Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Statistical approaches to measure heterogeneity in malaria transmission intensity: An epidemiological study on the Tanzania populations

João Torrado Malato

Mestrado em Bioestatística

Dissertação orientada por:

Prof. Doutor Nuno Sepúlveda

Prof.ª Doutora Marília Antunes

2018

"A análise dos dados por si só nunca nos serve de muito. Há que saber contar a sua história."

— *Nuno Sepúlveda*

"Ainda que não pareça, o ótimo é inimigo do bom. Respira fundo, sem medo!"

— *Marília Antunes*

Acknowledgements

I would like to thank everyone who directly and indirectly contributed to the realisation of this Biostatistics MSc thesis. First of all, my supervisor, Nuno for providing me with the opportunity to work, develop, and prepare my project and dissertation in the field of malaria epidemiology and modelling of infectious diseases. Moreover, thank you for all the advises, ideas, and continuous constructive criticism (as well as all the Wednesdays afternoon football games) that have already proven to be valuable teachings to take into the future. I also want to thank my co-supervisor, Marília, for all the Skype pep talks and helpful suggestions during the later stages of the dissertation that kept me motivated and helped to further improve the thesis. I am grateful for the opportunity to prepare this project at two institutions. First, the London School of Hygiene and Tropical Medicine (LSHTM). Particularly acknowledging all members from the Chris Drakeley's group, whom during my time abroad made me feel welcome to work in a specialised and fully dedicated research group. Cheers to you! And second, the Faculty of Sciences at the University of Lisbon (FCUL), that alongside with the Erasmus+ programme allowed me to travel abroad and learn new research fields, amazing organisations, and terrific people.

E agora em português. Quero agradecer a todos aqueles que me apoiaram no projecto e escrita da tese. Um obrigado a toda a minha família, principalmente aos meus pais e irmãos por me ajudarem com as viagens para a frente e para trás ao longo deste ano, e por constantemente zelarem pelo meu bem estar. Um beijo especialmente carinhoso à minha Mãe por todos os telefonemas, cartas e mimos que disponibilizou. Logo que concluir esta tese, será a primeira a quem vou ligar. Aos meus colegas de mestrado um muito obrigado. Especialmente ao Bruno, Heloísa, e Sofia, por me terem ajudado a entrar ná área da estatística ao longo dos dois anos do mestrado; e ao Constantino, por todos os debates, explicações e demonstrações, e por não esperar menos do que o melhor de mim, ao fazer de criticar esta tese um passatempo. Obrigado ainda à minha família alargada e aos mais variados amigos, que se preocuparam e me fizeram companhia ao longo do ano.

Mais que todos, mil obrigados à Dida, por tudo.
Obrigado Formiguita.

João Malato,
Lisboa, Outubro de 2018.

Resumo

A malária (paludismo) é uma doença infecciosa, reconhecida, nas últimas décadas, como sendo um dos maiores desafios para a saúde pública. A malária é endémica em grande parte da Africa Subsaariana, Sudeste asiático e América Latina, estimando-se cerca de 216 milhões de novos casos de infeção e 445 000 mortes, só em 2016. Os estudos produzidos nos últimos anos têm permitido abordagens efetivas para o controlo e para uma eliminação mais eficiente dos seus transmissores (os mosquitos do género *Anopheles*), bem como o desenvolvimento de tratamentos diretos na redução do parasita *Plamodium falciparum*, a espécie mais incidente nos países da Africa Subsaariana. Com estes avanços, a incidência de malária tem gradualmente vindo a ser reduzida, havendo cada vez mais áreas a transitar para estados de pré-eliminação e eliminação. Devido a estes desenvolvimentos, surgem novos desafios para a estimação da incidência dos locais e da intensidade de transmissão. As regiões onde a intensidade de transmissão é mantida baixa produzem infeções assintomáticas, tornando os indivíduos neste estado transmissores "invisíveis" de malária, uma vez que os métodos mais comuns para medir os níveis de malária baseiam-se na deteção objetiva de casos de infeção. Como alternativa a estas medidas, surge a serologia, aplicada em análises sero-epidemiológicas. Este método põe de lado a análise de indivíduos infetados/não infetados, passando a lidar com a exposição/não exposição dos indivíduos ao parasita *P. falciparum* e o desenvolvimento de anticorpos pelo sistema imunitário. A serologia mede e identifica a presença de anticorpos específicos para antigénios do parasita, podendo definir um gradiente para a intensidade de transmissão de uma dada população analisada, mesmo em locais de baixa incidência.

