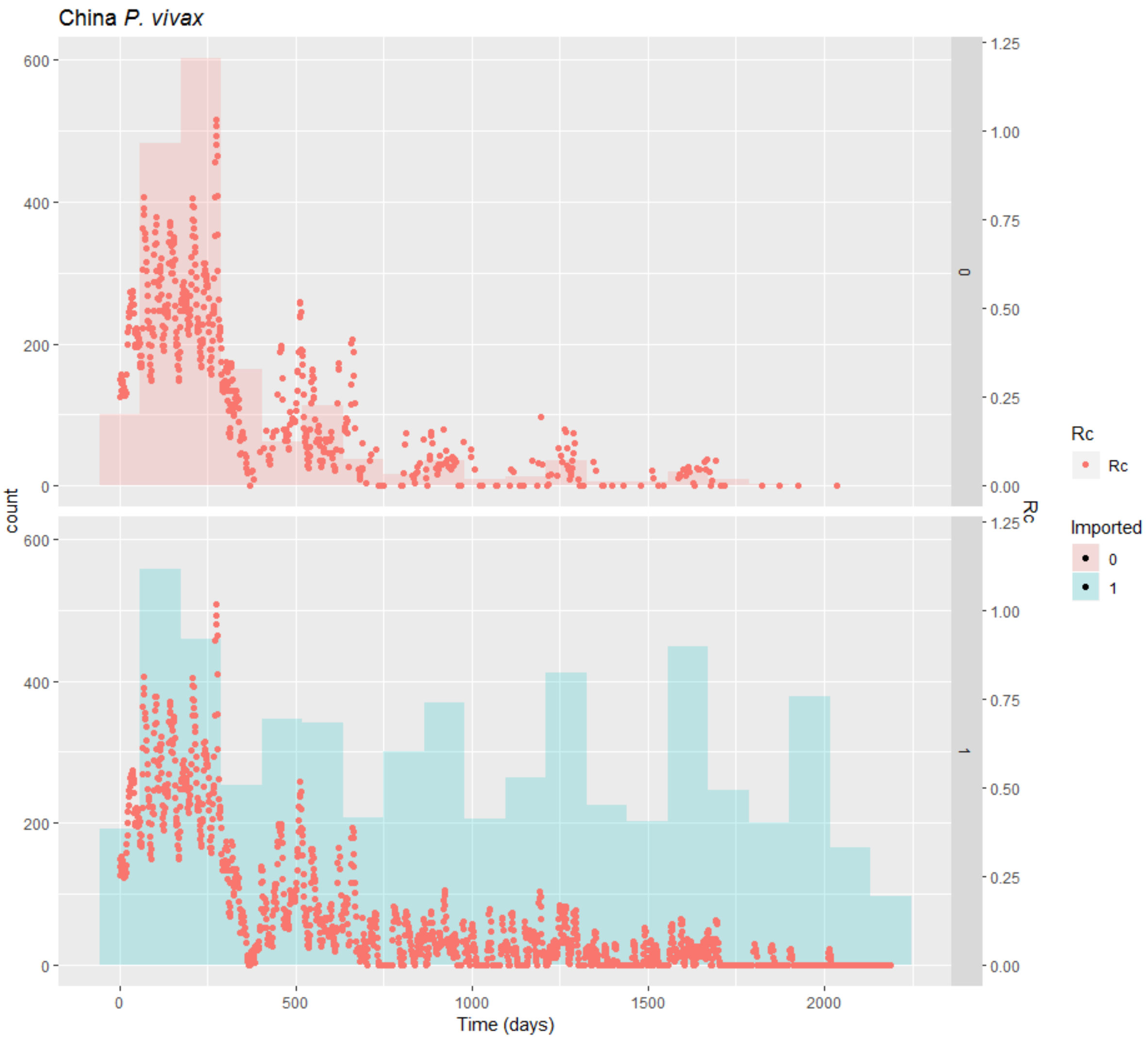


Figure 5.2: Temporal patterns of incidence and reproduction number estimates for P. vivax in China

Plot showing the relationship between estimated R_c (red points) and incidence (shaded histogram) over time for both imported (lower row, blue, imported =1) and local cases (upper row, red, Imported =0) for P. vivax in China

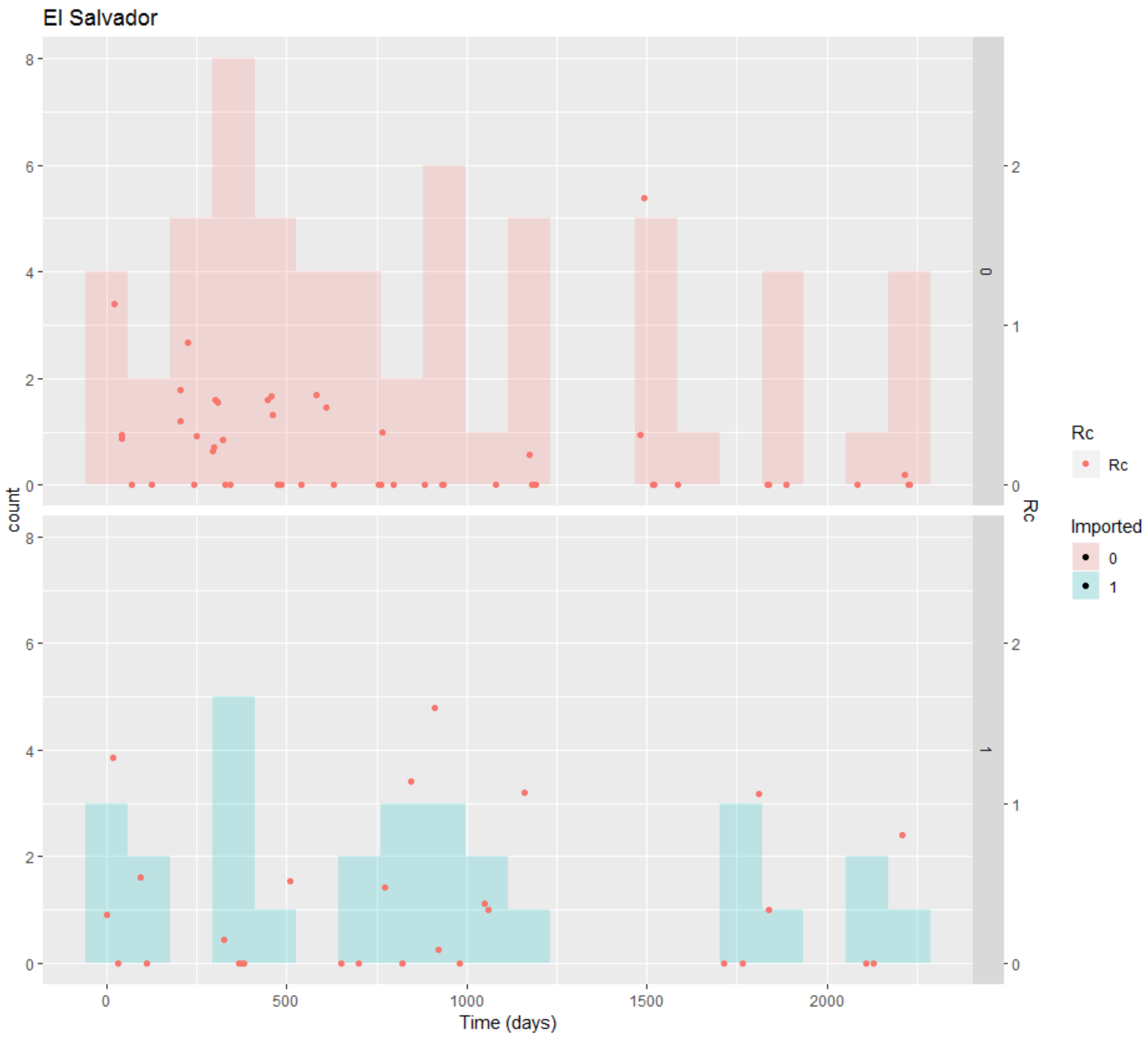


Figure 5.3: Temporal patterns of incidence and reproduction number estimates for El Salvador

Plot showing the relationship between estimated R_c (red points) and incidence (shaded histogram) over time for both imported (lower row, blue, imported =1) and local cases (upper row, red, Imported =0) for P. vivax in El Salvador

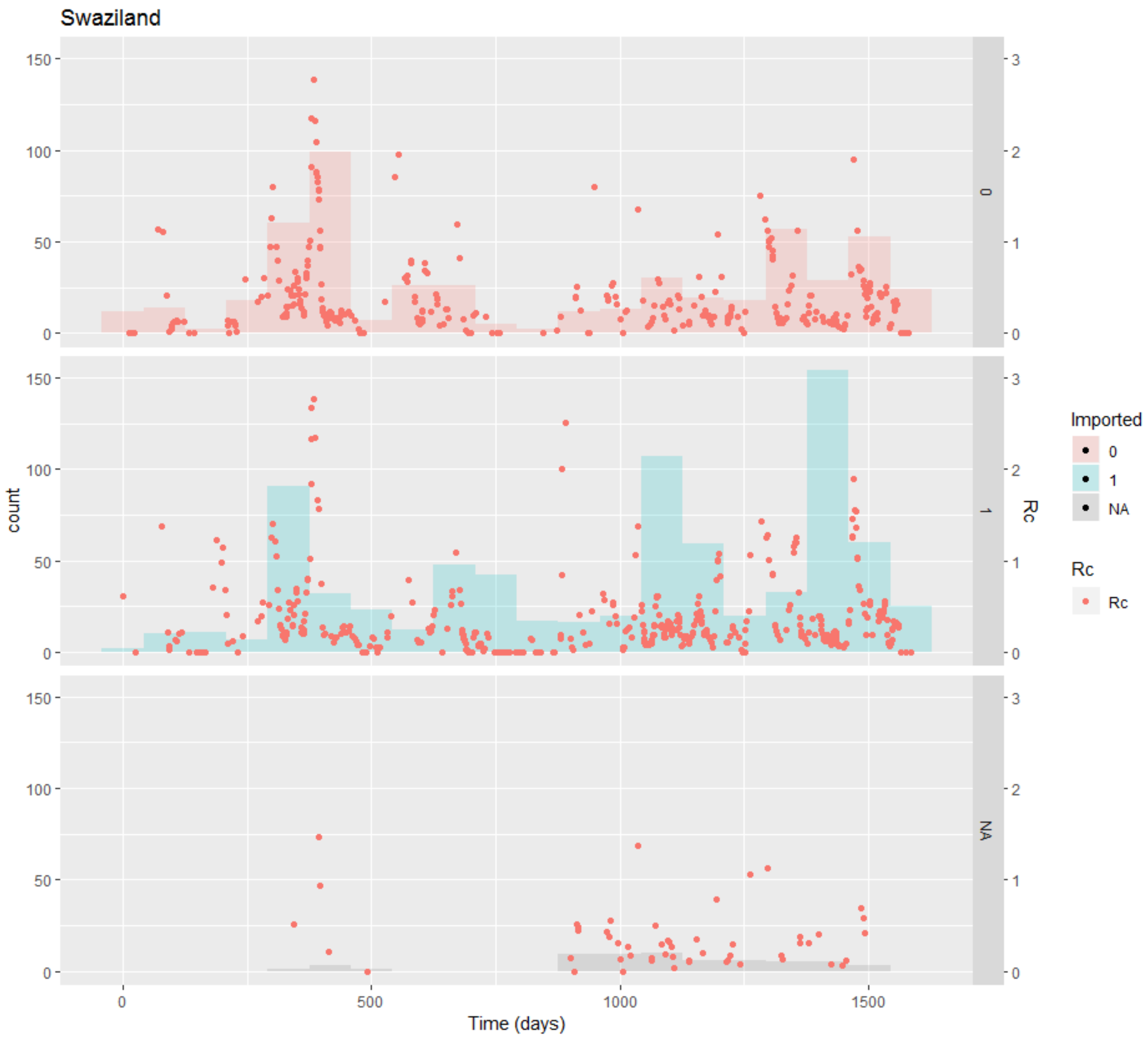


Figure 5.4: Temporal patterns of incidence and reproduction number estimates for Swaziland

Plot showing the relationship between estimated R_c *(red points) and incidence (shaded histogram) over time for imported (middle row, blue, Imported =1) and local cases (upper row, red, Imported =0) and "unknown" importation status for P. falciparum in Swaziland*

5.2.2 Transmission model specifics

In order to incorporate features other than time, I extended the algorithm applied to Yunnan Province in Chapter 4 by introducing a second function, f_2, which describes the relationship between space (or distance of any kind) and likelihood of transmission. An appropriate function such as a Gaussian kernel is defined and the parameter(s) shaping that distribution, β, are either fixed, or given a prior distribution and estimated from the data. Multiplied, together, this returns a single function:

$$f\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) = f_1(t_i|t_j;\ \alpha_{i,j}) \text{ x } f_2\left(x_i \middle| x_j; \beta\right) \ (1)$$

Determined by times t, spatial locations x, transmission rates $\alpha,$ spatial parameter(s) $\beta.$

As before, the hazard is defined as the pairwise likelihood divided by the survival term:

$$H = \frac{f\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right)}{s\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right)} \ (2)$$

To derive the survival function, one integrates across all distances and times as follows:

$$S\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) = (\int_0^\infty \int_0^{t_i - t_j} f_1\left(t_i \middle| t_j; \alpha_{j,i}\right)\ f_2\left(x_i \middle| x_j; \beta\right) dt\ dx \ (3)$$

The specific functions used in $f_1(t_i|t_j;\ \alpha_{i,j})$ and $f_2(x_i|x_j; \beta)$ will have large impacts on the outcomes of results and therefore the assumptions inherent in these choices must be made explicit and linked to the mechanisms of transmission.

To illustrate this approach by applying to several malaria line-lists, I will use a shifted Rayleigh distribution to model serial interval distributions, $f_1(t_i|t_j;\ \alpha_{i,j})$, as used in Chapters 3 and 4. For the second part of the likelihood which model the relationship between space and the likelihood of transmission $f_2(x_i|x_j; \beta)$, Gaussian and Exponential diffusion kernels were used (**Table 5.3**).

Table 5.3: Equations for f1, f2, hazard and survival for time -only, Gaussian and Exponential spatial kernels

	$\boldsymbol{f_1(t_i \vert t_j;\ \alpha_{i,j})}$	$\boldsymbol{f_2(x_i \vert x_j; \beta)}$	**Hazard**	**Survival**
Exponential	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}$	$e^{-\beta(x_i - x_j)}$	$\beta\alpha(t_i - t_j - \gamma)e^{-\beta(x_i - x_j)}$	$e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}\frac{1}{\beta}$
Gaussian	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}$	$e^{-\beta(x_i - x_j)^2}$	$\frac{2\sqrt{\beta}\alpha(t_i - t_j - \gamma)e^{-\beta(x_i - x}}{\sqrt{\pi}}$	$e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}\frac{\sqrt{\pi}}{2\sqrt{\beta}}$
Time only	$\alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}$	n/a	$\alpha(t_i - t_j - \gamma)$	$e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}$

Using a shifted Rayleigh distribution as before in Chapter 4 and an exponential kernel the pairwise likelihood of a case showing symptoms at t_i and at residence location x_i being infected by a case showing symptoms at time t_j and at residence location x_j, becomes

$$f(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \alpha(t_i - t_j - \gamma)e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}e^{-\beta(x_i - x_j)} \quad (4)$$

As shown in Chapter 2, the survival term simplifies to:

$$\mathrm{S}(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)}\frac{1}{\beta} \quad (5)$$

And the hazard simplifies to

$$H(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \beta\alpha(t_i - t_j - \gamma)e^{-\beta(x_i - x_j)} \quad (6)$$

For the Gaussian function, the pairwise likelihood of a case showing symptoms at t_i and at residence location x_i being infected by a case showing symptoms at time t_j and at residence location x_j is

$$f(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \alpha(t_i - t_j - \gamma) e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} e^{-\beta(x_i - x_j)^2} \quad (7)$$

The survival term is again determined by integrating the likelihood over all potential infection times and all distances

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = (\int_0^{\infty} \int_0^{t_i - t_j} \alpha(t_i - t_j - \gamma) e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} e^{-\beta(x_i - x_j)^2} \; dt \, dx \quad (8)$$

Integrating over time returns

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} \int_0^{\infty} e^{-\beta(x_i - x_j)^2} \; dx \quad (9)$$

Integrating over all distances gives

$$S(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} \frac{\sqrt{\pi}}{2\sqrt{\beta}} \quad (10)$$

Following equation 10, the hazard is equivalent to

$$H(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \frac{\alpha(t_i - t_j - \gamma) e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} e^{-\beta(x_i - x_j)^2}}{e^{-\frac{1}{2}\alpha(t_i - t_j - \gamma)} \frac{\sqrt{\pi}}{2\sqrt{\beta}}} \quad (11)$$

Which simplifies to

$$H(x_i, t_i | x_j, t_j; \alpha_{i,j}, \beta) = \frac{2\sqrt{\beta}\alpha(t_i - t_j - \gamma) e^{-\beta(x_i - x_j)^2}}{\sqrt{\pi}} \quad (12)$$

5.2.3 Modelling missing cases using ε edges

The vast majority of disease surveillance and outbreak response datasets will not be able to capture all cases due to asymptomatic infection, underreporting and movement of people in/out of the surveillance area. Therefore, it is important to consider the impact of missing information on results and identify potential missing sources of infection. In the work described in this chapter, as in chapter 2, we use Epsilon edges, ϵ_i , to identify potential sources of infection. Here, each hazard is estimated as a further competing edge of transmission from an unobserved source, $H_0(\epsilon_i)$. Depending on assumptions for the likelihood and extent of unobserved infection sources, the epsilon edge value can be set to a high or low value. When high, we assume high amounts of unobserved infection and unless two cases have a very high likelihood of being linked, we assume the case was from an unobserved source. When low, we assume little missing data and so cases are only linked to an outside source if they are very unlikely to be linked to an observed candidate infector.

Adding epsilon as a competing hazard and survival returns

$$f(\boldsymbol{t}, \boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\beta}) = \prod_{t_i \in \boldsymbol{t}} S_0(\epsilon_i) \prod_{I_k:t_k<t_i} S\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right) \Big(H_0(\epsilon_i) + \sum_{I_j:t_j<t_i} H\left(x_i, t_i \middle| x_j, t_j; \alpha_{i,j}, \beta\right)\Big) \quad (13)$$

The objective function is then

$$minimize_{\boldsymbol{\alpha},\boldsymbol{\epsilon}} - \log f(\boldsymbol{t}, \boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\beta}) \qquad subject\ to\ \boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\beta} > 0 \ \ \forall i, j \ (14)$$

Because this was carried out within a Bayesian framework the log posterior was maximised to obtain the maximum-a-posteriori estimates.

5.2.4 Implementation of algorithm

The algorithm was written in TensorFlow, implemented in R via the *rTensorflow* package. As in Chapter 4, a prior probability was defined for the parameter shaping the serial interval of malaria, informed by previous characterisations of the serial interval of malaria (Huber *et al.*, 2016) . Because clear data about how likely the cases in each context explored here were to have moved long distances or the likelihood a case has been infected by an unobserved source of infection were not available, several different parameterisations of the model were used to represent different scenarios (Table 5.4**)** , and a detailed sensitivity analysis was carried out (Section 5.2.6 and **Table 5.5**). The versions of the model which are described in Figures represent different patterns of human/parasite movement, ranging from a context where there may be small amounts of movement (almost all under 10km) to moderate amounts of movement/travel(almost all under 50km) to a less restrictive parameterisation, where near cases were more likely but far away cases were not completely excluded.

These datasets to different versions of the algorithm, as well as temporal-only algorithm described in Chapter 2 and applied in Chapter 4, to explore the impact of different assumptions about the impact of space on estimated R_c values and their variation over time and space. We also evaluate the performance of each approach by comparing differences in the second order AIC (ΔAICc), and the corresponding Akaike Weights.

5.2.5 Comparison of including spatial information for each dataset

Twelve scenarios (Table 5.4) were considered when defining parameters for each dataset. These scenarios consider three different levels of likelihood of transmission in relationship to Euclidian distance (due to the limited range of mosquito travel, this is considered in the context of human mobility), which was defined for both exponential and Gaussian kernels. These are illustrated in Figure 5.5 - Figure 5.7. Then the values for epsilon were set at 0.001 and 0.1, representing different levels of missing cases likely. This can be interpreted as the chance of a case having an unobserved source of infection. For example, 0.1 would represent P(unobserved source of infection) = 0.1. The results of simulations carried out in Chapter 2 demonstrated that setting correct, informative priors for epsilon returned accurate mean R_c values, whereas uninformative priors for epsilon returned slightly underestimated values of R_c when not all cases were observed. Varying standard deviations of the prior were also considered: 1 and 0.001, representing high and low confidence in missingness (informative and uninformative priors). Using the parameter definitions for each scenario for both Exponential and Gaussian Kernels

The timeseries of R and its spatial patterns were illustrated for each dataset and parameter combination and compared to the results of the time-only version of the algorithm (Algorithm 2, implemented in Chapter 4). The results were also mapped to compare how spatial patterns in R_c were affected by assumptions about space and unobserved infections.

In order to compare models quantifiably, the second order Akaike Information Criterion (AICc) was calculated using the equation $AIC_c = -2\log f(x) + 2K(\frac{n}{n-k-1})$, where $f(x)$ is the model likelihood, K is the number of parameters estimated and n is the sample size of the data used to fit the parameters. The AIC(Akaike, 1974) is used in model comparison, by creating a comparison of negative log likelihood that penalises increases in model parameters, to prevent overfitting. AICc is recommended for use with smaller datasets with larger numbers of parameters, and as the sample size n increases AICc converges to AIC(Hurvich and Tsai, 1989). The differences in AICc for each model, known as ΔAICc, were calculated to compare models. Typically, a ΔAICc of greater than 10 is considered strong evidence that that model performs worse than the model it is being compared to.

In addition, Akaike Weights were calculated, which are a measure of the relative likelihood of a model compared to the others considered. Akaike weights are determined by taking the normalised relative likelihood of a model which is $exp(-0.5 \ * \ \Delta AICc \, score \ )$, and then dividing by the sum of these values across all models to obtain a normalised result.

Table 5.4: Table illustrating the different scenarios and corresponding parameter values tested in scenario analysis

Scenario Description	Scenario	Beta (fixed)	Epsilon (prior)

Human movement unlikely, most movement under 10km
Missing cases more likely (but very uncertain)
	1	Gaussian = 0.005
		Exponential =0.1
			Mean = 0.1
			SD = 1

Human movement unlikely, most movement under 10km
Missing cases more likely (confident)
	2	Gaussian = 0.005
		Exponential =0.1
			Mean = 0.1
			SD = 0.001

Human movement unlikely, most movement under 10km
Missing cases less likely (but very uncertain)
	3	Gaussian = 0.005
		Exponential =0.1
			Mean = 0.001
			SD = 1

Human movement unlikely, most movement under 10km
Missing cases less likely (confident)
	4	Gaussian = 0.005
		Exponential =0.1
			Mean =0.001
			SD =0.001

Moderate human movement, most movement under 50km
Missing cases more likely (but very uncertain)
	5	Gaussian = 0.001
		Exponential =0.02
			Mean = 0.1
			SD = 1

Moderate human movement, most movement under 50km
Missing cases more likely (confident)
	6	Gaussian = 0.001
		Exponential =0.02
			Mean = 0.1
			SD = 0.001

Moderate human movement, most movement under 50km
Missing cases less likely (but very uncertain)
	7	Gaussian = 0.001
		Exponential =0.02
			Mean = 0.001
			SD = 1

Moderate human movement, most movement under 50km
Missing cases less likely (confident)
	8	Gaussian = 0.001
		Exponential =0.02
			Mean =0.001
			SD =0.001

Longer range human movement likely
Missing cases more likely (but very uncertain)
	9	Gaussian = 0.0001
		Exponential =0.01
			Mean = 0.1
			SD = 1

Longer range human movement likely
Missing cases more likely (confident)
	10	Gaussian = 0.0001
		Exponential =0.01
			Mean = 0.1
			SD = 0.001

Longer range human movement likely
Missing cases less likely (but very uncertain)
	11	Gaussian = 0.0001
		Exponential =0.01
			Mean = 0.001
			SD = 1

Longer range human movement likely
Missing cases less likely (certain)
	12	Gaussian = 0.0001
		Exponential =0.01
			Mean =0.001
			SD =0.001

5.2.6 Sensitivity analysis and comparison of prior choice on estimated results

In the scenario analysis above the distance shaping parameter is fixed. However due to the uncertainties in the relationship between distance and likelihood of transmission, in many contexts it may be useful to estimate β. To explore the relationship between the estimated epsilon edges, ϵ, and estimated shaping parameter, β, for the distance function. a detailed sensitivity analysis was carried out to explore the impact of a) prior choice for ϵ d) prior choice for β on both the maxmum-a-posteriori estimates for β and the estimated mean R_c.