Esta tese teve como objetivo descrever e estimar a intensidade de transmissão do parasita *P. falciparum* de uma amostra estratificada de 5058 indivíduos distribuídos em 21 vilas ao longo do Nordeste da Tanzânia. Estes dados foram originalmente recolhidos e aplicados num estudo de referência para a área de sero-epidemiologia, tendo como intuito estimar a intensidade de transmissão associada às variáveis altitude e precipitação.

Numa primeira abordagem, os principais fatores de risco associados à prevalência de infeção e heterogeneidade nos vários locais foram identificados. Através da construção e seleção do melhor modelo linear generalizado (generalised linear model, GLM), a influência destes determinantes de transmissão foi estudada. Nesta análise, determinantes representativos da altitude, agregado de vilas, grupo étnico, ou grupo etário, demonstraram ter uma influência significativa quando adicionados no modelo, tendo um efeito direto na probabilidade de infeção. O GLM também caracterizou os três determinantes de exposição usados, relativos aos três antigénios estudados ao longo do projeto: merozoite surface protein 1 (MSP1), merozoite surface protein 2 (MSP2) e apical merozoite antigen 1 (AMA1). A presença destes determinantes no modelo demonstrou a sua utilidade como bons indicadores de infeções de malária, aumentando muito a probabilidade de infeção de indivíduos sempre que estavam presentes.

Tendo identificado os anticorpos para os antigénios como uma alternativa ao estado de infeção das populações, a segunda parte da tese aplicou diferentes propostas sero-epidemiológicas para estudar a intensidade de transmissão das diferentes vilas estudadas. Para tal, diferentes modelos catalíticos reversíveis (reverse catalytic models, RCMs) foram propostos. Estes modelos baseiam-se na ideia de que indivíduos transitam entre dois estados serológicos (seronegativo e seropositivo), transitando de um para outro a diferentes taxas de transição, a taxa de seroconversão (seroconversion rate, SCR) e a taxa de serorreversão (seroreversion rate, SRR). A SCR representa a taxa média anual a que indivíduos de uma determinada idade (em anos) passam de seronegativos para seropositivos, após uma infeção. Já a SRR representa a taxa média anual a que indivíduos seropositivos regressam a um estado seronegativo devido ao decaimento gradual dos anticorpos.

Quatro RCMs foram aplicados aos dados serológicos. Um primeiro modelo M_0 assumiu as taxas SCR e SRR como constantes ao longo de todas as idades. Dois modelos consideraram SRR dependente da idade, $M_{1,1}$ e $M_{1,2}$. Os dois modelos assumiram a SCR de cada vila como constante ao longo da sequência de idades e a ocorrência de uma redução de SRR dada uma idade estimada. O modelo $M_{1,2}$ representava uma versão mais restrita, considerando que após a idade de redução, a SRR era igual a zero. Por fim, o RCM M_2 proposto considerava a ocorrência de algum efeito externo (e.g.: campanhas de intervenção e prevenção de malária nos locais estudados) que tenha ocorrido nas últimas décadas, influenciando a intensidade de transmissão. Este modelo assumiu a SRR estimada como constante ao longo dos anos, com uma variação na SCR, acontecendo um número estimado de anos antes da recolha das amostras.

Os resultados deste estudo mostraram que qualquer um dos modelos tem o potencial de descrever a intensidade de transmissão, bem como a seroprevalência das várias vilas estudadas. A análise dos resultados dos diferentes modelos mostrou que as propostas tidas como mais próximas da realidade (modelos $M_{1,2}$ e M_0) foram rejeitadas, na sua maioria, quando comparadas com o modelo de taxas constantes, M_0 (testes de razão de verosimilhanças, valores-p > 0.05). O modelo $M_{1,2}$, com SRR dependente da idade, foi apenas significativo numa minoria de vilas a altitudes intermédias (altitudes entre os 600 e os 1200 metros). A não rejeição da hipótese nula, aquando da comparação com o modelo M_2, demonstrou poucos episódios significativos onde a alteração de intensidade de transmissão foi observável. Este modelo foi apenas significativo em vilas com maiores taxas de transmissão estimadas, a altitudes baixas e intermédias.

Os RCM ainda que sejam modelos específicos para populações infinitas produziram estimativas paramétricas aceitáveis. Uma análise de correlação entre M_0 e $M_{1,2}$ demonstrou que o modelo estatisticamente preferido, tendencialmente subestimou as estimativas de SCR. Esta taxa, um *proxy* da intensidade de transmissão, é geralmente a medida de interesse nas análises sero-epidemiológicas. Situações de baixa intensidade de transmissão, que requerem uma maior precisão das estimativas, devem ter em conta estes resultados dados por M_0. O melhoramento dos modelos $M_{1,1}$ e $M_{1,2}$ poderá trazer novos resultados sobre a importância da SRR na estimação mais precisa das intensidades de transmissão. Já o modelo M_2 continuará a servir como uma ferramenta para controlo e evolução do estado serológico das populações intervencionadas.

Os dois modelos que consideram o efeito ao longo do tempo do sistema imunitário em regiões de malária endémica, $M_{1,1}$ e $M_{1,2}$, foram desenvolvidos paralelamente a esta tese, tendo sido propostos num artigo científico presentemente em avaliação.

Palavras-chave: malária, intensidade de transmissão, heterogeneidade de transmissão, epidemiologia, serologia, seroprevalência, taxa de seroconversão, taxa de seroreversão.

Abstract

To this day, malaria continues to be a worldwide cause of death and disease. With the recent decades bringing insightful research studies, campaigns for control and elimination became more efficient, gradually reducing the *Plasmodium falciparum* parasitic malaria across sub-Saharan African countries. Such actions have resulted in regions of pre-elimination going into elimination stages, where detectable symptomatic infections are almost vestigial. These scenarios may impose a new challenge, as the more usual methodologies do not consider apparent invisible individuals when estimating malaria transmission intensity and prevalence of infection. As a proposed alternative to this question, sero-epidemiology can be used to more accurately perform such inferences.

The objective of this thesis is then to estimate and characterise the *P. falciparum* transmission intensity from a sample of 5058 individuals structured by age groups, from across 21 villages in the Northeast Tanzania, with different prevalence levels. First, the principal transmission determinants influencing the infection heterogeneity were identified. Using the generalised linear models (GLMs) the study revealed the importance of some demographical risk factors when assessing the presence/absence of infection. Determinants such as the age group of the individuals, the altitude of a village – a known proxy of transmission intensity –, or the transect in which the villages were encompassed, were some of the more impactful demographical transmission determinants assessed. The detected antimalarial antibodies for the specific antigens MSP1, MSP2, and AMA1, used throughout this thesis, were also included in the GLMs and showed the importance these exposure determinants have as reasonable indicators of malaria infection. The inference then led to the use of different reverse catalytic models (RCMs), applied solely to the serological data sets of the three antigens collected.

The RCMs assume that individuals transit between two possible serological states (seronegative and seropositive) at distinct rates: seroconversion rate (SCR) and seroreversion rate (SRR). The SCR is the annual average rate by which the individuals of a certain age change from seronegative to seropositive, upon malaria infection. And the SRR is the annual average rate by which seropositive individuals of a certain age return to the seronegative state due to antibody decay. Focusing on different biological and epidemiological proposals – that might present an effect on the annual transitional rates between seronegative and seropositive individuals due to parasite exposure – four RCMs were tested. Model M_0 assumed the seroconversion rate (SCR) and the seroreversion rate (SRR) as constant transition values across all ages. Models $\mathrm{M}_{1,1}$ and $\mathrm{M}_{1,2}$ were built to adjust for the biological effect of gradually developing immunity over time, in a scenario of endemic malaria transmission. Both models assumed SCR as constant over time, with SRR being reduced to a second rate given an estimated cutoff. Model $\mathrm{M}_{1,2}$, a more restrictive version, assumed that the reduced SRR would be equal to zero, with no seropositive individuals transiting into a seronegative state after the age cutoff. Finally, model M_2 proposed the epidemiological effect that some event (e.g.: possible campaigns or interventions to prevent malaria) might have had on transmission intensity. This model assumed SRR as constant rate across all ages, with a change in SCR happening under an estimated cutoff sometime before the sampling.