To consider the effect of varying parameter values and explore their interactions, a range of distance and epsilon edge priors were considered. A truncated normal prior was used for both parameters, and the mean and standard deviation were varied. For ϵ the mean was varied between 1e-10 and 0.5, and the standard deviation was varied between 0.0001 and 0.1. For β, the mean for a Gaussian Kernel was varied between 0.00001 and 0.01 and for an exponential kernel the means considered ranged between 0.0001 and 0.1. For both the standard deviations varied between 0.0001 and 0.1 (**Table 5.5**). Every possible combination of the parameters were run for each dataset and both Gaussian and exponential spatial kernels, giving a total of 2400 parameter combinations tested per kernel, per dataset.

Table 5.5 Different parameters considered in sensitivity analysis

ϵ **mean**	ϵ **SD**	β **mean (Gaussian)**	β **mean (Exponential)**	β **SD**
1e-10	0.0001	0.00001	0.0001	0.0001
1e-5	0.001	0.0001	0.001	0.001
1e-3	0.01	0.001	0.01	0.01
1e-2	0.05	0.01	0.1	0.05
1e-1	0.1			0.1
0.5				

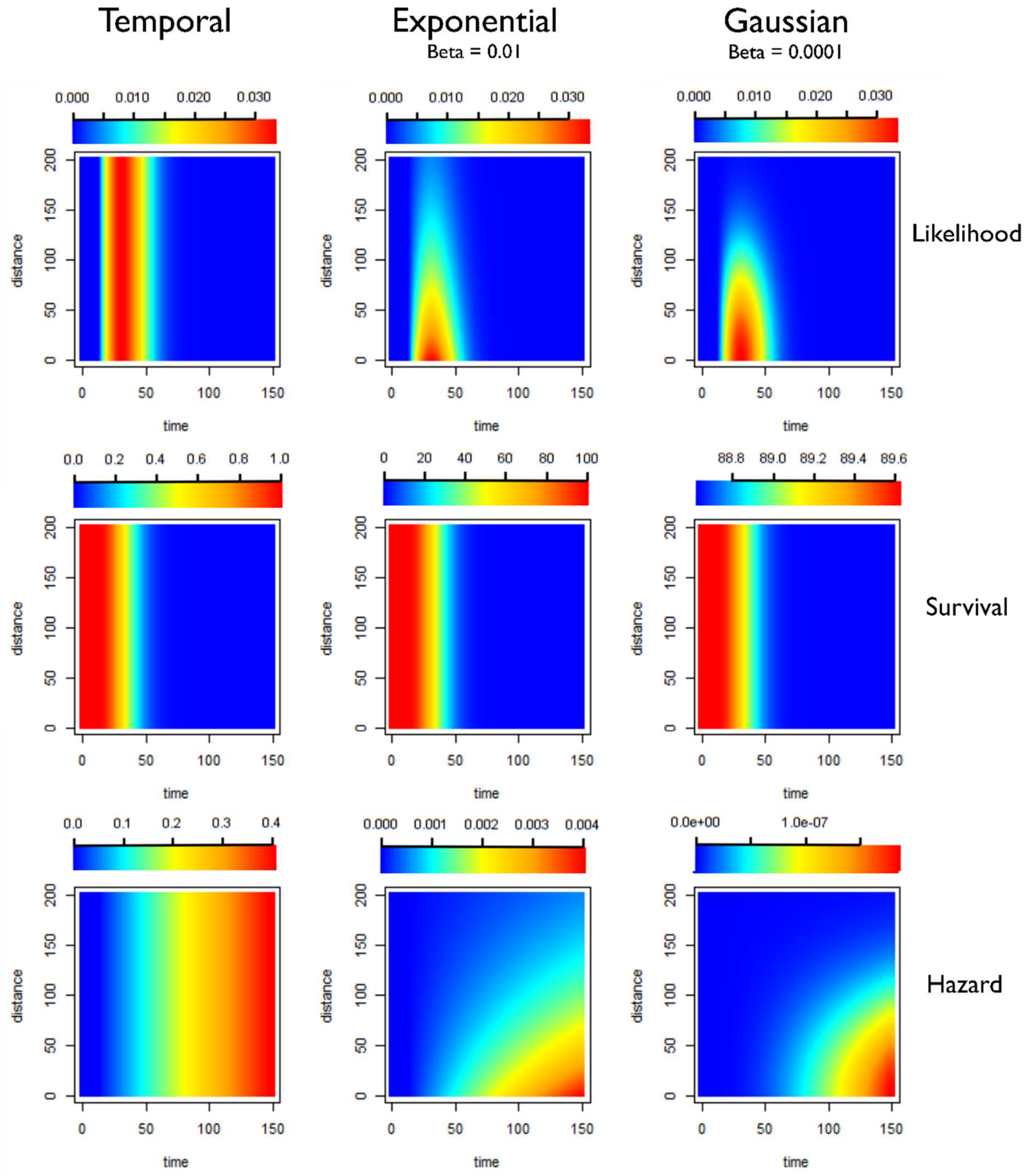


Figure 5.5 : Illustration of likelihoods, hazards and survivals for less restrictive kernels (longer range human movement likely). *Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example less restrictive values for beta, the shaping parameter for the distance kernels have been chosen, representing a context where there is more long-range movement of parasites.*

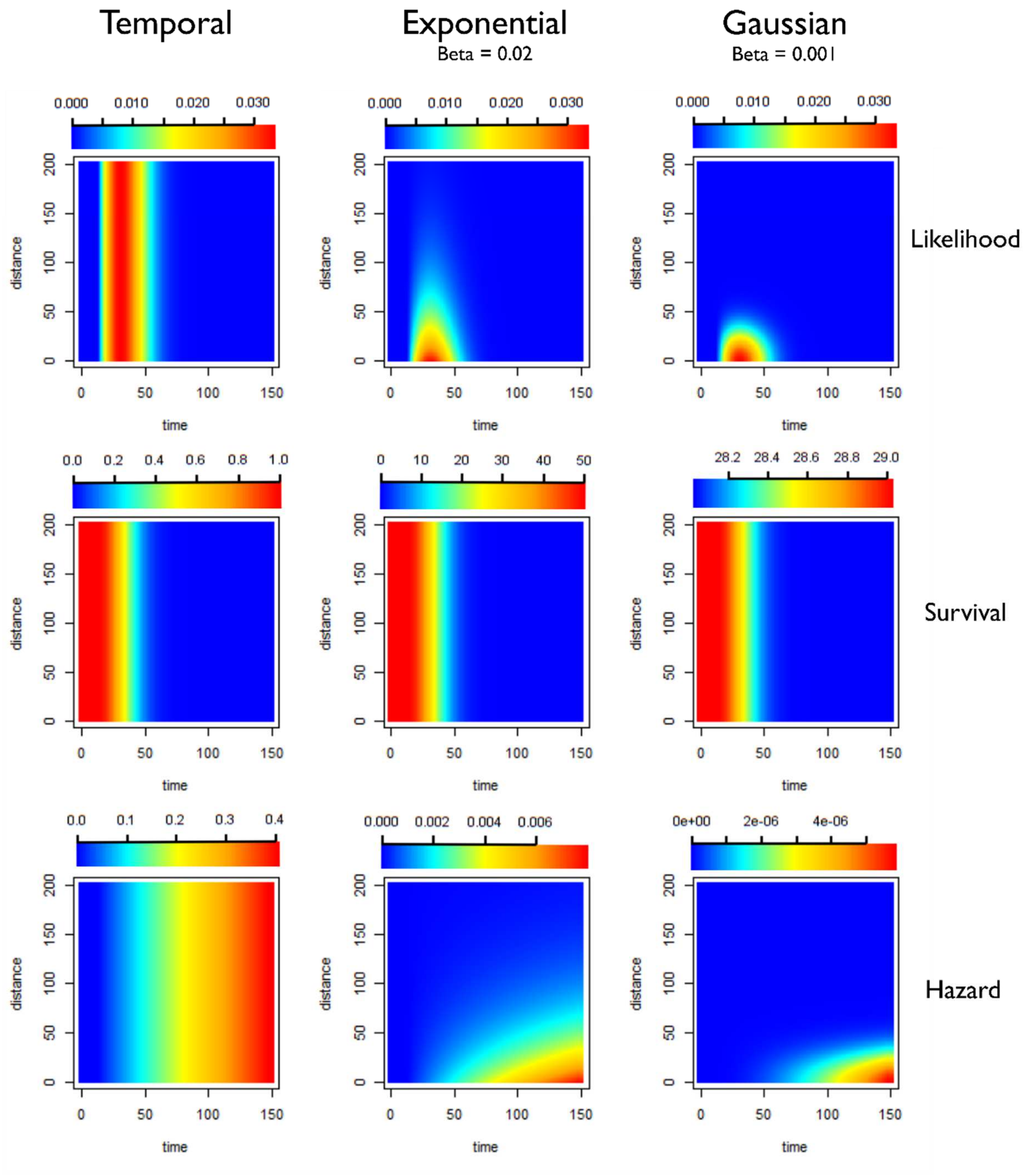


Figure 5.6: Illustration of likelihoods, hazards and survivals for moderately restrictive kernels (moderate human movement, most movement under 50km). *Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example values for beta, the shaping parameter for the distance kernels have been chosen to represent a context where there is more some movement of parasites, but where little movement is expected beyond 50-75km. The likelihood for the Gaussian Kernel is more concentrated, which could represent shorter range movement e.g. commutes, whereas the Exponential has a longer tail so could represent a mixture of short and longer range parasite movement.*

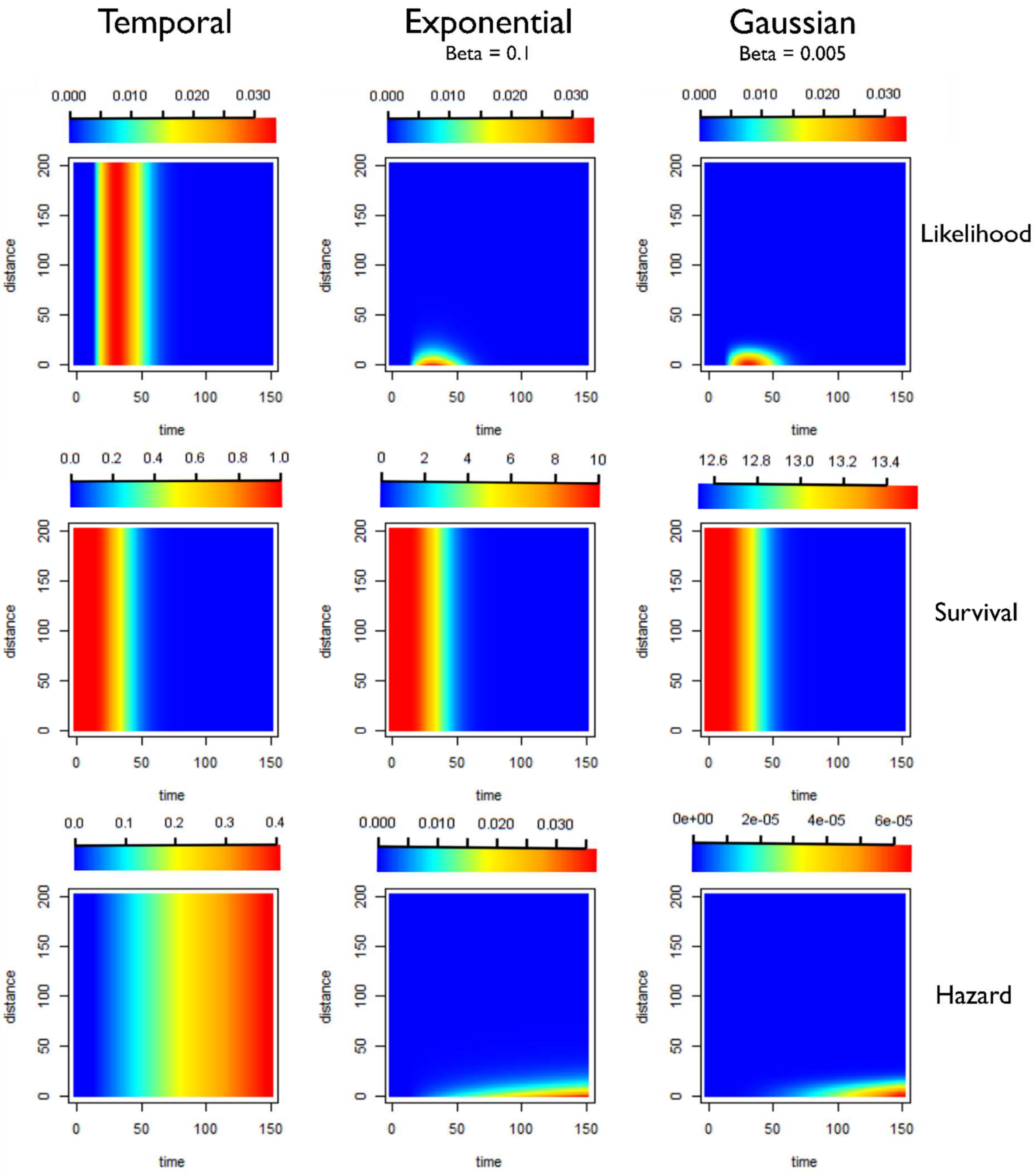


Figure 5.7: Illustration of likelihoods, hazards and survivals for highly restrictive kernels (Human movement unlikely, most movement under 10km). *Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example more restrictive values for beta, the shaping parameter for the distance kernels have been chosen, representing a context where there is very little movement of parasites, with very little movement beyond 10-20km expected.*

5.3 Results

The work presented in this chapter aimed to develop a framework to integrate distance information into the inference framework developed in Chapter 4 and then test the impact of varying assumptions about the relationship between location of cases and the likelihood of transmission as well as the impact of unobserved infection as modelled by epsilon edges, considering twelve scenarios, and applying them to four line-list datasets using two different spatial kernels.

5.3.1 Results of model comparison by ΔAICc across different scenarios

When ΔAICc scores were used to compare model results, all models which included distance had lower (and therefore better) ΔAICc scores than models which only included only time (**Table 5.6**). In addition, exponential kernels consistently outperformed equivalent scenarios using Gaussian kernels (**Table 5.6**). Two scenarios consistently returned the best ΔAICc results, namely Scenario 9 (El Salvador and Swaziland) and Scenario 11 (China, *P. vivax* and *P. falciparum*). Both scenarios assume longer range human movement likely and impose a smaller penalty on cases occurring larger distances. These scenarios also allow variation in epsilon edge values and use a very weakly informative prior on Epsilon edges, but with a different mean (0.1 for Scenario 9, 0.001 for Scenario 11). These results also return smaller mean R_c results than time-only versions of the model (Figure 5.8 – 5.12)

Table 5.6: Summary of ΔAICc results

Dataset	Best Model(s), by ΔAICc	Akaike Weight
Swaziland (Eswatini)	Scenario 9, Exponential	1
El Salvador	Scenario 9, Exponential	0.621540909785805
	Scenario 11, Exponential	0.37845909
China *P. vivax*	Scenario 11, Exponential	1
China *P. falciparum*	Scenario 11, Exponential	1

5.3.2 R_c estimates under different scenarios

Across all datasets, large differences in R_c estimates were found depending on both ε and β parameters. When β is higher, the assumption is that there is little movement of parasites within the country and therefore cases with residential addresses which are far away are unlikely to have infected each other. When this is the case and we assume there are unobserved sources of infection (either through a strongly informative prior on ε with mean 0.1, or an uninformative prior with a lower mean), then R_c values are very low. However if we assume there are little or no unobserved sources of infection, but continue to make restrictive assumptions about space, then most R_c very low but in the localities where there are cases we estimate much higher R_c values as there are no other possible infectors within a reasonable time and/or spatial area. This is illustrated in Figures 5.9 - 5.12.

When looking at the spatial patterns of R_c estimates under different scenarios several trends are seen across all datasets. Scenario 4 is particularly interesting to note because this scenario considers the most restrictive assumptions, both about space and unobserved sources of infection. Across datasets, Scenario 4 results in increased focality and higher R_cs within these foci, but in comparison lower R_cs in other areas. All of the best scenarios as measured by ΔAICc resulted in small R_c estimates, but where comparably larger R_c estimates were estimated, they were in localities identified as foci.

Table 5.7 : Full results of ΔAICc and Akaike Weights for each scenario, dataset and spatial kernel considered

	Dataset	Scenario	Kernel	ΔAICc	Akaike Weight
1	Swaziland	Time	Exp	7386293	0
2	Swaziland	1	Exp	3692506	0
3	Swaziland	2	Exp	3694136	0
4	Swaziland	3	Exp	3692532	0
5	Swaziland	4	Exp	3702805	0
6	Swaziland	5	Exp	1111429	0
7	Swaziland	6	Exp	1111429	0
8	Swaziland	7	Exp	1111432	0
9	Swaziland	8	Exp	1121581	0
10	Swaziland	9	Exp	0	1
11	Swaziland	10	Exp	1600	0
12	Swaziland	11	Exp	59.5	1.20E-13
13	Swaziland	12	Exp	10319	0
14	Swaziland	Time	Gauss	7386293	0
15	Swaziland	1	Gauss	3330479	0
16	Swaziland	2	Gauss	3332136	0
17	Swaziland	3	Gauss	3330507	0
18	Swaziland	4	Gauss	3340768	0
19	Swaziland	5	Gauss	2039970	0
20	Swaziland	6	Gauss	2039970	0
21	Swaziland	7	Gauss	2040024	0
22	Swaziland	8	Gauss	2050183	0
23	Swaziland	9	Gauss	193699	0
24	Swaziland	10	Gauss	195254.5	0
25	Swaziland	11	Gauss	193612	0
26	Swaziland	12	Gauss	203761	0
27	El Salvador	Time	Exp	50740.41	0
28	El Salvador	1	Exp	25281.75	0
29	El Salvador	2	Exp	25442.9	0
30	El Salvador	3	Exp	25283.46	0
31	El Salvador	4	Exp	25934.87	0
32	El Salvador	5	Exp	7610.738	0
33	El Salvador	6	Exp	7610.738	0
34	El Salvador	7	Exp	7613.293	0
35	El Salvador	8	Exp	8259.305	0
36	El Salvador	9	Exp	0	0.621541
37	El Salvador	10	Exp	160.5273	8.62E-36
38	El Salvador	11	Exp	0.992188	0.378459
39	El Salvador	12	Exp	651.3242	2.29E-142
40	El Salvador	Time	Gauss	50740.41	0

41	El Salvador	1	Gauss	22802.16	0
42	El Salvador	2	Gauss	22963.65	0
43	El Salvador	3	Gauss	22805.51	0
44	El Salvador	4	Gauss	23435.99	0
45	El Salvador	5	Gauss	13967.11	0
46	El Salvador	6	Gauss	13967.11	0
47	El Salvador	7	Gauss	13970.42	0
48	El Salvador	8	Gauss	14610.11	0
49	El Salvador	9	Gauss	1326.18	6.56E-289
50	El Salvador	10	Gauss	1486.707	9.88131291682493e-324
51	El Salvador	11	Gauss	1327.086	4.17E-289
52	El Salvador	12	Gauss	1976.551	0
53	China *P. vivax*	Time	Exp	1.2E+08	0
54	China *P. vivax*	1	Exp	59896712	0
55	China *P. vivax*	2	Exp	59898508	0
56	China *P. vivax*	3	Exp	59892352	0
57	China *P. vivax*	4	Exp	60131916	0
58	China *P. vivax*	5	Exp	18036424	0
59	China *P. vivax*	6	Exp	18036424	0
60	China *P. vivax*	7	Exp	18032128	0
61	China *P. vivax*	8	Exp	18145920	0
62	China *P. vivax*	9	Exp	4448	0
63	China *P. vivax*	10	Exp	4616	0
64	China *P. vivax*	11	Exp	0	1
65	China *P. vivax*	12	Exp	61832	0
66	China *P. vivax*	Time	Gauss	1.2E+08	0
67	China *P. vivax*	1	Gauss	54024360	0
68	China *P. vivax*	2	Gauss	54025048	0
69	China *P. vivax*	3	Gauss	54020028	0
70	China *P. vivax*	4	Gauss	54259728	0
71	China *P. vivax*	5	Gauss	33089320	0
72	China *P. vivax*	6	Gauss	33089320	0
73	China *P. vivax*	7	Gauss	33085048	0
74	China *P. vivax*	8	Gauss	33199024	0
75	China *P. vivax*	9	Gauss	3151784	0
76	China *P. vivax*	10	Gauss	3151928	0
77	China *P. vivax*	11	Gauss	3147272	0
78	China *P. vivax*	12	Gauss	3261120	0
79	China *P. falciparum*	Time	Exp	10959165	0
80	China *P. falciparum*	1	Exp	5479523	0
81	China *P. falciparum*	2	Exp	5479818	0
82	China *P. falciparum*	3	Exp	5479525	0
83	China *P. falciparum*	4	Exp	5485952	0

84	China *P. falciparum*	5	Exp	1651271	0
85	China *P. falciparum*	6	Exp	1651271	0
86	China *P. falciparum*	7	Exp	1651267	0
87	China *P. falciparum*	8	Exp	1654772	0
88	China *P. falciparum*	9	Exp	315	3.97E-69
89	China *P. falciparum*	10	Exp	290	1.06E-63
90	China *P. falciparum*	11	Exp	0	1
91	China *P. falciparum*	12	Exp	3505	0
92	China *P. falciparum*	Time	Gauss	10959165	0
93	China *P. falciparum*	1	Gauss	4941554	0
94	China *P. falciparum*	2	Gauss	4941848	0
95	China *P. falciparum*	3	Gauss	4941552	0
96	China *P. falciparum*	4	Gauss	4947978	0
97	China *P. falciparum*	5	Gauss	3027333	0
98	China *P. falciparum*	6	Gauss	3027333	0
99	China *P. falciparum*	7	Gauss	3027313	0
100	China *P. falciparum*	8	Gauss	3030842	0
101	China *P. falciparum*	9	Gauss	288014	0
102	China *P. falciparum*	10	Gauss	288308	0
103	China *P. falciparum*	11	Gauss	288046	0
104	China *P. falciparum*	12	Gauss	291521	0