The results showed that, depending on the antigen, the models could be used to describe the transmission intensity and seroprevalence of the assessed villages. However, the traditional RCM M_0 (transitional rates constant over time) was more often preferred when compared to the age-dependent $M_{1,2}$ (likelihood ratio test, p-values > 0.05). The age-dependent SRR model was only significant when applied to some villages at intermediate altitudes (600 meters to 1200 meters high). The traditional model was also chosen in favour of the model admitting a past change in transmission intensity, M_2, with the epidemiological model only identifying a change in few villages placed at low and intermediate altitudes. Despite the limited information to estimate some of the models' parameters, further analyses demonstrated that the statistically and overall more parsimonious model M_0 produced underestimations in its transitional rates, when compared to the more realistic model $M_{1,2}$. This underestimation could have a negative impact when estimating malaria transmission intensity in low transmission scenarios. Sided with the newly formulated strategies to advance sites in stages of malaria elimination and pre-elimination into eradication, serology serves as tool to more efficaciously measure transmission intensity. The improvement and application of the RCMs $M_{1,1}$ and $M_{1,2}$ could bring more information to the importance of a more precise estimation of the SRR.

Keywords: malaria, transmission intensity, transmission heterogeneity, epidemiology, serological data, seroprevalence, seroconversion rate, seroreversion rate.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Malaria is a parasitic disease described since ancient times, and until this day it continues to be a major health problem. Despite continuous worldwide efforts and investments, malaria is still the principal cause of death and disease caused by a parasite [1], with estimates for 216 million new infection cases and 445 000 deaths globally in 2016 [2]. The term malaria originates from the 18th century italian expression *mala aria*, meaning 'bad air', referring the foul air evaporating from stagnant waters of marshes that used to be thought as the origin of the disease [3]. The real causative agent of malaria was only discovered in 1897, after Ronald Ross identified parasites in a mosquito that had previously fed on an infected patient [4]. This understanding of the parasites' life cycle of development and transmission laid the foundations for specialised and more focused methods for malaria treatment and control.

The main carriers of malaria parasites known to affect humans are some species and subspecies of mosquitoes belonging to the genus *Anopheles*. More precisely the female mosquitoes, as they must take blood meals to support the development of successive batches of eggs [5]. Usually biting between sunset and sunrise, the infected mosquito transmits parasites of the genus *Plasmodium* into the human hosts' blood stream, where they travel to the liver to multiply. After five to fifteen days without apparent symptoms, the matured parasites re-emerge into the blood stream, targeting and invading the red blood cells [5]. By the time an infected individual might show the primary symptoms of malaria (such as chills, fever, abdominal discomfort, or muscle and joints aches), the parasites have already multiplied immensely, clogging blood vessels and rupturing blood cells [3]. Later stages of severe malaria usually cause anaemia, hypoglycaemia, acute renal failure, or coma (cerebral malaria), among other symptoms [6]. However, despite the infection, some individuals might remain asymptomatic, not suggesting a case for malaria infection. Nonetheless, if left untreated, severe malaria could be considered fatal in most cases [6].

1.1 Epidemiology of the burden of human malaria

There are four distinct *Plasmodium* parasite species known to infect humans: *P. falciparum*, *P. vivax*, *P. malariae*, and *P. ovale* [2]. Of all, *P. falciparum* is the most prevalent, being the principal cause of malaria morbidity and mortality [7]. Malaria endemicity varies geographically, from Africa to South east Asia to South America [2]. The differences in stability of transmission intensity depend on various environmental and demographical characteristics. Biological traits and preferences from the *Anopheles* mosquitoes are also an influence with implications on human hosts [8, 9]. Their spread can be delimited by climatic determinants such as temperature, altitude, rainfall patterns, or humidity, influencing the mosquitoes' activity and abundance [3]. Man made environmental changes like deforestation, extensive irrigation systems, or water dams can also cause transmission intensity to change. The combination of these aspects makes *P. falciparum* endemicity usually found in tropical, subtropical, and temperate regions like the sub-Saharan Africa [3], where it accounts for 99% of the estimated malaria cases in 2016 [2, 5].

The last decade has witnessed a rise in commitment to malaria control in endemic countries [10]. Effective actions in heavy malaria burdened sites have shown success in reducing the parasite's registered morbidity and mortality. Malaria incidence rate, i.e., the number of reported cases per year, has been decreasing globally since 2010 [2]. Fundings for disease control and prevention, although reportedly still less than required [2, 11], allowed for campaigns to include insecticide-treated mosquito nets, insecticide spraying actions, and facilitation of access to curative and preventive antimalarial drugs for millions of people at risk [2]. Amongst the intervened populations, control surveys serve as an important tool, allowing to estimate malaria transmission intensity across different regions. As malaria incidence is gradually reduced, it is required for such measuring approaches to remain as accurate as possible, since all developments made in this field require constant up to date data to keep formulating well informed actions for prevention and control [5].

1.2 Estimating malaria transmission intensity

1.2.1 Conventional measures to estimate transmission intensity

Transmission intensity is the frequency with which people living in an area are bitten by the infected *Anopheles* mosquitoes [12]. Campaigns for control and elimination require knowledge and stratification of malaria transmission intensity amongst the intervened populations [5, 6]. Based on the parasites' life cycle and its influencing determinants, several approaches can be used to estimate such transmission rate. Currently, the most used measures for malaria transmission focus on counting the number of detectable cases of infection. These measures are usually based on active case detection, where infected individuals are identified by active searches for infected patients, or passive case detection, where symptomatic patients come into health facilities seeking care for their illness [5, 13]. In both situations, analyses are only performed on individuals presenting symptoms that evidence a possible case of infection [5]. Measures such as the parasite rate (PR) – also known as prevalence of infection – for assessing the proportion/prevalence of individuals with blood-stage infections in a community, or the spleen rate, for identifying the prevalence of palpable enlarged spleens due to malaria infection (an effect more commonly observed in first time infected individuals), are examples of possible population-focused approaches. In order to be effective, these measures greatly depend on the diagnosis given by established services from the public, or private health sectors, as well as coordinated community services that are the first line of action for symptom assessment and treatment, by reporting the cases to health facilities [5]. Alternatively, estimation of transmission intensity can be done by studies measuring the density of *Anopheles* mosquitoes near the inspected populations. The proportion of infected mosquitoes in a region positively correlates to the capacity of these insects to transmit malaria within that area. This insect proportion also reflects the number of infected, and potentially infectious, human individuals [6]. The entomological inoculation rate (EIR) measures the number of infective mosquito bites received per person, in a population, over a defined period of time [5].

As interventions expand the number of populations inspected, heterogeneity in transmission intensity across different regions is likely to occur. When assessing transmission rates across different sites, reports given by measures such as PR or EIR allow to identify determinant variables related to the parasites, the mosquitoes, or even the human hosts [6]. The analysis of these potential risk inducing variables can be used to define areas with high transmission rates and act accordingly [6, 14].

Although practical and broadly used, measures dependent on infected individuals present limitations. Due to low transmission rates in certain environments, sampling infected mosquitoes and individuals can be challenging. These environments are characterised by a high number of non symptomatic cases of infection and a residual number of detectable infected mosquito bites. As asymptomatic, the undiagnosed individuals will remain invisible to the health system while still contributing to the cycle of malaria transmission [14]. Sites affected by seasonality that regularly shift between extreme high and low transmissions intensities also present a challenge to obtain accurate results [15, 16]. Regions where malaria incidence has been effectively reduced, or have recently been focused by campaigns, still need to be monitored in order to change interventions from malaria control and elimination to disease eradication. For sites where malaria incidence is currently low, alternative approaches might be favourable when estimating the transmission rates [17]. Serological antibody-based techniques can be used, as follows.

1.2.2 Serology as an epidemiological tool

Serology-based methods inspect the densities of existing antibodies and respective antigens circulating in the serum. Using serology, malaria transmission intensity can be assessed by identifying the levels of specific anti-malarial antibodies produced [17, 18]. Serology allows then to estimate the population level of disease transmission by appraising how a population boosts its immunity as a response to the presence or absence of infection.