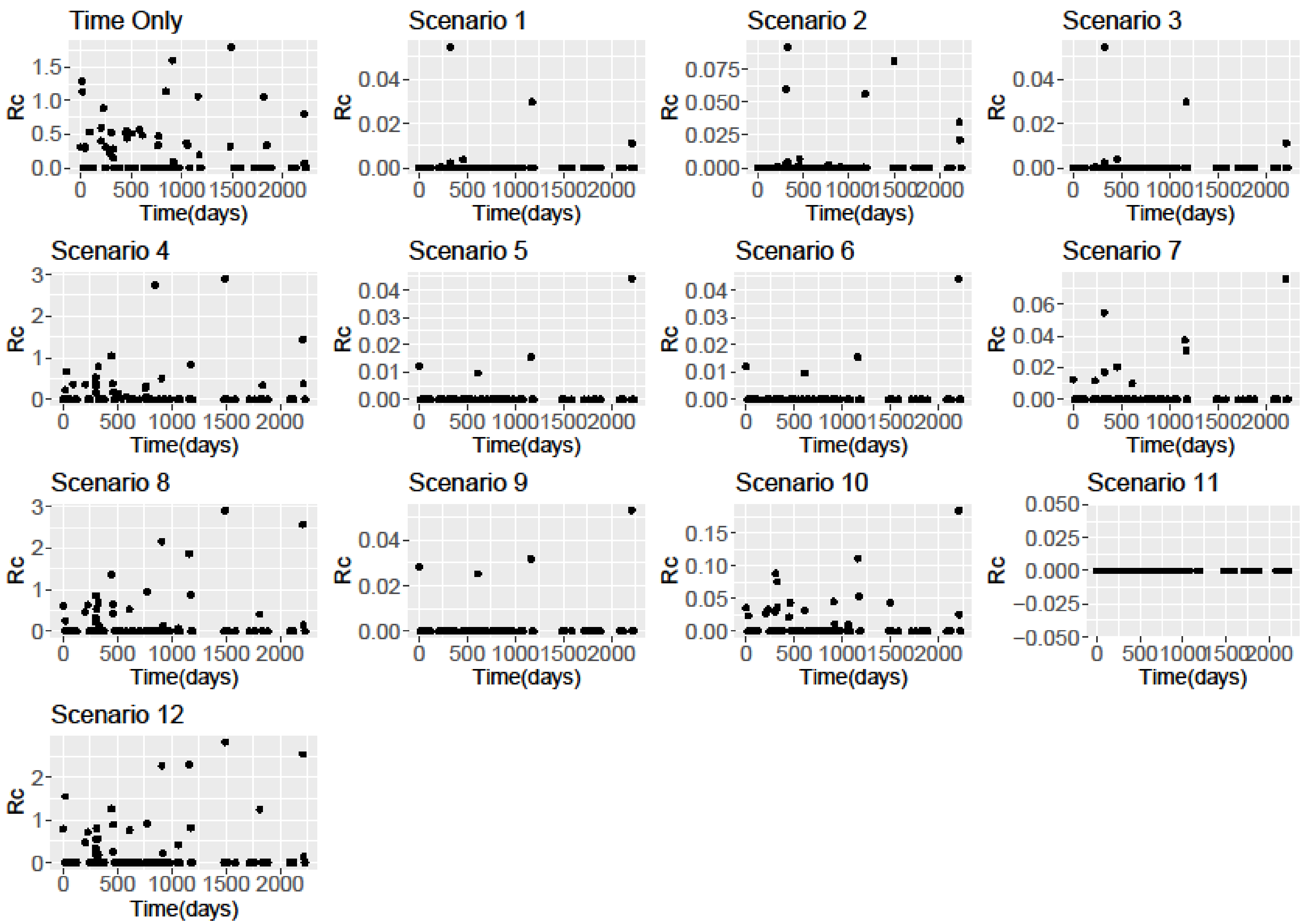


Figure 5.8: R_c estimates from El Salvador line list based on using the time-only scenario and Scenarios 1-12 with an exponential kernel

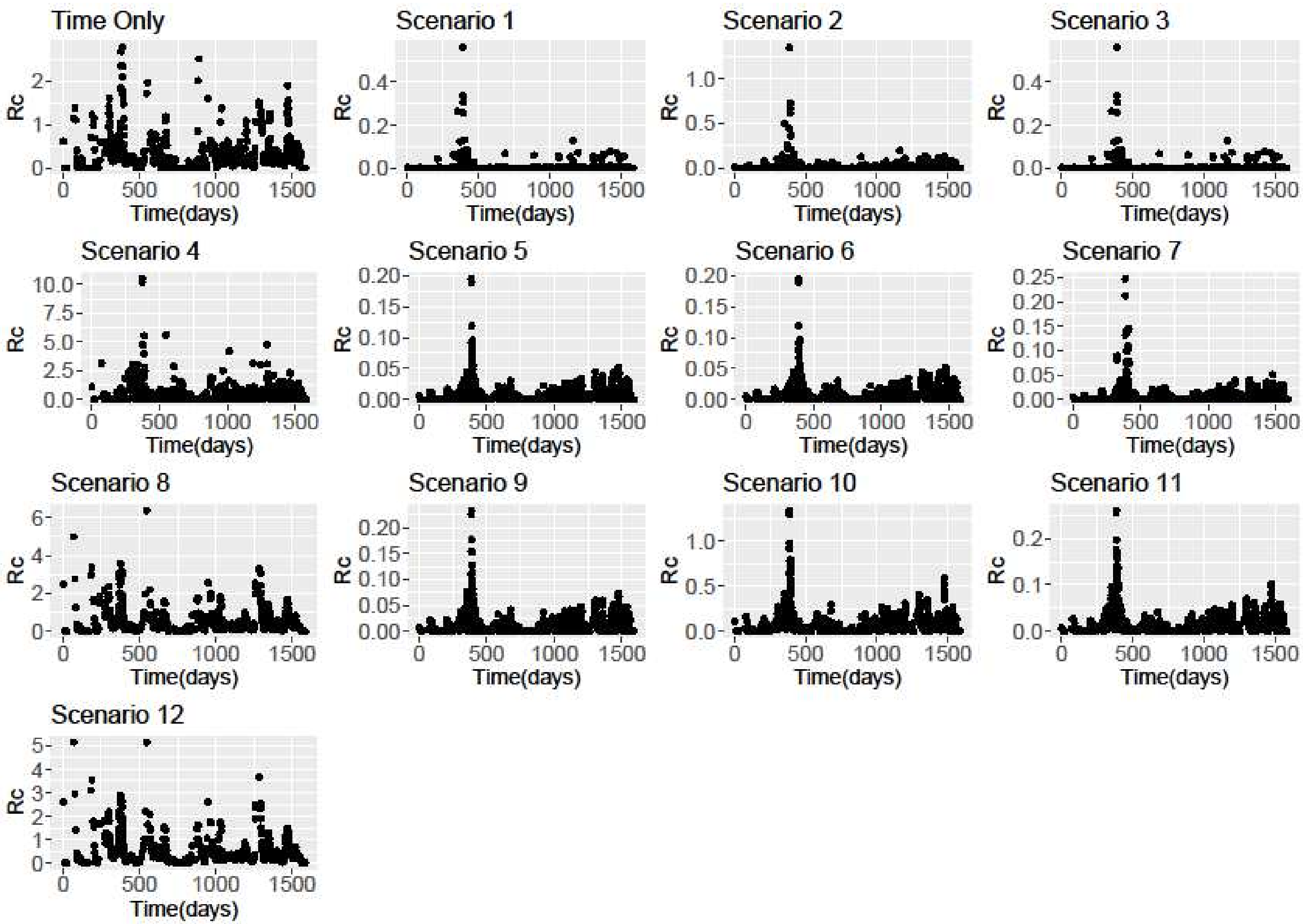


Figure 5.9 : R_c estimates from Eswatini line list based on using the time-only scenario and Scenarios 1-12 with an exponential kernel

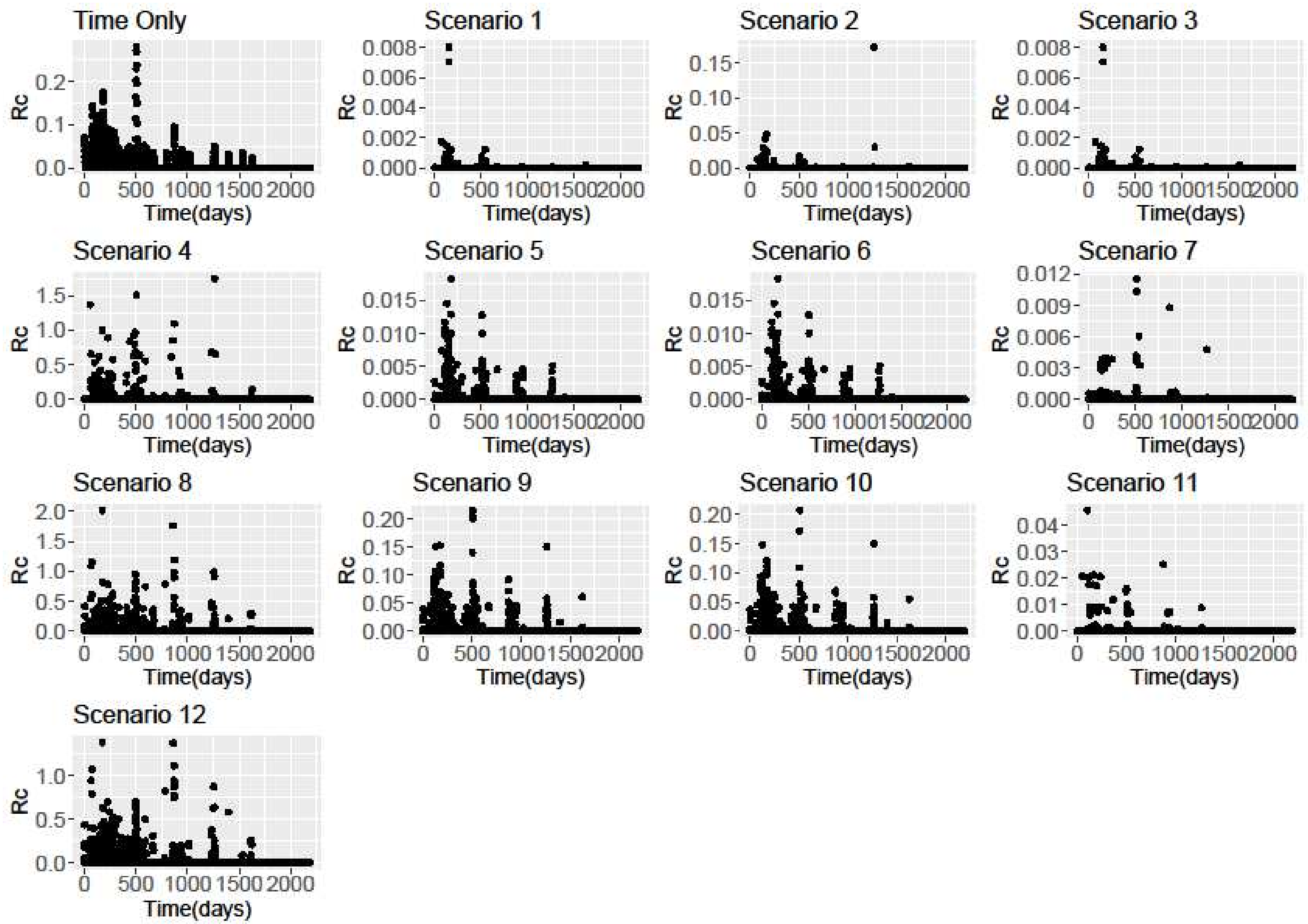


Figure 5.10: R_c estimates from China P. falciparum line list based on using the time-only scenario and Scenarios 1-12 with an exponential kernel

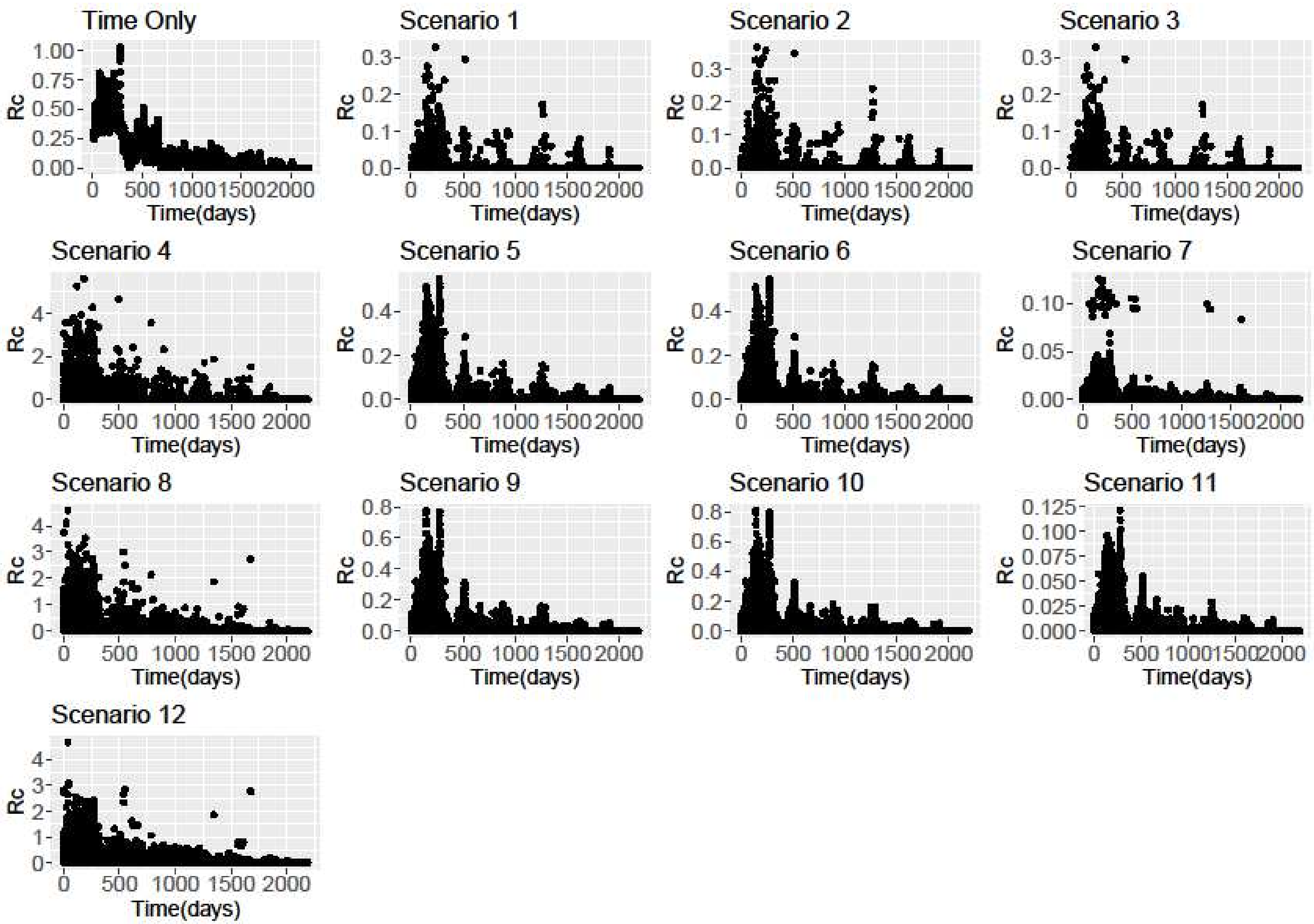


Figure 5.11: R_c *estimates from China P. vivax line list based on using the time-only scenario and Scenarios 1-12 with an exponential kernel*

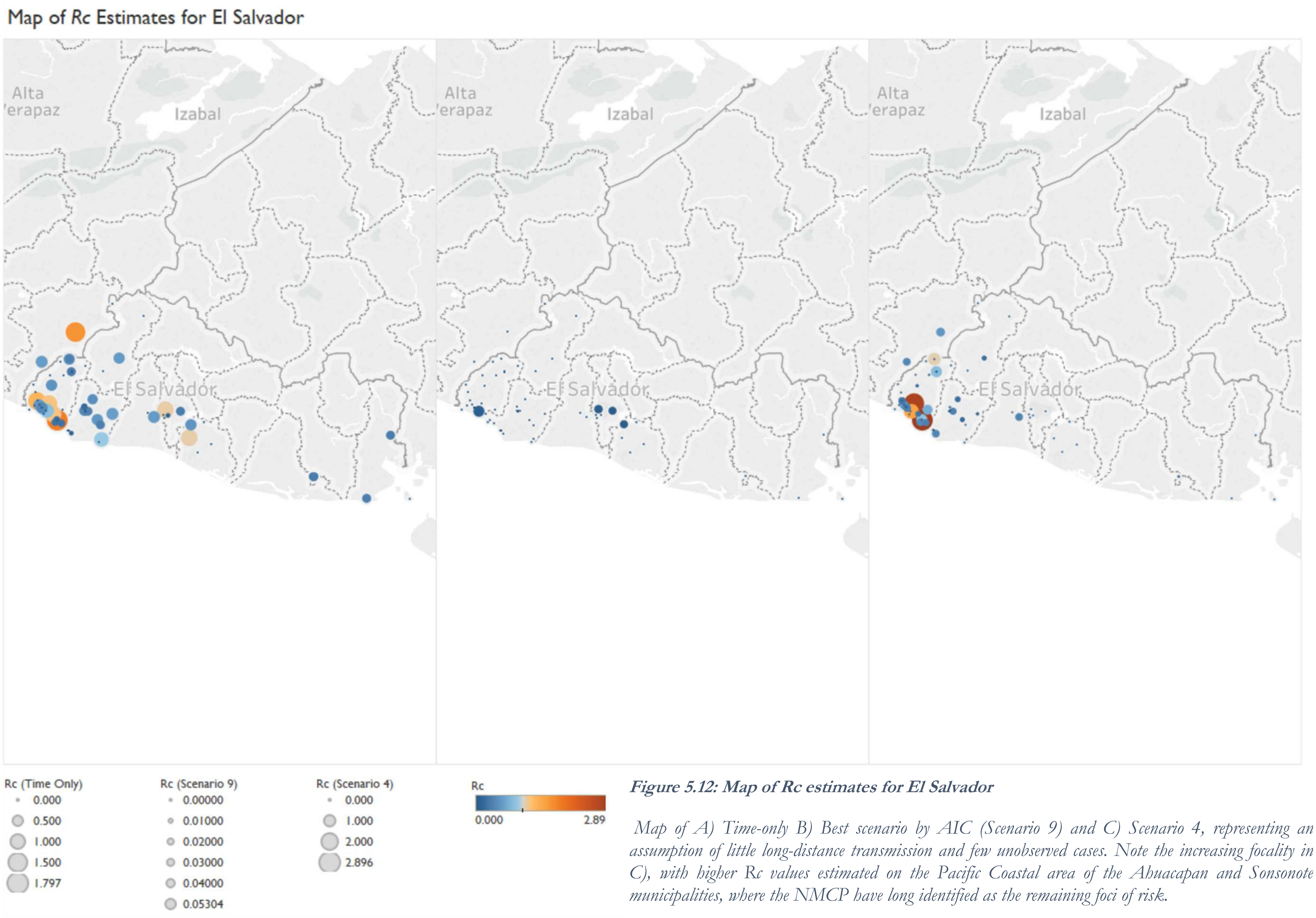


Figure 5.12: Map of Rc estimates for El Salvador

Map of A) Time-only B) Best scenario by AIC (Scenario 9) and C) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases. Note the increasing focality in C), with higher Rc values estimated on the Pacific Coastal area of the Ahuacapan and Sonsonote municipalities, where the NMCP have long identified as the remaining foci of risk.

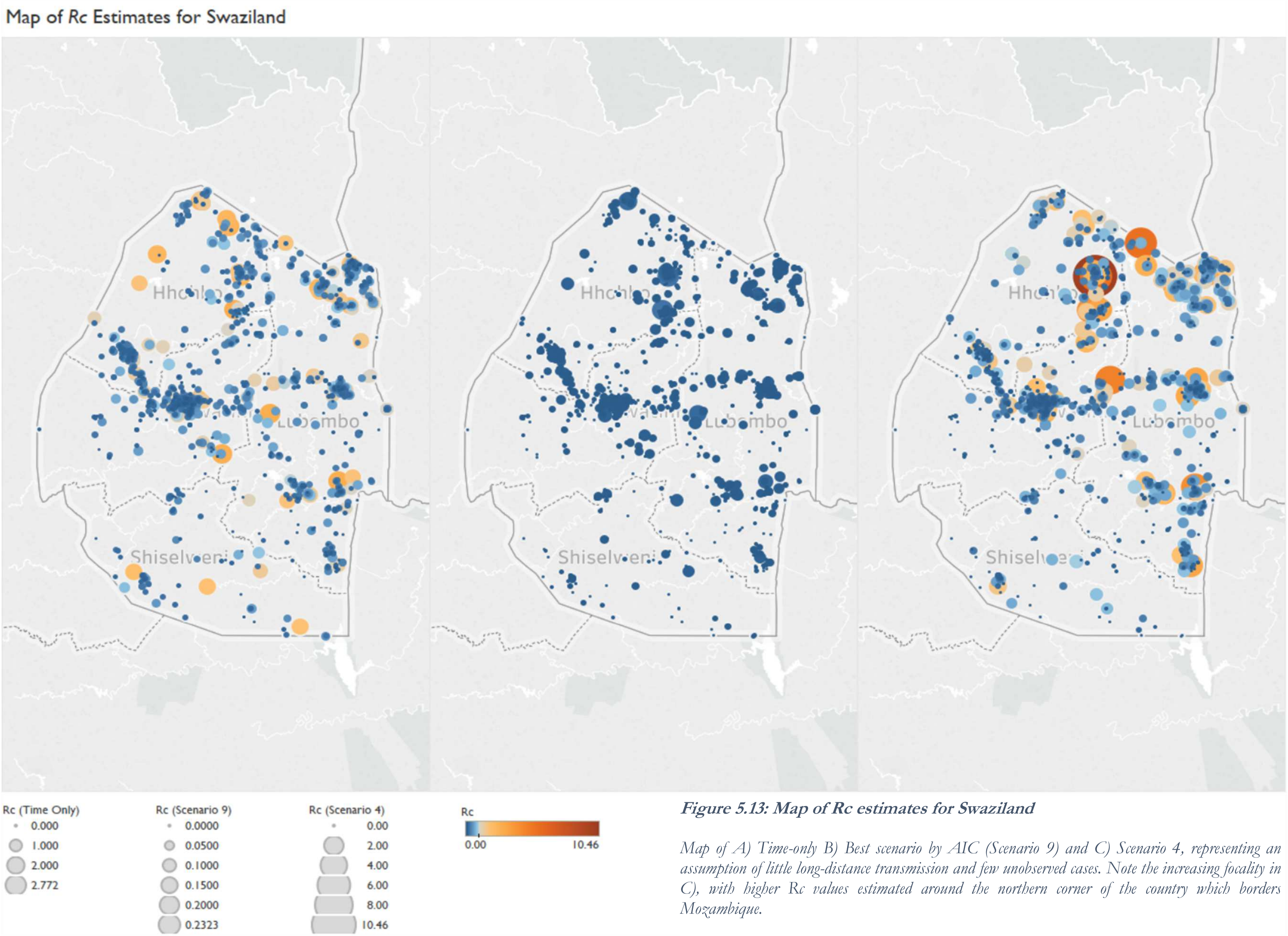


Figure 5.13: Map of Rc estimates for Swaziland

Map of A) Time-only B) Best scenario by AIC (Scenario 9) and C) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases. Note the increasing focality in C), with higher Rc values estimated around the northern corner of the country which borders Mozambique.

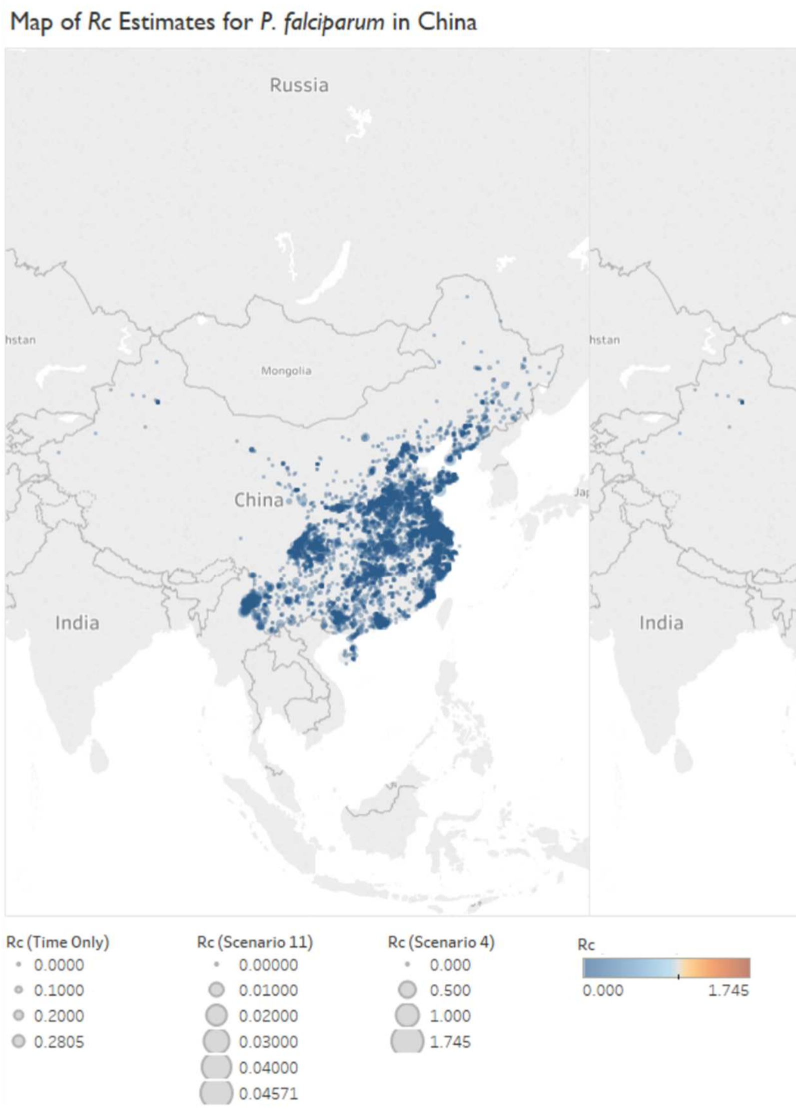


Figure 5.14 : Map of Rc estimates for P. falciparum in China

Map of A) Time-only B) Best scenario by AIC (Scenario 11) and C) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases.

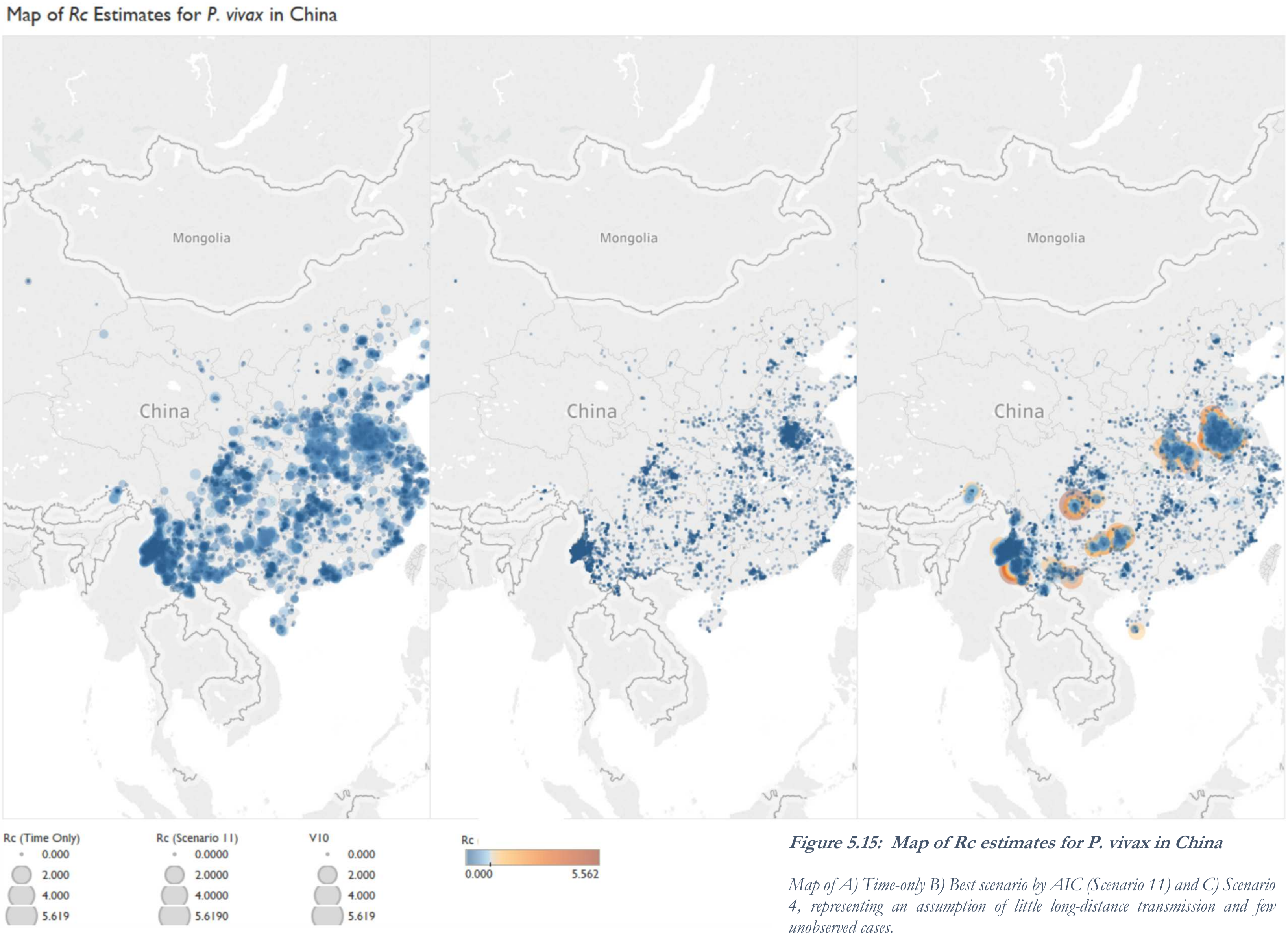


Figure 5.15: Map of Rc estimates for P. vivax in China

Map of A) Time-only B) Best scenario by AIC (Scenario 11) and C) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases.

5.3.3 Results of Sensitivity Analysis

5.3.4 El Salvador

For the line-list dataset from El Salvador, within the range of values explored in the sensitivity analysis (**Table 5.5**), regardless of how informative the prior was for either β, the distance shaping function, or for ε, the epsilon edge, β was always estimated as whatever the mean of the prior was set as between the prior mean values of 1e-4 and 1e-2 (Figure 5.16). However, when the mean value was set at 0.1, the estimated parameter converged at a slightly lower value of 0.075, with the exception of when the prior for ε was very low (all priors with mean ε of 1e-10 and also the more informative priors with mean 1e-5, when standard deviation was 1e-4). R_c is strongly shaped by the value of ε, with higher values of ε returning lower values of R_c, however R_c also declined with increasing values of β**.**

5.3.5 Eswatini

Very similar patterns to El Salvador were observed in the sensitivity analysis of the Eswatini dataset. Again, regardless of how informative the prior was for either ε or β, β was always estimated as whatever the mean of the prior was set as between the prior mean values of 1e-4 and 1e-2 (Figure 5.17). However, when the mean value was set at 0.1, the estimated parameter converged at a slightly lower value of 0.075, with the exception of when the prior for ε was very low (all priors with mean ε of 1e-10 and also the more informative priors with mean 1e-5, when standard deviation was 1e-4). Unlike El Salvador, for Eswatini, at higher values of ε (0.5 and 0.1) there are stark declines in R_c with increasing β.

5.3.6 China

For both *P. vivax* and *P. falciparum* datasets from China, within the parameter range explored in the sensitivity analysis, regardless of how informative the prior was for either β, the distance shaping function, or ε, the epsilon edge, β was always estimated as whatever the mean of the prior was set as (Figure 5.18 and Figure 5.19), suggesting a lack of identifiability or information within the data. When estimating R_c, and interesting interacting effect of ε (missing or unobserved infections) and β (distance) was seen. When β is low, although lower values of ε produce slightly higher mean R_c values, the difference in $R_c$estimates with varying prior values for ε is much smaller than when β is a higher value. In other words, when the prior for ε is low, 1e-10, R_c estimates do not vary as β changes, however when the prior for ε is much higher, then increasing β from 1e-4 to 0.1 reduces R_c estimates (from 0.21 to 0.01 for *P. vivax)*.

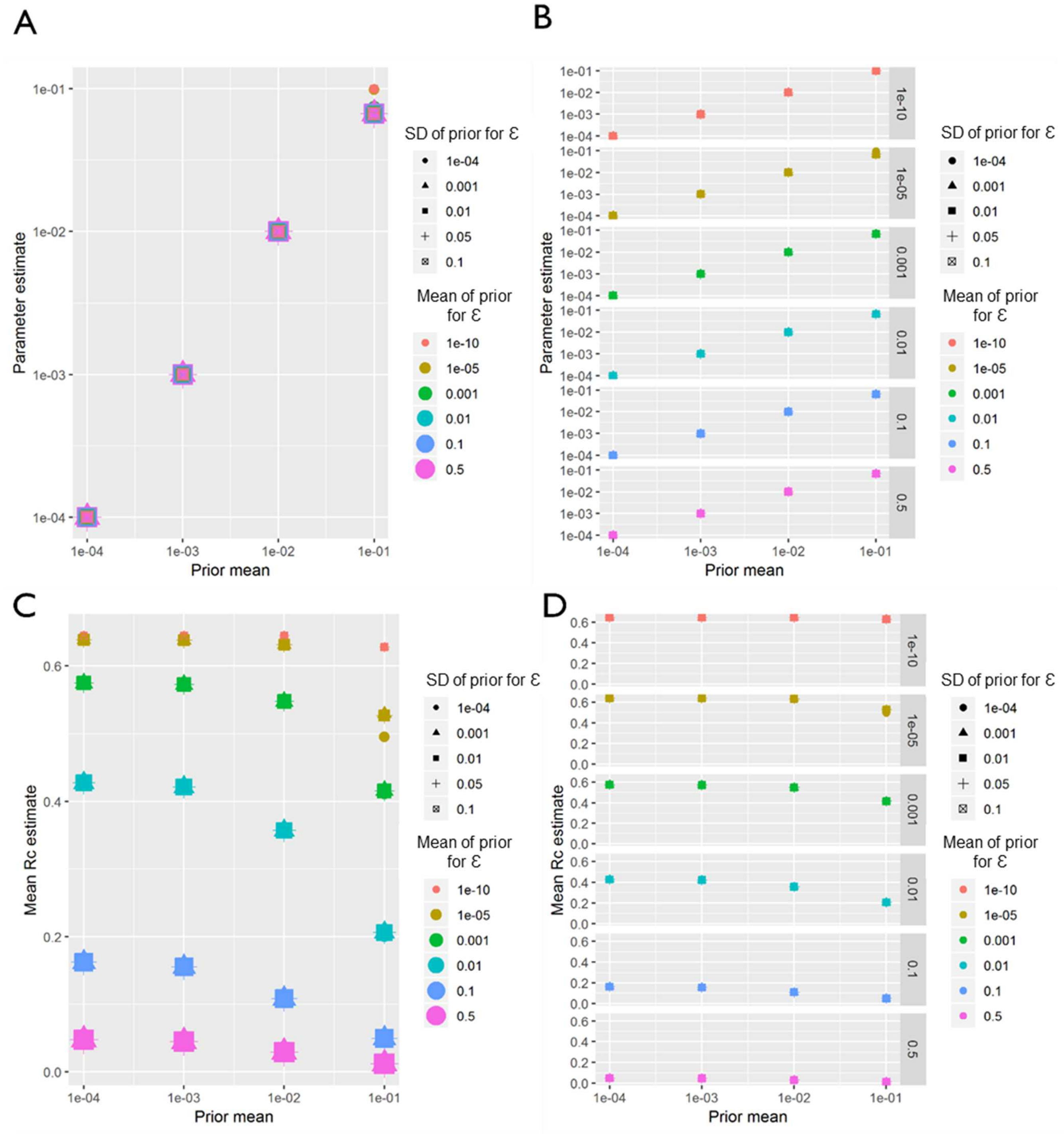


Figure 5.16 El Salvador sensitivity analysis.

Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε,which represents shapes represent different hazards of infection by an external, unobserved source. For A-D, the x-axis represents the prior mean used for β. A) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. B) shows the same results, stratified by the prior mean of ε for clarity. C) Shows the impact of priors for β and ε on the mean Rc estimate, and again D) shows the same result, stratified by the prior mean of ε.

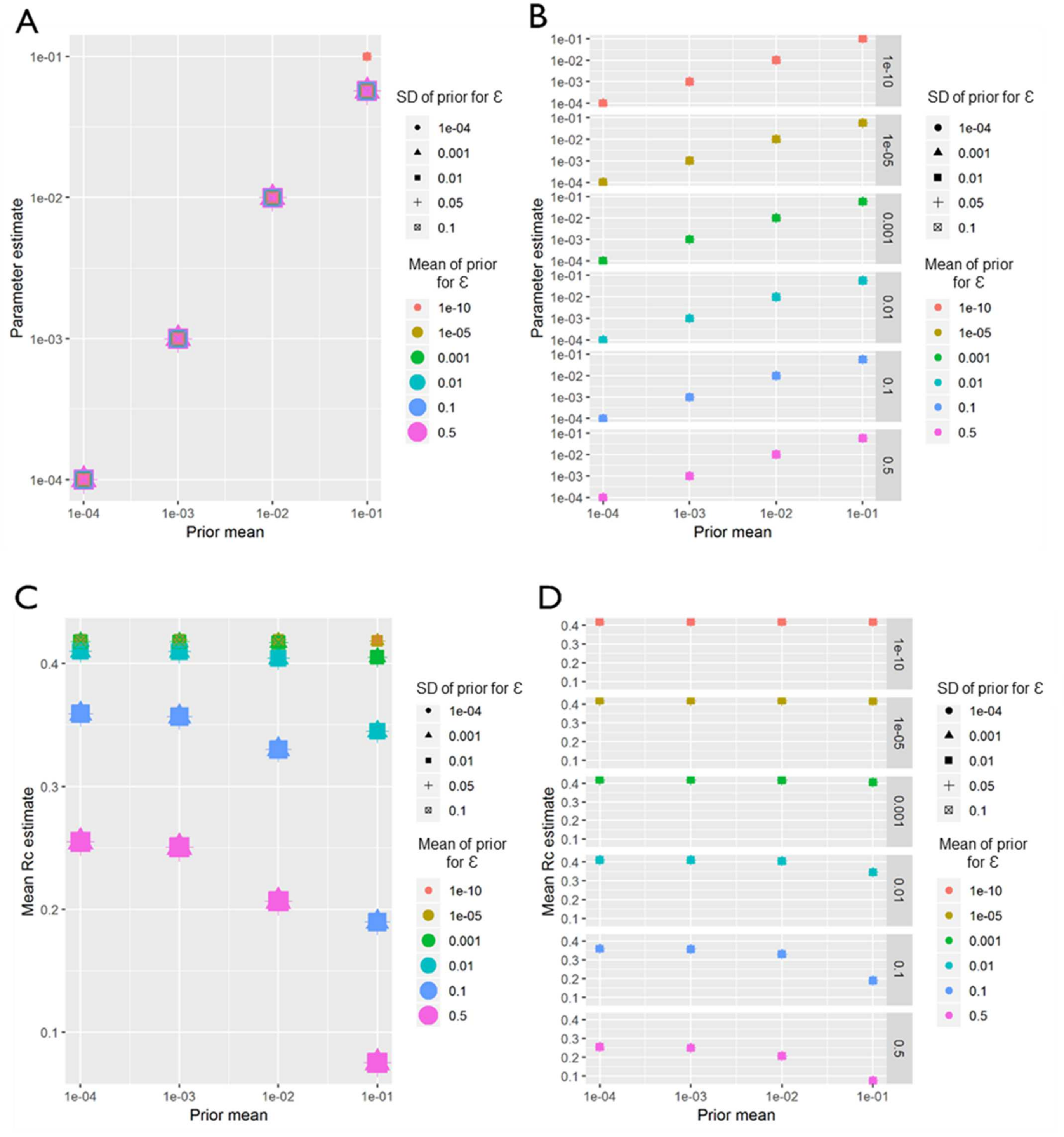


Figure 5.17: Eswatini sensitivity analysis.

Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε,which represents shapes represent different hazards of infection by an external, unobserved source. For A-D, the x-axis represents the prior mean used for β. A) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. B) shows the same results, stratified by the prior mean of ε for clarity. C) Shows the impact of priors for β and ε on the mean Rc estimate, and again D) shows the same result, stratified by the prior mean of ε.

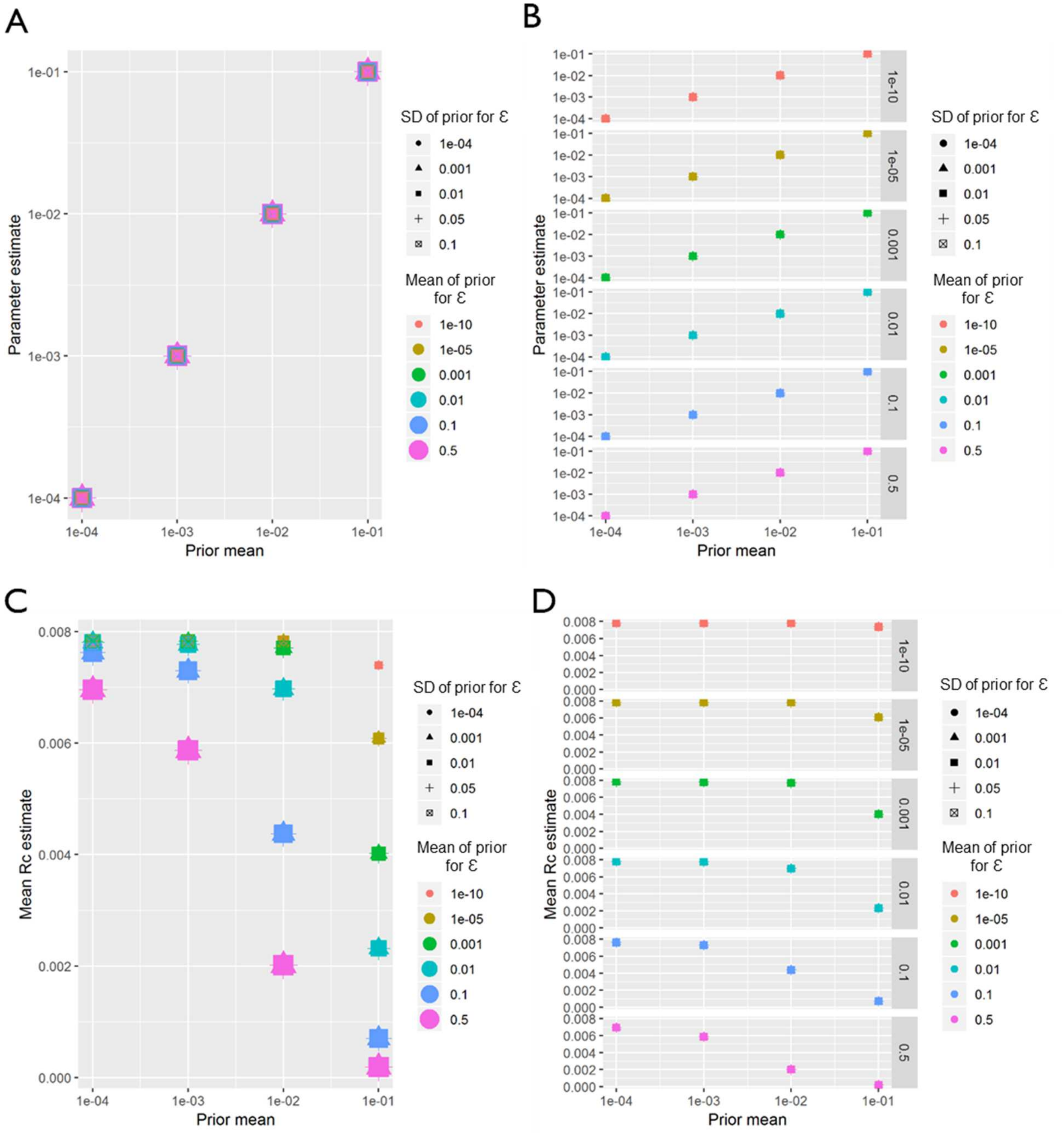


Figure 5.18: P. falciparum China Sensitivity Analysis

Sensitivity analysis showing the impact of varying the prior means for P. falciparum in China. Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, Ɛ,which represents shapes represent different hazards of infection by an external, unobserved source. For A-D, the x-axis represents the prior mean used for β. A) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. B) shows the same results, stratified by the prior mean of Ɛ for clarity. C) Shows the impact of priors for β and Ɛ on the mean Rc estimate, and again D) shows the same result, stratified by the prior mean of Ɛ.

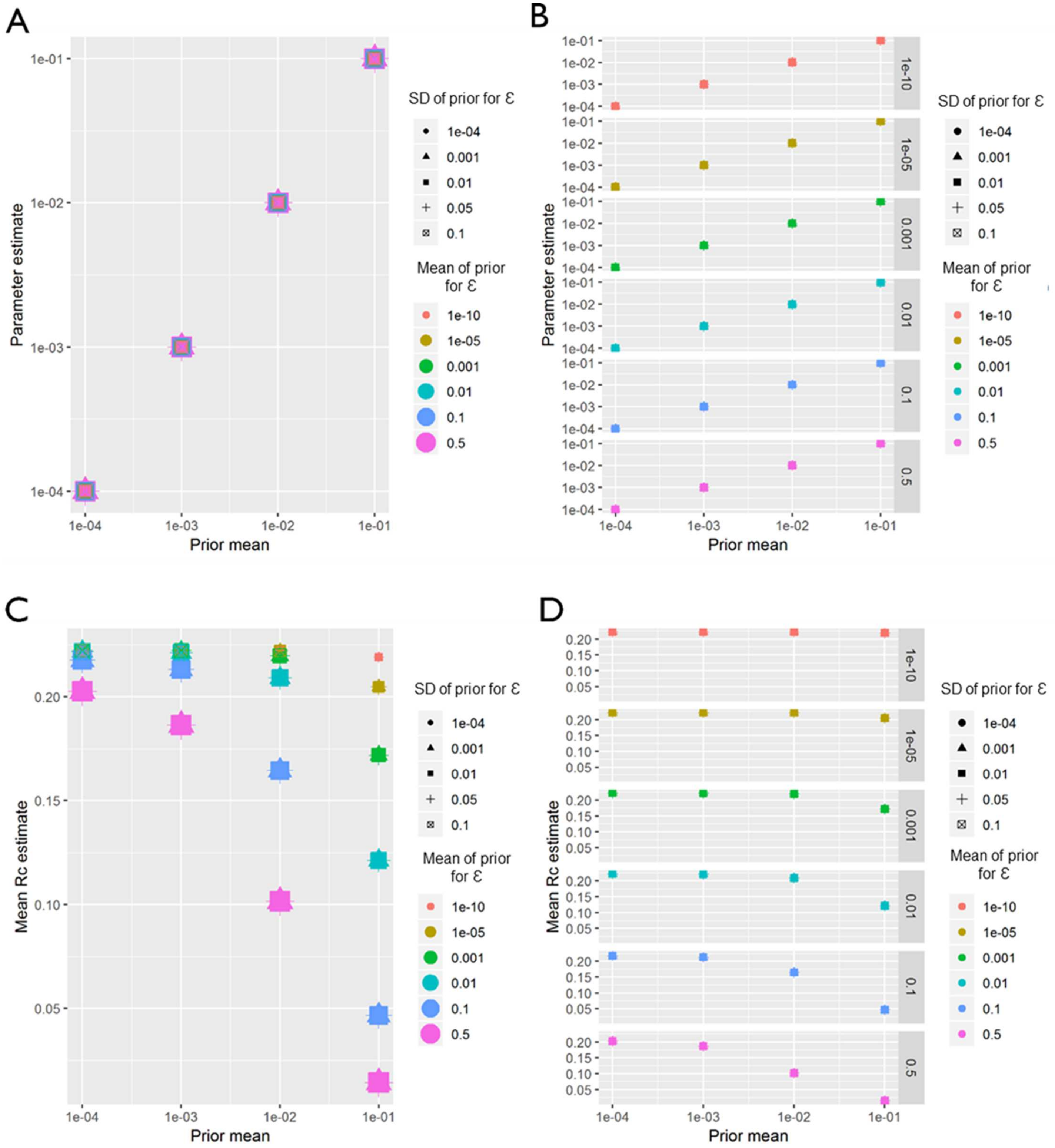


Figure 5.19: P. vivax China Sensitivity Analysis

Sensitivity analysis showing the impact of varying the prior means for P. vivax in China. Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, Ɛ,which represents shapes represent different hazards of infection by an external, unobserved source. For A-D, the x-axis represents the prior mean used for β. A) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. B) shows the same results, stratified by the prior mean of Ɛ for clarity. C) Shows the impact of priors for β and Ɛ on the mean Rc estimate, and again D) shows the same result, stratified by the prior mean of Ɛ.