Antibodies are specific proteins produced by the immune system, able to recognise and target particular foreign substances, the antigen molecules. During the course of natural infections to malaria parasites, individuals develop specific antibodies against the malarial antigens. With multiple episodes of infection over time, a protective immunity will gradually build up, reducing manifestations of severe disease [7]. Because this process of achieving effective protection takes time, the antimalarial immunity in malaria endemic countries is said to be 'age-dependent' [19]. As the immune system reacts to the presence of malaria parasites, the identification of specific antibodies in serologic tests reflects the cumulative (age-dependent) exposure to multiple infections over time [20]. Blood samples taken at a certain time point can provide information about whether or not the individual has been infected before that time point [21]. This ability allows serology to function as a proxy measure of historical malaria transmission, even in low transmission settings.

When applied in epidemiological studies, serological methods shift the focus away from epidemiological measures based on infection. The differentiation across multiple sites provides a better source of information than active or passive case detection that usually inspect only those who appear suspected of being infected, with possible biased results or inaccurate representative cases [22]. Serology has increasingly been incorporated in cross-sectional and longitudinal studies to monitor recent population changes in transmission intensity [23, 24, 25] and evaluate effectiveness of malaria eradication efforts [26].

The antibodies produced upon exposure to malaria parasites belong to the acquired immune system. The specific antimalarial antibodies, contrarily to some diseases such as some forms of the hepatitis virus, the mumps, or the rubella virus, wane over time in absence of infection. This means malaria does not cause long-lasting immunity. After a prolonged interval without reinfection, immunologically protected individuals can revert to an naïve status and once again become vulnerable to show symptoms. However, in malaria endemic sites, individuals might be exposed to a somewhat constant rate of infection from an early age. In those scenarios, the acquired immunity, i.e., the gradual learning of the immune system upon multiple exposures, grants antibody persistence due to continuous exposure over a long period time. In cases of endemic malaria, data from a single cross-sectional survey can be used to generate a point estimate of the current disease transmission intensity. The measure can also analyse potential historical changes in transmission intensity that led to a variation in exposure to the parasite [21].

1.3 Northeast Tanzania as a serological benchmark study

Serology and sero-epidemiological studies have already been assessed as good alternatives to analyse situations of malaria in stages of pre-elimination and elimination [17]. A benchmark example is the study whose data set is used throughout this thesis [27]. The study in Northeast Tanzania applied serology-based methods to analyse cohorts of patients from 24 distinct villages with varying intensities of transmission. Inferences made about the sites' seroprevalence – measure for the proportion of *P. falciparum*-specific antigen seropositive individuals detected in each community – allowed researchers to describe malaria transmission intensity from the different villages as function of altitude and estimated rainfall, confirming both variables to have a measurable impact on the disease's force of infection, i.e., the rate at which non immune, susceptible individuals become infected.

Following the described study, several published articles used and improved the sero-epidemiological methodologies and inferences. Based on the same data set alone, studies of methods and approaches in various research fields were developed. Some examples are studies on the serological analyses, inquiring about the trends in malaria endemicity [18], genetic studies on populations exposed to *P. falciparum* parasites [28, 29], and development of specific mathematical model, used for serological analyses [30].

1.4 Current challenges on malaria epidemiology

The multidisciplinary investment to control and aid populations hurt by the endemic malaria burden is visible [2]. Nowadays, severe malaria develops only in a minority of sites as effective campaigns have been able to control and reduce disease transmission intensity substantially [31]. Low transmission settings are now registered across various regions [23]. All measures presented here estimate malaria transmission intensity on the human population. However, none of them is a perfect indicator. Prevalence of malaria presents a characteristic pattern of increasing with age in young children under five years old, only to then decline throughout adolescence and adulthood (as individuals develop protective immunity). This age-dependent fluctuation is defined as 'peak-shift' and can be difficult to estimate, making approaches such as PR or EIR poor indexes of transmission intensity over time. These measures can be used as good alternatives to estimate recent infections. For serological analyses, some infections may be treated and clear before an immune response develops. This possible lack of immunity development, as well as waning immune responses from the malarial antibodies can affect the accuracy of seroprevalence.

1.5 Objectives and outline

Three *P. falciparum*-specific antigens were measured and analysed in order to estimate malaria transmission intensity: the merozoite surface protein 1 (MSP1), the merozoite surface protein 2 (MSP2), and the apical membrane antigen 1 (AMA1). The corresponding antibodies are known to not confer effective protection against malaria. Instead, they are used as serological markers due to their immunogenic profile, meaning they are expected to be detected and indicate exposure to malaria even in low transmission intensity settings [32, 33, 34].