5.4 Discussion

This chapter introduced a method which allows the flexible integration of distance metrics, either in the form of geographic distances, or other forms such as accessibility, with temporal information into a single inference framework. Twelve scenarios and corresponding parameter values were defined which represented varying likelihood of transmission over different geographic distances and likelihood of missing infections (as well as high and low confidence in this estimate). These scenarios were applied to four individual level datasets from malaria eliminating contexts and using two different spatial kernels. The estimated R_c values, their spatial and temporal distribution and the ΔAICc/Akaike weights for each model were compared alongside a time only model. These results suggest that including spatial information improved models as measured by AIC, compared to time only results. The prior/fixed values for both the distance function and epsilon value have very strong impacts on the estimated R_c, although relative temporal trends tend to stay consistent.

For all datasets considered, all model versions which used geographic information had lower ΔAICc values than the time only model. Based on the Akaike Weights and ΔAICc values for each model, large differences in ΔAICc were seen between different scenarios. Scenarios 9 and 11 produced the lowest ΔAICc values. These were parameterisations which penalised long range transmission the least where and the prior on epsilon edges was only weakly informative. These parameterisations also return much lower reproduction numbers than using time alone.

Exponential Kernels consistently outperformed Gaussian kernels as measured by ΔAICc. Although classic models of dispersion are as a diffusion process with Gaussian displacement, more leptokurtic or "fatter-tailed" probability distributions, where more of the probability density is concentrated in the tails of the function, are often found to better represent empirical dispersal patterns than traditional Gaussian kernels (Bateman, 1950). This "fatter-tail" in the exponential can be seen in Figure 5.5 - Figure 5.7.

However, there are many limitations to using ΔAICc in model comparison, particularly when estimation of some of the parameters are being carried out within a Bayesian context. We do not fix α_{ij} nor do we fix epsilon, but we do define priors and maximise the posterior rather than the log likelihood. Therefore, we are comparing negative log likelihoods from a maximised posterior, meaning we are not considering the information included in the prior. In addition, many α_{ij} values shrink to zero, however are still counted as parameters in the AIC estimation. Therefore, there is no recognition of which versions of the model produce fewer non-zero parameters. Whilst this difference in AIC is interesting to note, I would argue the broader trends in how R_c varies over time and space with different assumptions about both the spatial kernel and the number of unobserved sources of infection are more important to consider.

An interesting pattern which was noted across scenarios and across datasets was how including spatial information in the likelihood tended to increase the seasonality of temporal patterns in reproduction numbers and reduced noise in the temporal distribution of reproduction numbers. This could be suggestive of importation events leading to localised infections. Scenario 4 is also an interesting set of assumptions to consider as it assumes cases generally only infect cases near them and that unobserved cases of infection are unlikely. Under this assumption foci of infection are very clear and clear "sources" of infection.

The results of the sensitivity analysis reveal interesting differences between the different datasets and contexts contained in this dataset. For both El Salvador and Eswatini, which are both small countries (El Salvador has an area of 21,041 km^2 and Eswatini 17,364 km^2), at higher mean priors for β, the model converged on an estimate for β which was informed by the data. This was not the case for the dataset from China, which represents a much larger area geographically and where dynamics are likely to be strongly driven by importation. Given that for the kernels utilised in this chapter, increasing values of β lead to more restrictive assumptions about the scale of transmission, perhaps this difference is due to the different spatial scales at which the analysis was being carried out.

5.4.1 Limitations

There are several limitations to this approach and analysis. Firstly, there is a potential lack of identifiability between ε, the epsilon edge, and β, the shaping parameter of the spatial kernel. To give an intuitive example, say two cases occurred 50km from each other in space within a reasonable timeframe of symptom onset times for transmission to have occurred. Without strong prior information about what the spatial kernel may be, and/or how likely cases are to have an external source of infection, it is not clear whether these cases are linked by transmission (and there is some human travel/parasite movement, modelled by a less restrictive spatial kernel) or whether there are unobserved source(s) of infection leading to both cases. This is also exemplified in the results of the sensitivity analysis, where the mean of the prior for beta strongly shapes the final estimate of beta, and the epsilon value also shapes beta.

In the absence of reliable information about either of these values, strong assumptions must be made about either/both the likelihood of cases being infected by unobserved sources of infection and the relationship between distance and. Similar approaches (Wang, Ermon and Hopcroft, 2012) recommend fixing the kernel shaping parameter, and indeed approaches from others have also noted problems with unconstrained distance kernels in space-time diffusion modelling (Swapnil Mishra, personal correspondence). One potential way to address this is divide epsilon edge by the distance parameter $\frac{\varepsilon}{\beta}$, thereby linking the two parameters and thereby penalising increases in β.

Indeed, for similar approaches analysing the diffusion of twitter hashtags, it was recommended to fix the parameter beta, and the authors acknowledged potential challenges in estimating this parameter. Whilst the temporal aspect is not fixed, I view the utility in this method in excluding or penalising improbable transmission links between far away cases, rather than as a way of trying to determine what the spatial relationship between cases is for malaria transmission, or determining the relative contribution of space to malaria transmission.

An additional approach which could alleviate this problem is to collect internal travel history as part of surveillance in future data collection efforts. This may help tease apart the relationship between space and transmission. There also may be regions where there is more information to parameterise both the spatial scales of transmission and the likelihood of cases being unobserved (for example through looking at reporting rates, rates of relapse in the case of *P. vivax*, and prevalence of asymptomatic infection).

Secondly, as with all methods introduced in this thesis, the approach presented in this chapter was designed for application to near elimination and elimination settings, where surveillance and case management is very strong, numbers of cases are small, and therefore there is less overlap in potential infector/infectees, and changes in transmission are more apparent. If applying these approaches to contexts which are less far along the journey to elimination, the issue of identifiability may be even more of an issue as one cannot reasonably assume/fix epsilon edges to be a very small number. Asymptomatic infection will likely be more important to consider, more sophisticated methods to deal with missing cases will be required. There also will likely be a weaker signal in space and time, which may require the integration of additional information such as genetic distance. There also will be a transmission level above which these methods will no longer be useful, although we do not know what this exact level is.

Finally, due to there being no "ground truth" and Bayesian nature of model it is hard to rigorously compare model performance. ΔAICc and Akaike Weights are standard, however as mentioned previously, there are import limitations in using these metrics for model comparison. A useful future step would be to extend the simulations introduced in chapter 2 to space to investigate the impact of varying parameter values and the interaction between the shaping parameter of the spatial kernel and epsilon. Spatially explicit simulations may also reveal how tolerant the method is to missingness.

5.4.2 Future work

Currently, missing cases are dealt with in a relatively simple way, under the assumption that in the elimination settings used here, surveillance and control have been strong for an extended period of time as to ensure small case numbers and low prevalence of asymptomatic parasitaemia, and that the contribution of missing cases is small enough to be represented as a competing hazard. The latter assumption is supported by simulation results from Chapter 2 suggested that when missingness is unbiased, R_c estimates are not strongly affected but produce a slight underestimation in R_c. However, if missingness was biased, it is not clear how strongly this would affect results. Further simulations which model different forms of missing data/sampling schemes would be useful to reveal the potential impact of non-random missing data. These simulations could also model different sources of unobserved infection – for example missing cases caused by relapse of dormant *P. vivax,* unreported cases or asymptomatic infection.

Many of the potential ways to model and represent space discussed in the introduction section of this chapter (Section 5.1) have not been tested here due to the issues of identifiability seen even in simple models of space. Gravity, radiation, accessibility matrices all potential models of how space may affect the likelihood of transmission. As mosquitoes have a limited range and lifespan, developing better data and models of human movement, and how it varies in different cultural contexts and between different demographic groups, will provide useful information to appropriately parameterise and design the spatial component of the model.

Although the prior for the shaping parameter of the serial interval was selected under the assumption that the majority of cases are treated in a timely manner, In this analysis I have not explicitly utilised information about the time and location of treatment, although this is available in some contexts. This may be useful information to constrain the potential time window of infection occurring, as detailed information about infectivity and gametocyte carriage following treatment with anti-malarials is available (Bousema and Drakeley, 2011), although sub-optimal dosage, compliance and resistance have been associated with differing outcomes and therefore having additional information about treatment and prevalence of resistance would also be useful.

Another avenue for future work would be to adapt the approach to incorporate further sources of information, such as genetic markers of similarity between parasites. For the approach developed in this chapter to be useful in contexts which are not at or within a few years of elimination, incorporation of additional information into the inference framework will be required. This could be carried out either directly by incorporating an additional term or function in the likelihood or indirectly through informing the value of parameters and allowing them to vary between individuals. Previous work within the machine learning and network analysis community has successfully integrated diverse sources of information about texts such as language and similarity of context into very similar algorithms to the one presented here (Wang, Ermon and Hopcroft, 2012) .

5.5 Conclusion

Increasingly, line-list data contain spatial and other forms of information. Finding useful approaches to leverage the information contained within these diverse datasets will increasingly be useful in malaria surveillance and epidemiology (Pindolia *et al.*, 2012; Sturrock *et al.*, 2016; Wesolowski, Aimee R. Taylor, *et al.*, 2018) and developing a framework which flexibly takes on different forms of data within an integrated inference framework is a key aspect of this. There may be more useful information contained in genetic, and or travel, mobility data. However, as we have seen there can be issues of identifiability, which becomes increasingly relevant when there is not enough data available about key parameters in the model. Finding ways for leveraging multiple datasets, understanding their relationships, how they can enhance info contained in others, or used to build consensus is important.

In this chapter, I developed and tested an extension to the algorithm presented in Chapter 4, which flexibly allows the incorporation of distance or adjacency matrices describing the distance or connectivity between cases. This was applied to individual malaria case data from four eliminating and very low transmission contexts and a detailed sensitivity analysis was carried out. The results of these analyses suggest that including space improves model performance as measured by ΔAICc, and that, for the contexts considered here, the best performing models produce lower reproduction estimates than using temporal information only, likely in part due to estimating more unobserved sources of infection. However, this conclusion would be strengthened by more in-depth simulation studies. The approach presented here could be adapted to many different datasets and contexts, however issues of identifiability must be considered. The utility of this approach would be strengthened with further development of the methods of modelling unobserved sources of infection.

6

Discussion

6.1 Summary of aims and approach of thesis

In this thesis my aim was to introduce a new approach to measure malaria transmission in near elimination settings by extending, adapting and applying approaches used in network analysis of information spread through online social networks (Gomez-Rodriguez, Balduzzi and Schölkopf, 2011; Gomez-Rodriguez and Schölkopf, 2012) . With this approach, I utilised information about the time and location of cases showing symptoms of malaria to jointly infer the likelihood that a) each observed case was linked to another by transmission and b) that a case was infected by an external, unobserved source. This was carried out in a Bayesian (or in the case of Algorithm 1 used in Chapter 3, quasi-Bayesian) statistical framework to incorporate prior information about the relationship between time and the likelihood of infection occurring (Huber *et al.*, 2016). This information was then used to calculate individual reproduction numbers for each case, or how many new cases are expected to result from each case. When this number is above one, we expect transmission to continue, and below one we expect an outbreak to die out. In elimination settings, quantifying the distribution of individual reproduction numbers provides useful information about how quickly a disease may die out, and how the introduction of new cases through importation may affect ongoing transmission. These estimates were incorporated into timeseries analysis and forecasting models as well as geostatistical models to map how malaria transmission varied over space and time as well as considering timelines to elimination and the likelihood of resurgence of transmission once zero cases is achieved, as well as uncertainty in these estimates. I applied these approaches to previously unanalysed individual-level datasets of all recorded malaria cases from several eliminating contexts, including China and El Salvador.

6.2 Summary of key findings and their significance

In Chapter 3, I used the timing of symptom onset and prior distributions of the serial interval for treated, symptomatic malaria to estimate individual level reproduction numbers (R_c) for all reported and confirmed cases of malaria in El Salvador (2010 - early 2016). I then incorporated these results and the coordinates of geolocated cases into a binomial geostatistical model and explored estimates of risk of $R_c$exceeding one over space as well as time. I also fit the distribution of R_c values to several distributions to determine the expected mean R_c required to be 95% confident of observing no R_c s above zero and then fitted a Generalised Additive Model to explore the chance of the mean R_c reaching this value by 2020, based on the current declining trend. The results of this analysis suggested that whilst the average number of secondary malaria cases was below one (0.61, 95% CI 0.55–0.65), individual reproduction numbers often exceeded one during the observation period. There was an estimated decline in R_c between 2010 and 2016. However , based on the distribution of estimated R_c values during this period, characterised heterogeneity in the reproduction number using a Gamma distribution which, when fitted to the data, suggests a threshold mean $\mathcal{R}_c$ of 0.22, below which there would a less than 5% chance of any individual reproduction number exceeding one. Using the fitted trend in the mean $\mathcal{R}_c$, one would expect this level to be reached by 2023, assuming no change in the rate of importation.

In Chapter 4 I utilise an alternative framework which allows the analysis of much larger datasets within a Bayesian framework and extended and applied the approach to an individual-level dataset from China CDC. Using a geo-located individual-level dataset of cases recorded in Yunnan province between 2011 and 2016, I introduce a novel Bayesian framework to model a latent diffusion process and estimate the joint likelihood of transmission between cases and the number of cases with unobserved sources of infection. As in Chapter 3, this was used to estimate the case reproduction number, R_c, and used within spatio-temporal geostatistical models to map how transmission varied over time and space, estimate the timeline to elimination and the risk of resurgence. Using this approach, the estimated mean R_c between 2011 and 2016 was 0.171 (95% CI = 0.165, 0.178) for *P. vivax* cases and 0.089 (95% CI = 0.076, 0.103) for *P. falciparum* cases. From 2014 onwards, no cases were estimated to have an R_c value above one. An unobserved source of infection was estimated to be moderately likely (p>0.5) for 19/ 611 cases and high (p>0.8) for 2 cases, suggesting very high levels of case ascertainment. These estimates suggest that, if current intervention efforts are maintained, Yunnan is unlikely to experience sustained local transmission up to 2020. However, even with a mean R_c of 0.005 projected up to 2020, locally-acquired cases are possible due to high levels of importation. Testing the algorithm used in this chapter with simulated line-list datasets with varying levels of random missingness suggested the model can accurately return the mean R_c (+/- 0.05 when the probability of a case being observed is one, and +/- 0.15 when the probability of a case being observed is 0.3), and that setting a correct prior on the epsilon edge can return improved estimates of R_c.

In Chapter 5 I introduced a framework to incorporate additional distance metrics into the inference framework used in Chapter 4 and tested this algorithm on four line-list datasets, considering twelve scenarios and two spatial kernels describing the relationship between Euclidian distance between residences and likelihood of transmission occurring, as well as a detailed sensitivity analysis.

The contexts and datasets to which these methods were applied are very different ecologically, economically, demographically and culturally, yet both malaria control and elimination programmes provide useful success stories and pathways for other countries to learn from. My results are promising for both countries that are close to elimination certification – and indeed both El Salvador and China have reported zero locally acquired cases in 2018 (WHO, 2019b). However, the role of importation is important and it is interesting to note that in both contexts the highest estimate of R_c coincided with the highest periods of human movement. In the case of Yunnan province, recent studies of the dynamics of childhood diseases in South Western China found similar patterns, with highest transmission occurring during the time of the Spring Festival and in October, following the National Day holiday period (Saki Takahashi, personal correspondence). These are the only two periods with national week-long holidays in China. During the 40 day Chunyun period in China around the Spring Festival, there are over 3.6 billion passenger journeys estimated to occur (Wang *et al.*, 2014), and this period is described as the largest annual human migration in the world[4]. In October, the national day is also a holiday period associated with travel to visit family, and there are not obvious environmental reasons why this bimodal peak would be seen, although the rainy season and time of highest environmental suitability is May-October (Bi *et al.*, 2013)

The approaches develop in this thesis and the results of their application are relevant to elimination planning and certification in several ways.

Reproduction numbers directly relate to elimination in both a simple binary way, in terms of being above or below one, which is important for clear information for stratification and decision making, however by estimating individual reproduction numbers we also can identify the amount of individual variation, or variation over time and space, as well as looking at how close to zero estimates are, as an estimate of how quickly a disease will die out. Reproduction numbers have been useful metrics in a wide variety of outbreak scenarios to reveal characteristics of transmission, such as the amount of within community and within hospital transmission of Ebola (Faye *et al.*, 2015), changes in transmission intensity over time (Cori *et al.*, 2013) and assessing (Boëlle *et al.*, 2011), and in near-

[4] https://edition.cnn.com/travel/article/lunar-new-year-travel-rush-2019/index.html

elimination and elimination settings also reveal changes in transmission over space and time for malaria.

It is important to consider what these analyses reveal that is different from merely looking at incidence timeseries. In both El Salvador and China, the periods of year which had highest R_c estimates did not coincide with times of highest incidence, but instead periods of increased human movement. This could be explained by an increase in imported cases (classified correctly or incorrectly), leading to short stuttering chains of ongoing transmission. In addition, being able to map risks of R_c exceeding zero or one provide a clear stratification of risk, and can highlight areas where there may not be cases observed but where there may still be risk of resurgence of cases if importation were to occur.

The results of this thesis demonstrate how a network-based approach can provide additional insights into transmission in near elimination settings, identifying when R_c falls below 1, as well as estimating trajectories towards elimination. By incorporating these estimates into geostatistical models, this work also quantified where there was high and low uncertainty about there being minimal risk of ongoing transmission or resurgence, and how this has varied over time.

In addition, whilst there is a great deal of uncertainty associated with the forward projections for timelines to elimination, providing countries with these trajectories to elimination, and associated uncertainties, such as those produced during this thesis, can provide evidence to sustain current interventions and also highlight risks of resurgence.

Together, these provide helpful insights for elimination programmes, and the methods introduced in this thesis have attracted the attention of several national malaria elimination programmes who were interested in applying the approach to their data.

There are several important policy implications and considerations when implementing the findings described in this thesis. The current WHO definitions of elimination, namely of three consecutive years of zero locally acquired cases, is very difficult to achieve in countries where there are large amounts of importation. Although technically cases which are the result of infection by an imported case (introduced cases) are not classified as locally acquired by WHO definitions, in practice it is very difficult to classify cases as introduced. Approaches such as those introduced in this thesis could be used to identify the likelihood of cases being introduced cases rather than truly locally acquired cases, however this would require the imported case to have been observed. Furthermore, a country observing many importation events actually provides much stronger evidence that a country has achieved elimination and a low underlying receptivity to malaria if there are no or few resulting local cases than a country which has not experienced importation. Therefore, I would argue that the WHO should also factor importation into decision making when assessing the strength of evidence of elimination being achieved. The impact of importation also should be considered in relation to regional elimination. Certainly regional, international collaboration within both the Greater Mekong Subregion and the EMMIE initiative in Mesoamerica and Hispanola are thought to be key to ensuring the maintenance of very low cases of malaria in both El Salvador and China. The level of importation seen, particularly in China is high enough to render cross border collaboration essential.

There are both opportunities and limitations which must be considered when applying the approaches taken in this thesis to other contexts. As we have seen, when both spatial relationships and the prevalence of unobserved sources of infections are unknown, inferred reproduction numbers and their spatial distribution can vary depending on prior assumptions about their values. This approach would be suited to similar contexts with small numbers of cases and surveillance systems, such as the Cabo Verde islands, where there has been interest in using these approaches to analyse line lists of malaria outbreaks on the island (Dr Kimberley Linblade, personal communications)

Due to the uncertainty and assumptions made in this work, I do not advocate using these results to determine when and where to scale back interventions or surveillance. This is particularly notable with the forward forecasting and the risk maps produced in chapters 3 and 4. It is important to communicate this uncertainty clearly if this work were to be used in decision-making. For example, the standard deviation and lack of observations made in the maps of El Salvador mean there are large areas of the map where there is so much uncertainty that the mean values cannot be used for decision making, and the focus should be on the area bordering Guatemala, where there have been more observations and more certainty in estimates. However, one major policy implication of this research surrounds when and where to introduce enhanced surveillance. This work has identified times of year and localities where there is likely to be some risk of transmission with little uncertainty in these estimates. Where there is greater uncertainty in estimates due to a lack of data, active case detection or cross sectional surveys may be helpful to reduce this uncertainty.

6.3 Limitations

There are several limitations to the approaches taken here and to the datasets which these approaches were applied to. The frameworks developed in this thesis are designed to be general, flexible and adaptable to a variety of data types and elimination contexts. As a result, in order to adapt them to specific problems and datasets many assumptions are made about malaria in these contexts. In some cases, the ability to make recommendations based on the results presented here are limited by uncertainties in key parameter values. This is particularly true because this work has focussed on retrospective studies of historical surveillance data, and therefore it is not possible to collect additional data about the cases, for example through taking travel surveys of cases or collecting molecular data. However, given this is often the most widely available data collected by Ministries of Health or National Malaria Control Programmes, it is important to make the most of these data, show the impact of varying assumptions and illustrate what one might expect under different scenarios.

In all versions of the algorithm presented in this thesis, the minimum incubation period has been quite conservative, set at 15 days, to avoid erroneously excluding cases. In reality, the combined intrinsic and extrinsic incubation period is likely to be longer than 15 days (Boyd and Kitchen, 1937; Kitchen and Boyd, 1937; Nishiura *et al.*, 2007). Also, the assumption of the majority of cases being symptomatic, treated cases and therefore having a relatively less variable serial interval strongly impacts assumptions and the ability to infer connectivity between cases, as untreated malaria has a much wider range of potential serial interval lengths (Huber *et al.*, 2016) .

There also is uncertainty in the reporting rate. Both El Salvador and China have invested in strong surveillance in rural areas, and carry out both active and passive surveillance, and so I make the assumption in this thesis that the reporting rate is high. In addition, the results of the simulations carried out in Chapter 2 of the thesis suggest that the model is relatively robust to missing cases in terms of the R_c estimates, and that epsilon edges, if given an accurate informative prior, can account for this missingness. However, it is important to note that the simulations carried out in this thesis assumes missingness is proportional and unbiased. I also make the related assumption that overall case detection is high/unbiased in missingness. In reality, there are key groups who may be less likely to be detected yet more likely to contribute to onward infection in some contexts e.g. itinerant workers who spend some time in forests.

There is also uncertainty in asymptomatic infection prevalence, sub-microscopic infection prevalence and contribution of both to ongoing transmission. Meta-analyses and reviews (Okell *et al.*, 2012; Teun Bousema *et al.*, 2014; Tadesse *et al.*, 2018) can provide an indication of likely levels given the incidence and prevalence, but given the evidence in these works that asymptomatic and/or sub-microscopic infection prevalence and its contribution to ongoing transmission is highly varied, it is difficult to estimate whether asymptomatic reservoirs are contributing to ongoing transmission. In the contexts explored in this thesis, i.e. elimination settings, the numbers of cases are very small and sparse temporally, and incidence/prevalence has been maintained at a very low level over long periods of time. As a result, we assume that asymptomatic infections are unlikely to have a strong contribution to ongoing transmission, and any rare asymptomatic sources can be captured by ε-edges. However, in low transmission settings which have recently seen rapid declines in malaria incidence, or where there is a much higher incidence of cases, quantifying the asymptomatic reservoir will be important.

There is also uncertainty in the accuracy of imported/local classification. Whilst this classification has been carried out on the basis of epidemiological investigation and taking travel history, there may be inaccuracies in travel history, or in the case of *P. vivax*, an underlying infection which was acquired months prior. Others have found that when attempting to reconstruct transmission networks based on simulated surveillance data, assuming the travel history is correct produces better results than ignoring it or allowing it to be incorrect if no other information is available (Alex Perkins, personal correspondence).

There are aspects of *Plasmodium vivax* life history and epidemiology which I have made simplifying assumptions about in the analyses described in this thesis. The models do not explicitly model reinfection and relapse. This is likely to be a reasonable assumption in the contexts considered for this thesis, due to the lack of evidence for relapse in the electronic record, and through the policy of treating P. vivax cases with radical cure. As discussed in more depth in Chapter 4, relapse cases incorrectly identified as new cases would bias the results to estimate higher reproduction numbers than the true values, and therefore if this was the case would actually provide stronger evidence of low transmission levels achieved in both El Salvador and China. However, there are other aspects of *P. vivax* epidemiology which could have shaped model results and should be discussed. For example, with *P. vivax* there can be infectiousness before symptoms in first and subsequent relapse or balanced by declining parasitaemia with relapse (and the presence of individual infection registers for known vivax patients).

Finally, this analysis does not differentiate between data collected actively and passively, although reactive and targeted active case detection is used in surveillance in all the contexts presented in this thesis. Due to the non-random nature of reactive and targeted active case detection, data may be biased towards observing cases occurring in areas already identified as foci or higher risk, or close in time and space to other cases. Whilst this makes sense operationally, it means the datasets analysed may not be an unbiased sample of cases.

6.4 Future Directions

As discussed in Chapter 5, there is potential to incorporate more sophisticated models of human movement, such as gravity or radiation models, as well as accessibility matrices or friction surfaces. These methods are limited by the quality of the data available to parameterise population estimates using tools such as WorldPop (Tatem, 2017) , friction surfaces or population movement models, but provide approximate estimates which may help weight or exclude probable or unlikely locations of transmission pairs. In addition, simulations to test the assumptions and accuracy of the algorithms could be expanded to include space, to use a Rayleigh probability distribution for direct comparison. This would provide important and useful information to help tease apart some of the identifiability issues identified in Chapter 5, as well as demonstrate the impact of different patterns of movement/parasite dispersal on observed incidence. Combined with simulations of different forms of missing case data, these simulations would provide a clearer understanding of the applicability of the approaches introduced here to different epidemiological contexts.

In the contexts explored here, genetic and serological data were not available, however there is increasing evidence of their utility. Parasite genetic data have been found to provide useful information about latent processes such as past and current malaria transmission intensity (Nkhoma *et al.*, 2013; Wesolowski, *et al.*, 2018; Dalmat *et al.*, 2019) and the movement of parasites between populations (Chang *et al.*, 2019; Dalmat *et al.*, 2019; Tessema *et al.*, 2019). In near-elimination settings, genetic information may be most useful in identifying imported cases, however this is dependent upon the location of importation and the availability of reference genomes from importation population.

With current sequencing technologies it is now feasible to collect genetic data as part of routine surveillance systems. However, methods to relate the signal in genetic data to epidemiologically relevant metrics are lacking. Key questions remain as to what types of genetic data are the most useful to collect, which sampling frameworks are optimal to use and how to meaningfully integrate genetic data with other data streams, such as traditional surveillance, to infer parameters of interest.

As more countries reach the elimination stage for malaria and improve their surveillance, detection and response to malaria infection there is increasing applicability and utility of using methods such as this. In Chapter 5 we see the impact of uncertainty in both unobserved sources of infection and the distance kernel on performance of the methods. Based on these results, the sorts of contexts where this approach is suitable would be in contexts where there is good information about the travel patterns of people and/or where the amount of missingness can be quantified, however this could be better understood through more sophisticated simulations and investigations into the added benefit of incorporating additional sources of information.

In addition, in order for the approaches developed in this thesis to be utilised by control programmes, they would need to be packaged into an operationally useful tool. This would require collaboration and consultation with both control programmes and initiatives but would greatly improve the utility of approach.

However, there is a limit to what can be inferred from existing data. Further studies to better characterise asymptomatic reservoirs and their contributions to ongoing infection (Tadesse *et al.*, 2018), reporting rates, and patterns and incidence of relapse in *P. vivax* endemic areas (White *et al.*, 2016) are required and will help parameterise these models. The different causes and prevalence of unobserved infections can indicate how well a current surveillance system is capturing the true dynamics of infection, as well as which interventions may be required to achieve progress towards elimination. From an intervention standpoint, different interventions may be suited to different sources of unobserved infection. For example, if asymptomatic reservoirs are known to be a major driver of residual malaria transmission, then they can be targeted through active case detection programmes, or through mass drug administration. However, if the contribution of asymptomatic reservoirs is negligible, then interventions focused on detecting and treating symptomatic individuals (as well as vector control) can be prioritised. In addition, diverse approaches and data collection at different scales are required to understand travel patterns relevant to malaria transmission, from mechanistic modelling and large-scale data analysis e.g. of mobile phone data, to focused, on the ground studies.

More broadly, the case studies of China and El Salvador, countries reaching elimination, highlight the importance of regional and cross-border collaboration and initiatives. The importance of importation and cross-border movement in both contexts also highlights the utility of investing in reducing burden in neighbouring countries, thereby reducing the amount of importation into eliminating countries. Indeed, as China looks highly likely to reach three consecutive years of zero cases by 2020, there has been encouragement from the international community for China to make financial aid commitments to malaria control efforts elsewhere. Although this has been framed as being in celebration of China's achievements, investment in malaria control in high transmission countries could also be beneficial in reducing the likelihood of resurgence via importation.

6.5 Conclusions

Although malaria is still responsible for a great deal of death and illness in many parts of the world, many national control programmes have made great strides in controlling malaria and now are able to aim for elimination. However, in order to monitor progress towards elimination and plan interventions, it is crucial to measure malaria transmission and how it varies over space and time. In this thesis, I introduced an approach to flexibly incorporate line-list data to quantify reproduction numbers and how they varied over space and time, applying two individual level datasets from elimination countries. The results highlight the successes achieved by both China and El Salvador – the only two E-2020 countries to have zero locally acquired cases in 2018 which have not yet been certified as eliminated. This work shows the importance of considering not only environmental factors for seasonal patterns in malaria transmission, but the potential for human culture and movement patterns to also play a role in transmission dynamics in elimination settings. These tools could be of use to other national malaria control programmes to assess trajectories towards elimination based on recent historical line-list data.

References

Abadi, M. *et al.* (2015) *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Available at: www.tensorflow.org. (Accessed: 24 August 2018).

Akaike, H. (1974) 'A New Look at the Statistical Model Identification', *Automatic Control, IEEE Transactions on*, 19(6), pp. 716–723. doi: 10.1109/TAC.1974.1100705.

Alonso, P. (2016) *A Framework for Malaria Elimination*, *WHO*. World Health Organization. Available at: http://www.who.int/malaria/publications/atoz/9789241511988/en/ (Accessed: 29 May 2017).

Bateman, A. J. (1950) 'Is gene dispersion normal?', *Heredity*, 4(3), pp. 353–363. doi: 10.1038/hdy.1950.27.

Battle, K. E. *et al.* (2014) 'Geographical variation in Plasmodium vivax relapse.', *Malaria journal*. Malaria Journal, 13(1), p. 144. Available at: http://www.malariajournal.com/content/13/1/144.

Battle, K. E. *et al.* (2019) 'Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study', *The Lancet*. Elsevier, 394(10195), pp. 332–343. doi: 10.1016/s0140-6736(19)31096-7.

Beier, J. C. *et al.* (2012) 'Attractive toxic sugar bait (ATSB) methods decimate populations of Anopheles malaria vectors in arid environments regardless of the local availability of favoured sugar-source blossoms', *Malaria Journal*. BioMed Central, 11(1), p. 31. doi: 10.1186/1475-2875-11-31.

Bejon, P. *et al.* (2010) 'Stable and Unstable Malaria Hotspots in Longitudinal Cohort Studies in Kenya', *PLoS Medicine*. Edited by T. A. Smith. Public Library of Science, 7(7), p. e1000304. doi: 10.1371/journal.pmed.1000304.

Bejon, P. *et al.* (2014) 'A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots.', *eLife*. eLife Sciences Publications, Ltd, 3, p. e02130. doi: 10.7554/eLife.02130.

Bhatt, S. *et al.* (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature*. Nature Research, 526(7572), pp. 207–211. doi: 10.1038/nature15535.

Bi, Y. *et al.* (2013) 'Impact of climate variability on Plasmodium vivax and Plasmodium falciparum malaria in Yunnan Province, China', *Parasites & Vectors*. BioMed Central, 6(1), p. 357. doi: 10.1186/1756-3305-6-357.

Biggs, J. *et al.* (2017) 'Serology reveals heterogeneity of Plasmodium falciparum transmission in northeastern South Africa: implications for malaria elimination', *Malaria Journal*. BioMed Central, 16(1), p. 48. doi: 10.1186/s12936-017-1701-7.

Blumberg, S. *et al.* (2013) 'Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains', *PLoS Computational Biology*. Edited by N. Ferguson. Public Library of Science, 9(5), p. e1002993. doi: 10.1371/journal.pcbi.1002993.

Boëlle, P.-Y. *et al.* (2011) 'Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: a review.', *Influenza and other respiratory viruses*, 5(5), pp. 306–16. doi: 10.1111/j.1750-2659.2011.00234.x.

Bousema, T. *et al.* (2012) 'Hitting hotspots: spatial targeting of malaria for control and elimination.', *PLoS medicine*. Public Library of Science, 9(1), p. e1001165. doi: 10.1371/journal.pmed.1001165.

Bousema, T *et al.* (2014) 'Asymptomatic malaria infections: detectability, transmissibility and public health relevance', *Nature reviews Microbiology*. Available at: http://www.nature.com/nrmicro/journal/v12/n12/abs/nrmicro3364.html (Accessed: 6 June 2017).

Bousema, T. and Drakeley, C. (2011) 'Epidemiology and Infectivity of Plasmodium falciparum and Plasmodium vivax Gametocytes in Relation to Malaria Control and Elimination', *CLINICAL MICROBIOLOGY REVIEWS*, 24(2), pp. 377–410. doi: 10.1128/CMR.00051-10.

Boyd, M. F. and Kitchen, S. F. (1937) 'The Duration of the Intrinsic Incubation Period in Falciparum Malaria in Relation to Certain Factors Affecting the Parasites 1', *The American Journal of Tropical Medicine and Hygiene*. The American Society of Tropical Medicine and Hygiene, s1-17(6), pp. 845–848. doi: 10.4269/ajtmh.1937.s1-17.845.

Breban, R., Vardavas, R. and Blower, S. (2007) 'Theory versus data: how to calculate R0?', *PloS one*, 2(3), p. e282. doi: 10.1371/journal.pone.0000282.

Britton, T. and O 'Neill, P. D. (2002) 'Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure'. Available at: http://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/b5270c2d37901915c125795700461cda/$FILE/1467-9469.00296.pdf (Accessed: 24 May 2017).

Brockmann, D., Hufnagel, L. and Geisel, T. (2006) 'The scaling laws of human travel', *Nature*. Nature Publishing Group, 439(7075), pp. 462–465. doi: 10.1038/nature04292.

Brookmeyer, R., Gail, M. and Gail, M. (1994) *AIDS epidemiology: a quantitative approach*. Available at: https://books.google.com/books?hl=en&lr=&id=IT7nCwAAQBAJ&oi=fnd&pg=PR13&ots=SeY3dEOkrU&sig=8MEijhUw1mNOAragvraG1_LGmKQ (Accessed: 16 August 2019).

Cameron, D. and Jones, I. G. (1983) 'John Snow, the broad street pump and modern epidemiology', *International Journal of Epidemiology*, 12(4), pp. 393–396. doi: 10.1093/ije/12.4.393.

Cao, J. *et al.* (2014) 'Communicating and Monitoring Surveillance and Response Activities for Malaria Elimination: China's "1-3-7" Strategy', *PLoS Medicine*. Public Library of Science, 11(5), p. e1001642.

doi: 10.1371/journal.pmed.1001642.

Carter, K. H. *et al.* (2015) 'Malaria in the Americas: Trends from 1959 to 2011', *Am. J. Trop. Med. Hyg. Pan American Health Organization/World Health Organization*. The American Society of Tropical Medicine and Hygiene, 92(2), pp. 302–316. doi: 10.4269/ajtmh.14-0368.

Carter, R., Mendis, K. N. and Roberts, D. (2000) 'Spatial targeting of interventions against malaria', *Bulletin of the World Health Organization*, pp. 1401–1411.

Cauchemez, S. *et al.* (2006) 'Estimating in Real Time the Efficacy of Measures to Control Emerging Communicable Diseases', *American Journal of Epidemiology*, 164(6), pp. 591–597. doi: 10.1093/aje/kwj274.

Cauchemez, S. *et al.* (2011) 'Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), pp. 2825–30. doi: 10.1073/pnas.1008895108.

Cauchemez, S. *et al.* (2016) 'Unraveling the drivers of MERS-CoV transmission.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(32), pp. 9081–6. doi: 10.1073/pnas.1519235113.

Cauchemez, S. and Donnelly, C. (2009) 'Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States', *… England Journal of …*. Available at: http://www.nejm.org/doi/full/10.1056/NEJMoa0905498 (Accessed: 20 November 2015).

Chang, H.-H., Wesolowski, A., Sinha, I., Jacob, Christopher G., Mahmud, A., Uddin, D., Zaman, S. I., Hossain, M. A., Faiz, M. A., Ghose, A., *et al.* (2019) 'Mapping imported malaria in Bangladesh using parasite genetic and human mobility data', *eLife*. eLife Sciences Publications, Ltd, 8, p. e43481. doi: 10.7554/elife.43481.

Chis Ster, I., Singh, B. K. and Ferguson, N. M. (2009) 'Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in Great Britain', *Epidemics*, 1(1), pp. 21–34. doi: 10.1016/j.epidem.2008.09.001.

Chiyaka, C., Tatem, A. J., Cohen, J. M., Gething, P. W., Johnston, G., Gosling, R., Laxminarayan, R., *et al.* (2013) 'The stability of malaria elimination', *Science*, pp. 909–910. doi: 10.1126/science.1229509.

Chuquiyauri, R. *et al.* (2012) 'Socio-demographics and the development of malaria elimination strategies in the low transmission setting', *Acta Tropica*, 121(3), pp. 292–302. doi: 10.1016/j.actatropica.2011.11.003.

Churcher, T. S. *et al.* (2014) 'Measuring the path toward malaria elimination', *Science*, 344(6189), pp. 1230–1232. doi: 10.1126/science.1251449.

Cibulskis, R. E. *et al.* (2016) 'Malaria: Global progress 2000 - 2015 and future challenges.', *Infectious* *diseases of poverty*. BioMed Central, 5(1), p. 61. doi: 10.1186/s40249-016-0151-8.

Clements, A. C. *et al.* (2013) 'Further shrinking the malaria map: How can geospatial science help to achieve malaria elimination?', *The Lancet Infectious Diseases*, pp. 709–718. doi: 10.1016/S1473-3099(13)70140-3.

Cohen, J. M. *et al.* (2010) 'How absolute is zero? An evaluation of historical and current definitions of malaria elimination', *Malaria Journal*. BioMed Central, 9(1), p. 213. doi: 10.1186/1475-2875-9-213.

Cohen, J. M. *et al.* (2012) 'Malaria resurgence: a systematic review and assessment of its causes', *Malaria Journal*. BioMed Central, 11(1), p. 122. doi: 10.1186/1475-2875-11-122.

Cori, A. *et al.* (2013) 'A new framework and software to estimate time-varying reproduction numbers during epidemics.', *American journal of epidemiology*, 178(9), pp. 1505–12. doi: 10.1093/aje/kwt133.

Corran, P. *et al.* (2007) 'Serology: a robust indicator of malaria transmission intensity?', *Trends in Parasitology*, 23(12), pp. 575–582. doi: 10.1016/j.pt.2007.08.023.

Cottam, E. M. *et al.* (2008) 'Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus', *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1637), pp. 887–895. doi: 10.1098/rspb.2007.1442.

Cotter, C. *et al.* (2013) 'The changing epidemiology of malaria elimination: new strategies for new challenges', *The Lancet*, 382(9895), pp. 900–911. doi: 10.1016/S0140-6736(13)60310-4.

Cruz Marques, A. (1987) 'Human migration and the spread of malaria in Brazil', *Parasitology Today*. Elsevier Current Trends, 3(6), pp. 166–170. doi: 10.1016/0169-4758(87)90170-0.

Csardi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal*, Complex Sy, p. 1695. Available at: http://igraph.sf.net.

Dalmat, R. *et al.* (2019) 'Use cases for genetic epidemiology in malaria elimination', *Malaria Journal*, p. 163. doi: 10.1186/s12936-019-2784-0.

Dalrymple, U., Mappin, B. and Gething, P. W. (2015) 'Malaria mapping: understanding the global endemicity of falciparum and vivax malaria.', *BMC medicine*. BioMed Central, 13, p. 140. doi: 10.1186/s12916-015-0372-x.

Dewasurendra, R. L. *et al.* (2017) 'Effectiveness of a serological tool to predict malaria transmission intensity in an elimination setting', *BMC Infectious Diseases*, 17(1), p. 49. doi: 10.1186/s12879-016-2164-0.

Dondorp, A. M. *et al.* (2009) 'Artemisinin Resistance in *Plasmodium falciparum* Malaria', *New England Journal of Medicine*. Massachusetts Medical Society , 361(5), pp. 455–467. doi: 10.1056/NEJMoa0808859.

Duane, S. *et al.* (1987) 'Hybrid monte carlo', *Physics letters B*. Available at: http://www.sciencedirect.com/science/article/pii/037026938791197X (Accessed: 31 May 2017).

Dublin, L. I. and Lotka, A. J. (1925) 'On the True Rate of Natural Increase', *Journal of the American Statistical Association*. Taylor & Francis, Ltd.American Statistical Association, 20(151), p. 305. doi: 10.2307/2965517.

Ernst, K. C. *et al.* (2006) 'Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors.', *Malaria journal*. BioMed Central, 5(1), p. 78. doi: 10.1186/1475-2875-5-78.

Eubank, S. *et al.* (2004) 'Modelling disease outbreaks in realistic urban social networks', *Nature*, 429(6988), pp. 180–184. doi: 10.1038/nature02541.

Faye, O. *et al.* (2015) 'Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study', *The Lancet Infectious Diseases*, 15(3), pp. 320–326. doi: 10.1016/S1473-3099(14)71075-8.

Feachem, R. G. A. *et al.* (2019) 'Malaria eradication within a generation: ambitious, achievable, and necessary', *The Lancet*, 394(10203), pp. 1056–1112. doi: 10.1016/S0140-6736(19)31139-0.

Feachem, R. G. A., Phillips, A. A. and Targett, G. rey A. (2009) 'Shrinking the Malaria Map: A Prospectus on Malaria Elimination'. Global Health Science.

Feng, J. *et al.* (2015) 'Analysis of malaria epidemiological characteristics in the people's republic of China, 2004-2013', *American Journal of Tropical Medicine and Hygiene*, 93(2), pp. 293–299. doi: 10.4269/ajtmh.14-0733.

Feng, J. *et al.* (2018) 'Ready for malaria elimination: zero indigenous case reported in the People's Republic of China', *Malaria Journal*. BioMed Central, 17(1), p. 315. doi: 10.1186/s12936-018-2444-9.

Feng, X. Y. *et al.* (2014) 'Surveillance and response to drive the national malaria elimination program', in *Advances in Parasitology*. Academic Press, pp. 81–108. doi: 10.1016/B978-0-12-800869-0.00004-4.

Ferguson, N. M., Donnelly, C. A. and Anderson, R. M. (2001) 'Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain', *Nature*, 413(6855), pp. 542–548. doi: 10.1038/35097116.

Fine, P. E. M. (2003) 'The interval between successive cases of an infectious disease.', *American journal of epidemiology*, 158(11), pp. 1039–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14630599 (Accessed: 12 May 2017).

Fraser, C. (2007) 'Estimating individual and household reproduction numbers in an emerging epidemic.', *PloS one*. Public Library of Science, 2(8), p. e758. doi: 10.1371/journal.pone.0000758.

Garrett-Jones, C. (1964) 'The human blood index of malaria vectors in relation to epidemiological assessment', *Bulletin of the World Health Organization*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2554803/ (Accessed: 19 September 2017).

Gething, P. W. *et al.* (2011) 'Modelling the global constraints of temperature on transmission of Plasmodium falciparum and P. vivax.', *Parasites & vectors*, 4, p. 92. doi: 10.1186/1756-3305-4-92.

Gething, P. W. *et al.* (2014) 'Declining malaria in Africa: improving the measurement of progress', *Malaria Journal*. BioMed Central, 13(1), p. 39. doi: 10.1186/1475-2875-13-39.

Ghani, A. *et al.* (2009) 'The Early Transmission Dynamics of H1N1pdm Influenza in the United Kingdom.', *PLoS currents*, 1, p. RRN1130. doi: 10.1371/currents.RRN1130.

Global Partnership to Roll Back Malaria (2000) *The African Summit to Roll Back Malaria*. Abuja, Nigeria. Available at: http://www.who.int/iris/handle/10665/67815.

Gomez-Rodriguez, M., Balduzzi, D. and Sch??lkopf, B. (2011) 'Uncovering the temporal dynamics of diffusion networks', in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 561–568.

Gomez-Rodriguez, M. G. and Schölkopf, B. (2012) 'Submodular Inference of Diffusion Networks from Multiple Trees', in *ICML '12: Proceedings of the 29th International Conference on Machine Learning*.

Gomez-Rodriguez, M., Leskovec, J. and Krause, A. (2010) 'Inferring networks of diffusion and influence', *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*, 5(4), pp. 1019–1028. doi: 10.1145/1835804.1835933.

Gomez Rodriguez, M. *et al.* (2014) 'Uncovering the structure and temporal dynamics of information propagation'. doi: 10.1017/nws.2014.3.

González, M. C., Hidalgo, C. A. and Barabási, A.-L. L. (2008) 'Understanding individual human mobility patterns.', *Nature*. Nature Publishing Group, 453(7196), pp. 779–82. doi: 10.1038/nature06958.

Green, P. J. (1995) 'Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination', *Biometrika*, 82(4), pp. 711–732. Available at: http://links.jstor.org/sici?sici=0006-3444%28199512%2982%3A4%3C711%3ARJMCMC%3E2.0.CO%3B2-F (Accessed: 11 June 2017).

Greenhouse, B. *et al.* (2018) 'Taking Sharper Pictures of Malaria with CAMERAs: Combined Antibodies to Measure Exposure Recency Assays', *The American Journal of Tropical Medicine and Hygiene*. The American Society of Tropical Medicine and Hygiene, 99(5), pp. 1120–1127. doi: 10.4269/ajtmh.18-0303.

Griffin, J. T. *et al.* (2010) 'Reducing Plasmodium falciparum malaria transmission in Africa: A model-based evaluation of intervention strategies', *PLoS Medicine*. Edited by S. Krishna. Public Library of Science, 7(8), p. e1000324. doi: 10.1371/journal.pmed.1000324.

Griffin, J. T. (2016) 'Is a reproduction number of one a threshold for Plasmodium falciparum malaria elimination?', *Malaria journal*. BioMed Central, 15(1), p. 389. doi: 10.1186/s12936-016-1437-9.

Griffin, J. T. *et al.* (2016) 'Potential for reduction of burden and local elimination of malaria by reducing Plasmodium falciparum malaria transmission: A mathematical modelling study', *The Lancet Infectious Diseases*. Elsevier, 16(4), pp. 465–472. doi: 10.1016/S1473-3099(15)00423-5.

Grillet, M. E. *et al.* (2019) 'Venezuela's humanitarian crisis, resurgence of vector-borne diseases, and implications for spillover in the region', *The Lancet Infectious Diseases*, 19(5), pp. e149–e161. doi: 10.1016/S1473-3099(18)30757-6.

Guyant, P. *et al.* (2015) 'Malaria and the mobile and migrant population in Cambodia: a population movement framework to inform strategies for malaria control and elimination', *Malar J*, 14, p. 252. doi: 10.1186/s12936-015-0773-5.

Hay, S. I. *et al.* (2000) 'Annual Plasmodium falciparum entomological inoculation rates (EIR) across Africa: literature survey, internet access and review', *Trans R Soc Trop Med Hyg.*, 94(2), pp. 113–127. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3204456/pdf/ukmss-36405.pdf (Accessed: 6 August 2017).

Hay, S. I. *et al.* (2010) 'Developing global maps of the dominant anopheles vectors of human malaria', *PLoS Medicine*. Oxford University Press, 7(2), p. e1000209. doi: 10.1371/journal.pmed.1000209.

Hay, S. I. *et al.* (2010) 'Estimating the global clinical burden of Plasmodium falciparum malaria in 2007.', *PLoS medicine*, 7(6), p. e1000290. doi: 10.1371/journal.pmed.1000290.

Hay, S. I., Smith, D. L. and Snow, R. W. (2008) *Measuring malaria endemicity from intense to interrupted transmission*, *The Lancet Infectious Diseases*. doi: 10.1016/S1473-3099(08)70069-0.

Helb, D. A. *et al.* (2015) 'Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(32), pp. E4438–E4447. doi: 10.1073/pnas.1501705112.

Herrera, S. *et al.* (2015) 'Prospects for malaria elimination in Mesoamerica and Hispaniola.', *PLoS neglected tropical diseases*. Public Library of Science, 9(5), p. e0003700. doi: 10.1371/journal.pntd.0003700.

Howes, R. E. *et al.* (2016) 'Global Epidemiology of Plasmodium vivax.', *The American journal of tropical medicine and hygiene*. American Society of Tropical Medicine and Hygiene, pp. 16–0141. doi: 10.4269/ajtmh.16-0141.

Hu, T. *et al.* (2016) 'Shrinking the malaria map in China: Measuring the progress of the National Malaria Elimination Programme', *Infectious Diseases of Poverty*, 5(1). doi: 10.1186/s40249-016-0146-5.

Huang, J. X. *et al.* (2015) 'Spatio-temporal analysis of malaria vectors in national malaria surveillance sites in China', *Parasites and Vectors*. doi: 10.1186/s13071-015-0741-5.

Huber, J. H. *et al.* (2016) 'Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria', *Malaria Journal*. BioMed Central, 15(1), p. 490. doi: 10.1186/s12936-016-1537-6.

Hurvich, C. M. and Tsai, C. L. (1989) 'Regression and time series model selection in small samples', *Biometrika*. Narnia, 76(2), pp. 297–307. doi: 10.1093/biomet/76.2.297.

Jombart, T. *et al.* (2014) 'Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data', *PLoS Computational Biology*. Edited by M. M. Tanaka. Public Library of Science, 10(1), p. e1003457. doi: 10.1371/journal.pcbi.1003457.

Keeling, M. J. *et al.* (2003) 'Modelling vaccination strategies against foot-and-mouth disease', *Nature*, 421(6919), pp. 136–142. doi: 10.1038/nature01343.

Keeling, M. J. and Eames, K. T. D. (2005) 'Networks and epidemic models.', *Journal of the Royal Society, Interface*. The Royal Society, 2(4), pp. 295–307. doi: 10.1098/rsif.2005.0051.

Kempe, D., Kleinberg, J. and Tardos, É. (2003) 'Maximizing the spread of influence through a social network', in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. New York, New York, USA: ACM Press, p. 137. doi: 10.1145/956750.956769.

Khuller, S., Moss, A. and Naor, J. (1999) 'The budgeted maximum coverage problem', *Information Processing Letters*.

Kitchen, S. F. and Boyd, M. F. (1937) 'A Consideration of the Duration of the Intrinsic Incubation Period in Vivax Malaria in Relation to Certain Factors Affecting the Parasites 1', *The American Journal of Tropical Medicine and Hygiene*. The American Society of Tropical Medicine and Hygiene, s1-17(3), pp. 437–444. doi: 10.4269/ajtmh.1937.s1-17.437.

Lai, S. *et al.* (2016) 'Plasmodium falciparum malaria importation from Africa to China and its mortality: an analysis of driving factors', *Scientific Reports*. Nature Publishing Group, 6(1), p. 39524. doi: 10.1038/srep39524.

Lai, S. *et al.* (2017) 'Malaria in China, 2011–2015: An observational study', *Bulletin of the World Health Organization*. World Health Organization, 95(8), pp. 564–573. doi: 10.2471/BLT.17.191668.

Lai, S. *et al.* (2019) 'Changing epidemiology and challenges of malaria in China towards elimination', *Malaria Journal*. BioMed Central, 18(1), p. 107. doi: 10.1186/s12936-019-2736-8.

Li, J., Blakeley, D. and Smith, R. J. (2011) 'The failure of R0.', *Computational and mathematical methods in medicine*, 2011, p. 527610. doi: 10.1155/2011/527610.

Liu, Q. *et al.* (2012) 'Dispersal Range of Anopheles sinensis in Yongcheng City, China by Mark-Release-Recapture Methods', *PLoS ONE*. Edited by J. Pinto. Public Library of Science, 7(11), p. e51209. doi: 10.1371/journal.pone.0051209.

Lowe, R. *et al.* (1975) 'Studies on flight range and survival of An albimanus in El Salvador', *Mosquito News*, pp. 160–168. Available at: https://www.cabdirect.org/cabdirect/abstract/19762900545 (Accessed: 15 August 2019).

Lowe, R. E. and C. E. S. H. J. H. (1974) 'Studies of flight range and survival of Anopheles albimanus Wiedmann in El Salvador I. Dispersal and Survival During the Dry Season', *Mosquito News*, 34(4), pp. 389–393. Available at: https://www.cabdirect.org/cabdirect/abstract/19762900789 (Accessed: 15 August 2019).

Lu, G. *et al.* (2014) 'Malaria outbreaks in China (1990-2013): A systematic review', *Malaria Journal*. doi: 10.1186/1475-2875-13-269.

Lynch, C. and Roper, C. (2011) 'The Transit Phase of Migration: Circulation of Malaria and Its Multidrug-Resistant Forms in Africa', *PLoS Medicine*. Public Library of Science, 8(5), p. e1001040. doi: 10.1371/journal.pmed.1001040.

Macdonald, G. (1952) 'The analysis of equilibrium in malaria.', *Tropical diseases bulletin*, 49(9), pp. 813–29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12995455 (Accessed: 14 August 2017).

MacDonald, G. (1956) 'Epidemiological basis of malaria control', *Bull World Health Org*, (15), pp. 613–626.

MacDonald, G. (1957) *The Epidemiology and Control of Malaria.* London, Oxford Univ. Pr. Available at: https://www.cabdirect.org/cabdirect/abstract/19581000237 (Accessed: 13 June 2017).

Maia, M. F. *et al.* (2018) 'Attractive toxic sugar baits for controlling mosquitoes: a qualitative study in Bagamoyo, Tanzania', *Malaria Journal*. BioMed Central, 17(1), p. 22. doi: 10.1186/s12936-018-2171-2.

Marshall, J. M. *et al.* (2016) 'Key traveller groups of relevance to spatial malaria transmission: a survey of movement patterns in four sub-Saharan African countries', *Malaria Journal*. BioMed Central, 15(1), p. 200. doi: 10.1186/s12936-016-1252-3.

Marshall, J. M. *et al.* (2018) 'Mathematical models of human mobility of relevance to malaria transmission in Africa', *Scientific Reports*. Nature Publishing Group, 8(1), p. 7713. doi:

10.1038/s41598-018-26023-1.

Mbogo, C. M. *et al.* (2003) ‘Spatial and temporal heterogeneity of Anopheles mosquitoes and Plasmodium falciparum transmission along the Kenyan coast’, *American Journal of Tropical Medicine and Hygiene*, 68(6), pp. 734–742. doi: 10.4269/ajtmh.2003.68.734.

Ménard, R. *et al.* (2013) ‘Looking under the skin: the first steps in malarial infection and immunity’, *Nature Reviews Microbiology*. Nature Research, 11(10), pp. 701–712. doi: 10.1038/nrmicro3111.

Mendis, K. *et al.* (2009) ‘From malaria control to eradication: The WHO perspective’, *Tropical Medicine and International Health*. Blackwell Publishing Ltd, pp. 802–809. doi: 10.1111/j.1365-3156.2009.02287.x.

Meyer, S. and Held, L. (2014) ‘Power-law models for infectious disease spread’, *Annals of Applied Statistics*, 8(3), pp. 1612–1639. doi: 10.1214/14-AOAS743.

Millar, S. B. and Cox-Singh, J. (2015) ‘Human infections with Plasmodium knowlesi-zoonotic malaria’, *Clinical Microbiology and Infection*. Elsevier, pp. 640–648. doi: 10.1016/j.cmi.2015.03.017.

Ministerio de Salud El Salvador (MINSAL) (2015) ‘Plan estrategio nacional multisectoral de eliminacion de la malaria en El Salvador, 2016 - 2020’. Available at: http://www.proyectomesoamerica.org:8088/smsp/phocadownload/Institucional/PlanesNacionales/PNMalaria/SLV PN Malaria.pdf (Accessed: 28 September 2017).

Moiroux, N. *et al.* (2012) ‘Changes in anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in benin’, *Journal of Infectious Diseases*. Oxford University Press, 206(10), pp. 1622–1629. doi: 10.1093/infdis/jis565.

Molineaux, L. and Gramiccia, G. (1980) ‘The Garki Project. Research on the epidemiology and control of malaria in the Sudan Savanna of West Africa’, *The Garki Project. Research on the epidemiology and control of malaria in the Sudan Savanna of West Africa.* doi: 10.1016/0035-9203(81)90085-7.

Morelli, M. J. *et al.* (2012) ‘A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data’, *PLoS Computational Biology*. Edited by C. Fraser, 8(11), p. e1002768. doi: 10.1371/journal.pcbi.1002768.

Nagraj, V. *et al.* (2018) ‘epicontacts: Handling, visualisation and analysis of epidemiological contacts’, *F1000Research*, 7, p. 566. doi: 10.12688/f1000research.14492.2.

Nájera, J. A., González-Silva, M. and Alonso, P. L. (2011) ‘Some lessons for the future from the global malaria eradication programme (1955-1969)’, *PLoS Medicine*. Public Library of Science, p. e1000412. doi: 10.1371/journal.pmed.1000412.

Nemhauser, G., Wolsey, L. and Fisher, M. (1978) ‘An analysis of approximations for maximizing submodular set functions—I’, *Mathematical Programming*.

Nishiura, H. *et al.* (2007) 'Estimates of short- and long-term incubation periods of Plasmodium vivax malaria in the Republic of Korea', *Transactions of the Royal Society of Tropical Medicine and Hygiene*. Narnia, 101(4), pp. 338–343. doi: 10.1016/j.trstmh.2006.11.002.

Nkhoma, S. C. *et al.* (2013) 'Population genetic correlates of declining transmission in a human pathogen', *Molecular Ecology*, 22(2), pp. 273–285. doi: 10.1111/mec.12099.

Oesterholt, M. J. a M. *et al.* (2006) 'Spatial and temporal variation in malaria transmission in a low endemicity area in northern Tanzania.', *Malaria journal*, 5, p. 98.

Okell, L. C. *et al.* (2012) 'Factors determining the occurrence of submicroscopic malaria infections and their relevance for control.', *Nature communications*. Nature Publishing Group, 3, p. 1237. doi: 10.1038/ncomms2241.

OpenStreetMap contributors (2017) *No Title*. Available at: https://www.openstreetmap.org/.

Patel, J. C. *et al.* (2014) 'Genetic Evidence of Drug-Resistant Malarial Strain from the Democratic Republic of the Congo Imported to Guatemala', *Emerging Infectious Diseases*, 20(6), pp. 932–940. doi: 10.3201/eid2006.131204.

Perkins, T. A. *et al.* (2015) 'Estimating Drivers of Autochthonous Transmission of Chikungunya Virus in its Invasion of the Americas', *PLoS Currents*. Public Library of Science. doi: 10.1371/currents.outbreaks.a4c7b6ac10e0420b1788c9767946d1fc.

Pindolia, D. K. *et al.* (2012) 'Human movement data for malaria control and elimination strategic planning', *Malaria Journal*. BioMed Central, 11(1), p. 205. doi: 10.1186/1475-2875-11-205.

Pothin, E. *et al.* (2016) 'Estimating malaria transmission intensity from Plasmodium falciparum serological data using antibody density models', *Malaria Journal*. BioMed Central, 15(1), p. 79. doi: 10.1186/s12936-016-1121-0.

R Core Team (2016) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria. Available at: https://www.r-project.org/.

Reiner, R. C. *et al.* (2015) 'Mapping residual transmission for malaria elimination', *eLife*. eLife Sciences Publications Limited, 4(DECEMBER2015), p. e09520. doi: 10.7554/eLife.09520.

Roberts, L. and Enserink, M. (2007) 'Did They Really Say ... Eradication?', *Science*, 318(5856). Available at: http://science.sciencemag.org/content/318/5856/1544.full (Accessed: 6 June 2017).

Rosas-Aguirre, A. *et al.* (2016) 'Epidemiology of Plasmodium vivax Malaria in Peru.', *The American journal of tropical medicine and hygiene*, 95(Suppl 6), pp. 133–144. doi: 10.4269/ajtmh.16-0268.

Ross, R. (1908) 'Prevention of Malaria in Mauritius', *WATERLOW AND SONS LIMITED*.

Ross, R. (1911) ‘Some quantitative studies in epidemiology’, *Nature*, (87), pp. 466–467.

Routledge, I. *et al.* (2018) ‘Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting’, *Nature Communications*. Nature Publishing Group, 9(1), p. 2476. doi: 10.1038/s41467-018-04577-y.

RTS,S Clinical Trials Partnership (2015) ‘Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial’, *The Lancet*, 386(9988), pp. 31–45. doi: 10.1016/S0140-6736(15)60721-8.

Rue, H., Martino, S. and Chopin, N. (2009) ‘Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2), pp. 319–392. doi: 10.1111/j.1467-9868.2008.00700.x.

Ruktanonchai, N. W. *et al.* (2016) ‘Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data’, *PLoS Computational Biology*, 12(4). doi: 10.1371/journal.pcbi.1004846.

Salje, H. *et al.* (2016) ‘How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(47), pp. 13420–13425. doi: 10.1073/pnas.1611391113.

Salje, H., Cummings, D. A. T. and Lessler, J. (2016) ‘Estimating infectious disease transmission distances using the overall distribution of cases’, *Epidemics*, 17, pp. 10–18. doi: 10.1016/j.epidem.2016.10.001.

El Salvador Ministerio de Salud (2011) *Informe de vigilancia y control de enfermedades transmitidas por vectores: Dengue, Malaria, Chagas, y Leishmaniasis.*

Schneider, K. *et al.* (2016) *Malaria Elimination in El Salvador: A Historical and Epidemiological Perspective*. Available at: http://www.path.org/publications/detail.php?i=2644 (Accessed: 25 September 2016).

Sepúlveda, N., Paulino, C. D. and Drakeley, C. (2015) ‘Sample size and power calculations for detecting changes in malaria transmission using antibody seroconversion rate’, *Malaria Journal*. BioMed Central, 14(1), p. 529. doi: 10.1186/s12936-015-1050-3.

Shi, B. *et al.* (2017) ‘Risk assessment of malaria transmission at the border area of China and Myanmar’, *Infectious Diseases of Poverty*. BioMed Central, 6(1), p. 108. doi: 10.1186/s40249-017-0322-2.

Simini, F. *et al.* (2012) ‘A universal model for mobility and migration patterns’, *Nature*. Nature Publishing Group, 484(7392), pp. 96–100. doi: 10.1038/nature10856.

Siv, S. *et al.* (2016) 'Plasmodium vivax Malaria in Cambodia.', *The American journal of tropical medicine and hygiene*, 95(Suppl 6), pp. 97–107. doi: 10.4269/ajtmh.16-0208.

Sluydts, V. *et al.* (2014) 'Spatial clustering and risk factors of malaria infections in Ratanakiri Province, Cambodia.', *Malaria journal*. BioMed Central, 13, p. 387. doi: 10.1186/1475-2875-13-387.

Smith, D. L. *et al.* (2012) 'Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens', *PLoS Pathogens*. Edited by C. E. Chitnis. Harrison and Sons, Ltd, p. e1002588. doi: 10.1371/journal.ppat.1002588.

Smith, D. L. *et al.* (2013) 'A sticky situation: the unexpected stability of malaria elimination.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1623), p. 20120145. doi: 10.1098/rstb.2012.0145.

Snow, J. (1855) 'On the mode of communication of cholera'. Available at: https://books.google.co.uk/books?hl=es&lr=&id=-N0_AAAAcAAJ&oi=fnd&pg=PA1&dq=Snow+J.+On+the+mode+of+communication+of+cholera.+London:+John+Churchill%3B+1855.&ots=mVTgCpFrPX&sig=NOSGQfl6jOG48kPzmLGvIf20Ozc (Accessed: 19 May 2017).

Sokhna, C., Ndiath, M. O. and Rogier, C. (2013) 'The changes in mosquito vector behaviour and the emerging resistance to insecticides will challenge the decline of malaria', *Clinical Microbiology and Infection*, 19(10), pp. 902–907. doi: 10.1111/1469-0691.12314.

Sorto, Ó. R. *et al.* (2015) 'Prevalence and intensity of infection by soil-transmitted helminths and prevalence of malaria among schoolchildren in El Salvador', *Biomedica*, 35(3), pp. 407–418. doi: 10.7705/biomedica.v35i3.2408.

Stan Development Team (2016) 'RStan: the R interface to Stan'. Available at: http://mc-stan.org/.

Sturrock, H. J. W. *et al.* (2013) 'Targeting asymptomatic malaria infections: active surveillance in control and elimination.', *PLoS medicine*. Public Library of Science, 10(6), p. e1001467. doi: 10.1371/journal.pmed.1001467.

Sturrock, H. J.W. *et al.* (2013) 'Targeting Asymptomatic Malaria Infections: Active Surveillance in Control and Elimination', *PLoS Medicine*. Public Library of Science, 10(6), p. e1001467. doi: 10.1371/journal.pmed.1001467.

Sturrock, H. J. W. *et al.* (2016) 'Mapping Malaria Risk in Low Transmission Settings: Challenges and Opportunities', *Trends in Parasitology*. Elsevier, pp. 635–645. doi: 10.1016/j.pt.2016.05.001.

Surjadjaja, C., Surya, A. and Baird, J. K. (2016) 'Epidemiology of Plasmodium vivax in Indonesia.', *The American journal of tropical medicine and hygiene*, 95(69), pp. 121–132. doi: 10.4269/ajtmh.16-0093.

Szmaragd, C. *et al.* (2009) 'A Modeling Framework to Describe the Transmission of Bluetongue Virus within and between Farms in Great Britain', *PLoS ONE*. Edited by S. J. Cornell. Public Library of Science, 4(11), p. e7741. doi: 10.1371/journal.pone.0007741.

Tadesse, F. G. *et al.* (2018) 'The Relative Contribution of Symptomatic and Asymptomatic Plasmodium vivax and Plasmodium falciparum Infections to the Infectious Reservoir in a Low-Endemic Setting in Ethiopia', *Clinical Infectious Diseases*, 66(12), pp. 1883–1891. doi: 10.1093/cid/cix1123.

Tatem, A. J. *et al.* (2010) 'Ranking of elimination feasibility between malaria-endemic countries', *The Lancet*, pp. 1579–1591.

Tatem, A. J. (2017) 'WorldPop, open data for spatial demography', *Scientific Data*, p. 170004. doi: 10.1038/sdata.2017.4.

Taylor, S. J. and Letham, B. (2017) 'Business Time Series Forecasting at Scale', *RDH*. PeerJ Inc., 35(8), pp. 48–90. doi: 10.7287/peerj.preprints.3190v2.

Tessema, S. *et al.* (2019) 'Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa', *eLife*, 8. doi: 10.7554/eLife.43510.

United Nations (2015) *Millennium Development Goals*, *Millenium Development Goals and Beyond 2015*.

Virlogeux, V. *et al.* (2015) 'Estimating the Distribution of the Incubation Periods of Human Avian Influenza A(H7N9) Virus Infections', *American Journal of Epidemiology*. Narnia, 182(8), pp. 723–729. doi: 10.1093/aje/kwv115.

Walker, P. *et al.* (2012) 'Outbreaks of H5N1 in poultry in Thailand: the relative role of poultry production types in sustaining transmission and the impact of active surveillance in control', *Journal of The Royal Society Interface*, 9(73), pp. 1836–1845. doi: 10.1098/rsif.2012.0022.

Walker, P. G. T. *et al.* (2010) 'A Bayesian Approach to Quantifying the Effects of Mass Poultry Vaccination upon the Spatial and Temporal Dynamics of H5N1 in Northern Vietnam', *PLoS Computational Biology*. Edited by L. A. Meyers. Public Library of Science, 6(2), p. e1000683. doi: 10.1371/journal.pcbi.1000683.

Walker, P. G. T. *et al.* (2016) 'Estimating the most efficient allocation of interventions to achieve reductions in Plasmodium falciparum malaria burden and transmission in Africa: a modelling study.', *The Lancet. Global health*. Elsevier, 4(7), pp. e474-84. doi: 10.1016/S2214-109X(16)30073-0.

Wallinga, J. and Teunis, P. (2004) 'Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures', *American Journal of Epidemiology*, 160(6), pp. 509–516. doi: 10.1093/aje/kwh255.

Wang, L., Ermon, S. and Hopcroft, J. E. (2012) 'Feature-Enhanced Probabilistic Models for Diffusion Network Inference', in *Lecture Notes in Computer Science (including subseries Lecture Notes in*

Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 499–514. doi: 10.1007/978-3-642-33486-3_32.

Wang, X. *et al.* (2014) 'Tracing The Largest Seasonal Migration on Earth', *Arxiv*, (1411.0983). Available at: https://arxiv.org/ftp/arxiv/papers/1411/1411.0983.pdf (Accessed: 2 January 2019).

Wangdi, K. *et al.* (2015) 'Cross-Border Malaria: A Major Obstacle for Malaria Elimination', *Advances in Parasitology*. doi: 10.1016/bs.apar.2015.04.002.

Warrell, D. A. and Gilles, H. M. (2002) 'Essential malariology.', *Essential malariology.* Arnold, (Ed. 4). Available at: https://www.cabdirect.org/cabdirect/abstract/20023074578 (Accessed: 13 May 2017).

Watts, D. and Strogatz, S. (1998) 'Collective dynamics of'small-world'networks', *nature*. Available at: http://search.proquest.com/openview/afbb88ac45f3437067fbc694e38687a3/1?pq-origsite=gscholar&cbl=40569 (Accessed: 2 June 2017).

Weiss, D. J. *et al.* (2018) 'A global map of travel time to cities to assess inequalities in accessibility in 2015', *Nature*. Nature Publishing Group, 553(7688), pp. 333–336. doi: 10.1038/nature25181.

Weiss, D. J. *et al.* (2019) 'Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study', *The Lancet*. Elsevier, 394(10195), pp. 322–331. doi: 10.1016/S0140-6736(19)31097-9.

Welch, D., Bansal, S. and Hunter, D. R. (2011) 'Statistical inference to advance network models in epidemiology.', *Epidemics*. NIH Public Access, 3(1), pp. 38–45. doi: 10.1016/j.epidem.2011.01.002.

Wesolowski, A. *et al.* (2012) 'Quantifying the impact of human mobility on malaria.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 338(6104), pp. 267–70. doi: 10.1126/science.1223467.

Wesolowski, A., Taylor, Aimee R., *et al.* (2018) 'Mapping malaria by combining parasite genomic and epidemiologic data', *BMC Medicine*, 16(1), p. 190. doi: 10.1186/s12916-018-1181-9.

White, M. T. *et al.* (2016) 'Variation in relapse frequency and the transmission potential of Plasmodium vivax malaria', *Proceedings of the Royal Society B: Biological Sciences*, 283(1827), p. 20160048. Available at: http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2016.0048.

WHO (2015) 'Guidelines for the treatment of Malaria', *World Health Organization*, 3, pp. 1–313. doi: 10.1016/0035-9203(91)90261-V.

WHO (2017a) 'Malaria prevention works, let's close the gap', *World Health Organization*. World

Health Organization. Available at: https://www.who.int/malaria/publications/atoz/malaria-prevention-works/en/.

WHO (2017b) 2017). *A framework for malaria elimination*. World Health Organization. Available at: https://apps.who.int/iris/bitstream/handle/10665/254761/9789241511988-eng.pdf

WHO (2018a) 'World malaria report 2018', *World Health Organization*. World Health Organization.doi: http://www.who.int/malaria/publications/world-malaria-report-2017/report/en/.

WHO (2018b) *Global report on insecticide resistance in malaria vectors: 2010-2016*, *WHO*. World Health Organization. Available at: https://www.who.int/malaria/publications/atoz/9789241514057/en/

WHO (2019a) *Guidelines for Malaria Vector Control*, *Guidelines for Malaria Vector Control*. World Health Organization. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30844152.

WHO (2019b) *The E-2020 initiative of 21 malaria-eliminating countries, 2019 progress report*, *WHO*. World Health Organization. Available at: https://www.who.int/malaria/publications/atoz/e-2020-progress-report-2019/en/ (Accessed: 17 September 2019).

WHO Global Malaria Progamme (2016) *Eliminating Malaria*. World Health Organization. doi: 10.1080/03071845709419205.

Xia, Z.-G. *et al.* (2014) 'Lessons from Malaria Control to Elimination: Case Study in Hainan and Yunnan Provinces', *Advances in Parasitology*. Academic Press, 86, pp. 47–79. doi: 10.1016/B978-0-12-800869-0.00003-2.

Yalew, W. G. *et al.* (2017) 'Current and cumulative malaria infections in a setting embarking on elimination: Amhara, Ethiopia.', *Malaria journal*, 16(1), p. 242. doi: 10.1186/s12936-017-1884-y.

Yang, D. *et al.* (2017) 'Spatiotemporal epidemic characteristics and risk factor analysis of malaria in Yunnan Province, China', *BMC Public Health*. BioMed Central, 17(1), p. 66. doi: 10.1186/s12889-016-3994-9.

Yang, G. J. *et al.* (2012) 'Malaria surveillance-response strategies in different transmission zones of the People's Republic of China: Preparing for climate change', *Malaria Journal*. doi: 10.1186/1475-2875-11-426.

Yekutil, P. (1980) 'Eradication of infectious diseases. A critical study.', *Contributions to epidemiology and biostatistics.* S. Karger AG, Arnold-Böcklin-Strasse 25, CH-4011, Basel, Switzerland., 2.

Yin, J. H. *et al.* (2014) 'Historical patterns of malaria transmission in China', *Advances in Parasitology*, 86, pp. 1–19. doi: 10.1016/B978-0-12-800869-0.00001-9.

Ypma, R. J. F. *et al.* (2012) ‘Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data’, 279(1728), pp. 444–450. doi: 10.1098/rspb.2011.0913.

Zhang, J. *et al.* (2016) ‘Effectiveness and impact of the cross-border healthcare model as implemented by non-governmental organizations: case study of the malaria control programs by health poverty action on the China-Myanmar border’, *Infectious Diseases of Poverty*. BioMed Central, 5(1), p. 80. doi: 10.1186/s40249-016-0175-0.

Zhang, S. *et al.* (2017) ‘Anopheles Vectors in Mainland China While Approaching Malaria Elimination.’, *Trends in parasitology*. Elsevier, 33(11), pp. 889–900. doi: 10.1016/j.pt.2017.06.010.

Zhang, S. Sen *et al.* (2018) ‘Monitoring of malaria vectors at the China-Myanmar border while approaching malaria elimination’, *Parasites and Vectors*. BioMed Central, 11(1), p. 511. doi: 10.1186/s13071-018-3073-4.

Zhou, S.-S. *et al.* (2015) ‘China’s 1-3-7 surveillance and response strategy for malaria elimination: Is case reporting, investigation and foci response happening according to plan?’, *Infectious Diseases of Poverty*. BioMed Central, 4(1), p. 55. doi: 10.1186/s40249-015-0089-2.

Zhou, Sheng *et al.* (2016) ‘Trends of imported malaria in China 2010-2014: Analysis of surveillance data’, *Malaria Journal*. BioMed Central, 15(1), p. 39. doi: 10.1186/s12936-016-1093-0.

Zhou, Z. J. (1981) ‘The malaria situation in the People’s Republic of China.’, *Bulletin of the World Health Organization*. World Health Organization, 59(6), pp. 931–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6978199 (Accessed: 11 January 2019).

Appendix

Appendix 1: Sensitivity Analysis for Chapter 3

I explored the sensitivity of my approach by varying the threshold likelihood for linking cases, ϵ, and the threshold gain in marginal likelihood used to define the number of edges to create, *K*. We consider several scenarios, illustrated in Figure A1:

Scenario 1: epsilon = 0.01 and tolerance for edges = 0.003

Scenario 2: epsilon = 0.003 and tolerance for edges = 0.003

Scenario 3: epsilon = 0.007 and tolerance for edges = 0.003

Scenario 4: epsilon = 0.007 and tolerance for edges = 0.005

Scenario 5: epsilon= 1e-10, tolerance for edges = 1e-10

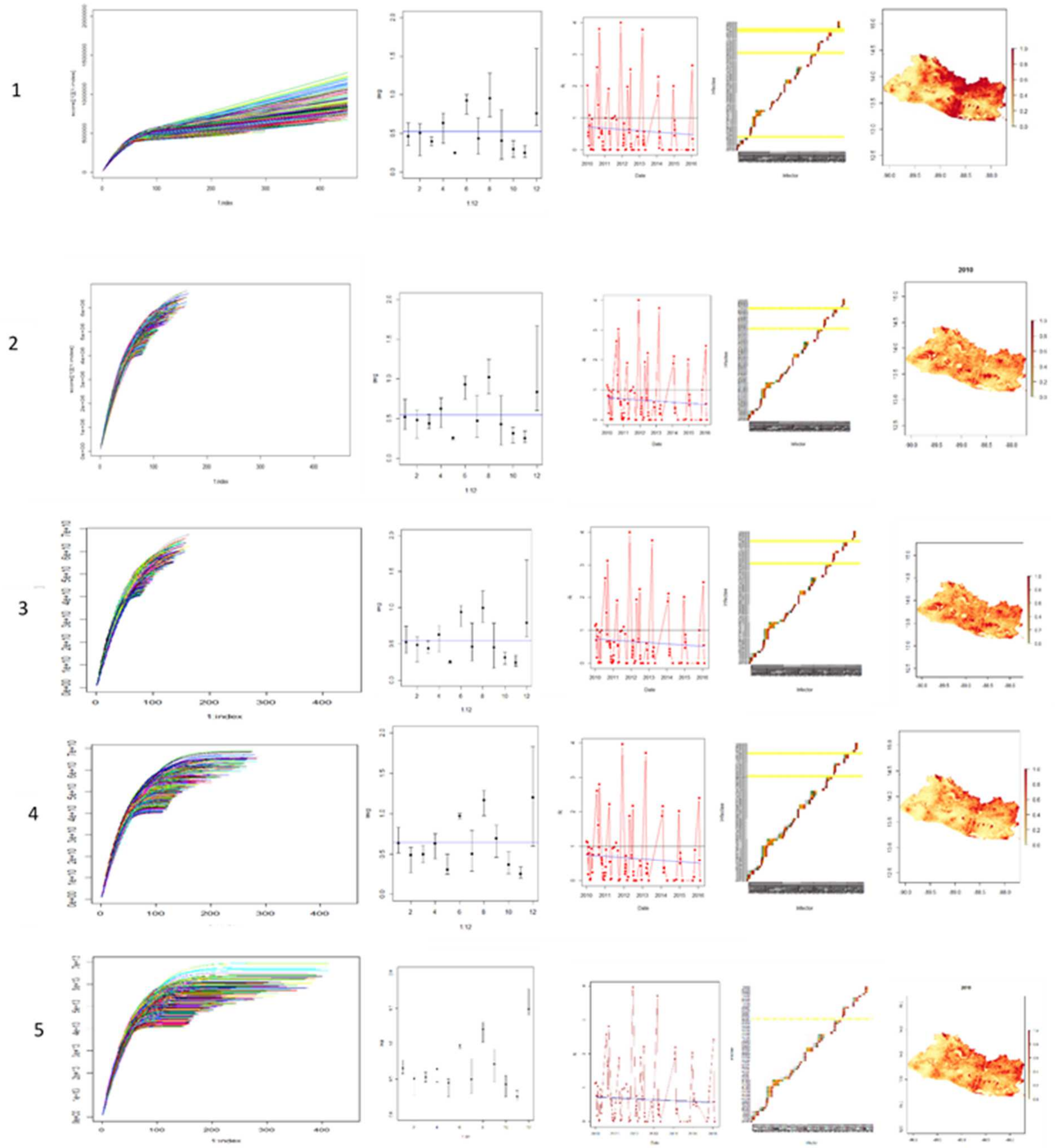
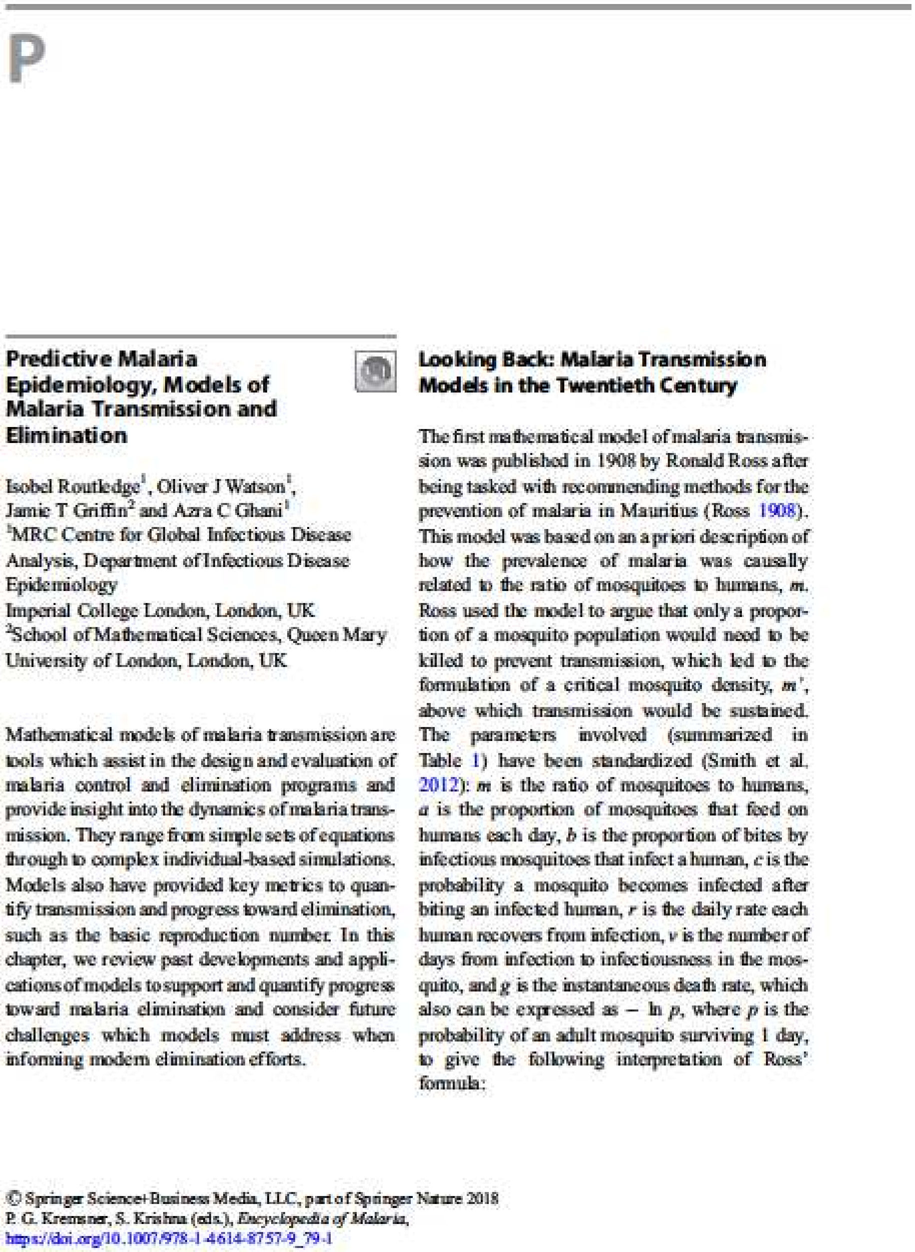


Figure A1 : Each row, numbered 1-5, shows model results for the correspondingly numbered scenarios : Scenario 1: epsilon = 0.01 and tolerance for edges = 0.003, Scenario 2: epsilon = 0.003 and tolerance for edges = 0.003, Scenario 3: epsilon = 0.007 and tolerance for edges = 0.003, Scenario 4: epsilon = 0.007 and tolerance for edges = 0.005, Scenario 5: epsilon= 1e-10, tolerance for edges = 1e-10. Each column shows a different model output A) The marginal gain in tree likelihood from adding edges, B, the estimated Rc *by month. C)* R_c *over time D) A matrix, E) Maps*

Appendix 2: Associated publication, Chapter 1



P

Predictive Malaria Epidemiology, Models of Malaria Transmission and Elimination

Isobel Routledge[1], Oliver J Watson[1], Jamie T Griffin[2] and Azra C Ghani[1]
[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology
Imperial College London, London, UK
[2]School of Mathematical Sciences, Queen Mary University of London, London, UK

Mathematical models of malaria transmission are tools which assist in the design and evaluation of malaria control and elimination programs and provide insight into the dynamics of malaria transmission. They range from simple sets of equations through to complex individual-based simulations. Models also have provided key metrics to quantify transmission and progress toward elimination, such as the basic reproduction number. In this chapter, we review past developments and applications of models to support and quantify progress toward malaria elimination and consider future challenges which models must address when informing modern elimination efforts.

Looking Back: Malaria Transmission Models in the Twentieth Century

The first mathematical model of malaria transmission was published in 1908 by Ronald Ross after being tasked with recommending methods for the prevention of malaria in Mauritius (Ross 1908). This model was based on an a priori description of how the prevalence of malaria was causally related to the ratio of mosquitoes to humans, m. Ross used the model to argue that only a proportion of a mosquito population would need to be killed to prevent transmission, which led to the formulation of a critical mosquito density, m', above which transmission would be sustained. The parameters involved (summarized in Table 1) have been standardized (Smith et al. 2012): m is the ratio of mosquitoes to humans, a is the proportion of mosquitoes that feed on humans each day, b is the proportion of bites by infectious mosquitoes that infect a human, c is the probability a mosquito becomes infected after biting an infected human, r is the daily rate each human recovers from infection, v is the number of days from infection to infectiousness in the mosquito, and g is the instantaneous death rate, which also can be expressed as $-\ln p$, where p is the probability of an adult mosquito surviving 1 day, to give the following interpretation of Ross' formula:

P. G. Kremsner, S. Krishna (eds.), *Encyclopedia of Malaria*,
https://doi.org/10.1007/978-1-4614-8757-9_79-1

Appendix 3: Associated publication, Chapter 3

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04577-y OPEN

Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting

Isobel Routledge[1], José Eduardo Romero Chevéz[2], Zulma M. Cucunubá[1], Manuel Gomez Rodriguez[3], Caterina Guinovart[4], Kyle B. Gustafson[5], Kammerle Schneider[4], Patrick G.T. Walker[1], Azra C. Ghani[1] & Samir Bhatt[1]

In 2016 the World Health Organization identified 21 countries that could eliminate malaria by 2020. Monitoring progress towards this goal requires tracking ongoing transmission. Here we develop methods that estimate individual reproduction numbers and their variation through time and space. Individual reproduction numbers, R_c, describe the state of transmission at a point in time and differ from mean reproduction numbers, which are averages of the number of people infected by a typical case. We assess elimination progress in El Salvador using data for confirmed cases of malaria from 2010 to 2016. Our results demonstrate that whilst the average number of secondary malaria cases was below one (0.61, 95% CI 0.55–0.65), individual reproduction numbers often exceeded one. We estimate a decline in R_c between 2010 and 2016. However we also show that if importation is maintained at the same rate, the country may not achieve malaria elimination by 2020.

[1] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK. [2] Ministry of Health (MINSAL), Calle Arce No.827, San Salvador, El Salvador. [3] Max Planck Institute for Software Systems, E1 5, Campus, 66123 Saarbrücken, Germany. [4] MACEPA, PATH, Seattle, Washington 98121, USA. [5] Institute for Disease Modeling, Bellevue, WA 98005, USA. Correspondence and requests for materials should be addressed to I.R. (email: i.routledge15@imperial.ac.uk)

Appendix 4: Associated papers, Chapter 4

New Results

Tracking progress towards malaria elimination in China: estimates of reproduction numbers and their spatiotemporal variation

Isobel Routledge, Shengjie Lai, Katherine E Battle, Azra C Ghani, Manuel Gomez-Rodriguez, Kyle B Gustafson, Swapnil Mishra, Joshua L Proctor, Andrew J Tatem, Zhongjie Li, Samir Bhatt

doi: https://doi.org/10.1101/628842

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

China reported zero locally-acquired malaria cases in 2017 and 2018. Understanding the spatio-temporal pattern underlying this decline, especially the relationship between locally-acquired and imported cases, can inform efforts to maintain elimination and prevent re-emergence. This is particularly pertinent in Yunnan province, where the potential for local transmission is highest. Using a geo-located individual-level dataset of cases recorded in Yunnan province between 2011 and 2016, we jointly estimate the case reproduction number, R_c, and the number of unobserved sources of infection. We use these estimates within spatio-temporal geostatistical models to map how transmission varied over time and space, estimate the timeline to elimination and the risk of resurgence. Our estimates suggest that, maintaining current intervention efforts, Yunnan is unlikely to experience sustained local transmission up to 2020. However, even with a mean R_c of 0.005 projected for the year 2019, locally-acquired cases are possible due to high levels of importation.

Lai *et al. Malar J* (2019) 18:107
https://doi.org/10.1186/s12936-019-2736-8

Malaria Journal

RESEARCH Open Access

Changing epidemiology and challenges of malaria in China towards elimination

Shengjie Lai[1,2,3], Junling Sun[2], Nick W. Ruktanonchai[1,4], Sheng Zhou[2], Jianxing Yu[2,5], Isobel Routledge[6], Liping Wang[2], Yaming Zheng[2], Andrew J. Tatem[1,4] and Zhongjie Li[2*]

Abstract

Background: Historically, malaria had been a widespread disease in China. A national plan was launched in China in 2010, aiming to eliminate malaria by 2020. In 2017, no indigenous cases of malaria were detected in China for the first time. To provide evidence for precise surveillance and response to achieve elimination goal, a comprehensive study is needed to determine the changing epidemiology of malaria and the challenges towards elimination.

Methods: Using malaria surveillance data from 2011 to 2016, an integrated series of analyses was conducted to elucidate the changing epidemiological features of autochthonous and imported malaria, and the spatiotemporal patterns of malaria importation from endemic countries.

Results: From 2011 to 2016, a total of 21,062 malaria cases with 138 deaths were reported, including 91% were imported and 9% were autochthonous. The geographic distribution of local transmission have shrunk dramatically, but there were still more than 10 counties reporting autochthonous cases in 2013–2016, particularly in counties bordering with countries in South-East Asia. The importation from 68 origins countries had an increasing annual trend from Africa but decreasing importation from Southeast Asia. Four distinct communities have been identified in the importation networks with the destinations in China varied by origin and species.

Conclusions: China is on the verge of malaria elimination, but the residual transmission in border regions and the threats of importation from Africa and Southeast Asia are the key challenges to achieve and maintain malaria elimination. Efforts from China are also needed to help malaria control in origin countries and reduce the risk of introduced transmission.

Keywords: Malaria, Epidemiology, Elimination, Importation, China, Africa, Southeast Asia

Background

Plasmodium malaria, transmitted via the bites of female *Anopheles* mosquitoes, is one of the most prevalent parasitic diseases affecting mankind. Although the global malaria burden has fallen from an estimated 239 million cases occurred worldwide in 2010 to 219 million cases in 2017, no significant progress in reducing global malaria cases was made for the first time in the last decade, especially between 2015 and 2017 [1–4]. However, the progress of eliminating malaria in China seems to be encouraging.

Malaria was once widespread in China, with more than 90% population in China were estimated at risk of infection in the 1940s, and it was still highly endemic in China between 1950s and 1970s, with the highest record of 24 million cases reported in 1970 [5, 6]. Due to the widely use of anti-malarial medications, along with the unprecedented socioeconomic changes and urbanization in China, the incidence of malaria decreased gradually from 1980 to 2000, with only 20 cases per one million residents in 2000 [5, 6]. Although the resurgence of malaria occurred in central China between 2001 and 2006 [7, 8], the efforts of intensified control since 2007 resulted in a dip in the number of cases, reducing to less than 6

*Correspondence: lizj@chinacdc.cn
[2] Division of Infectious Disease, Key Laboratory of Surveillance and Early-Warning on Infectious Disease, Chinese Center for Disease Control and Prevention, Beijing, China
Full list of author information is available at the end of the article