To analyse the Tanzania data set, different statistical approaches were applied. Using infection and serological samples from the different *P. falciparum* antigens, one expects to identify the principal determinants influencing the prevalence of infection, as well as estimate seroconversion rate – average rate at which individuals become positive for the antimalarial antibodies, upon exposure to *P. falciparum* parasites. Based on the aphorism that all models are wrong but some are useful [35], different statistical models were fit to the data. First, by making use of the generalised linear models and more commonly used statistical approaches to study the detected infection cases and prevalence of infection. Afterwards, the seroprevalence was studied by applying specific serological models to the different known antigens.

As a thesis in Biostatistics, this project was structured in a way that would focus different academic objectives, while maintaining a coherent line of thought. With its fundaments in the matters of malaria and malaria sero-epidemiology, this project granted the opportunity to:

- Work with a well recorded cross-sectional data, used in renowned sero-epidemiological studies;
- Build different generalised linear models to characterise the risk factors and study effects causing heterogeneity in transmission intensity levels;
- Make use of the specific reverse catalytic models to corroborate the previous heterogeneity inferences, and create serological profiles for different villages based on different biological and epidemiological assumptions;
- Describe and propose an innovative extension of the more broadly used reverse catalytic models;
- Study the implications of applying different models to the same data set by changing between epidemiological strategies.

The following chapter will specify the situation of malaria in the two studied regions of the Northeast Tanzania, as well as introduce, and describe the collected data (Chapter 2). The statistical theory used is described further ahead (Chapter 3). The models' analyses and inferences are then presented, firstly focusing the infection status of each individual as the outcome of interest. Afterwards, by applying the alternative statistical methodologies onto the serological outcomes (Chapters 4 and 5, respectively). Lastly, the different methodologies and results are discussed, attending the statistical and epidemiological backgrounds of this project (Chapter 6).

Chapter 2

Study design and description

Including the island of Zanzibar, Tanzania has almost 42 million inhabitants. Tanzania is an East African country with moderate malaria endemicity [36]. Throughout the past decades, Tanzania has made important progresses towards malaria control [37]. From early 2000 until 2010, initiatives provided campaigns and interventions such as distribution of bed nets for mosquito control and improvement of diagnostics and treatment that allowed the country to gradually reduce its proportion of communities living in areas of intense transmission [38]. Recent data suggests almost 60% of the population currently lives in low transmission settings [38]. The data used throughout this project were collected between 2001 and 2002 as part of a program investigating the burden of malaria and its transmission intensity across 24 villages in Tanzania [27].

2.1 Tanzania data set

2.1.1 Study sites on Northeast Tanzania

Within the Northeast Tanzania, the study encompassed the Kilimanjaro region, an inland area with the highest mountain in Africa, and the Tanga region, with the Usambara mountains and coastal plains near the Indian Ocean. Six transects were delimited: Rombo, North, and South Pare in the Kilimanjaro region, and West Usambara 1, 2, and 3 in the Tanga region. Each transect containing four villages located at different altitudes. One at high altitude (>1200m), two at intermediate altitude (600m−1200m), and one the closest to a low altitude (<600m) (Figure 2.1). All transects comprised nearby areas with varying transmission intensity, representing increasing prevalence from high to low altitude. Each village was measured for its mean altitude, daily mean temperature, and rainfall estimates, derived from meteorological stations across both regions.

2.1.2 Cross-sectional survey

Two cross-sectional surveys were conducted in each village. A first one after the short rainy season in November of 2001, and a second during the following year in June, after the long rains. In addition to the local geographical data, for each one of the 24 villages the study objectives were to collect clinical and anthropometric data, as well as blood samples from a total of approximately 250 inhabitants, all structured by age: 80 with ages between 0−4 years old, 80 who were between 5−14 years old, and 90 who were between 15−45 years old. The corresponding populational sampling proportions were then approximately 30%, 30%, and 40%. Attention so that the sexes kept the same ratio across villages was also taken in account, being achieved in the younger age groups, although approximately 70% of the 15−45 years old group surveyed were women. Within each transect a selection criteria was taken to minimise differences in ethnicity. This granted a dominant ethnic group between villages in the same transect, reducing the genetic diversity across those geographically closer sites. Seasonal migration, and access to health care were also points taken into account during the population sampling.

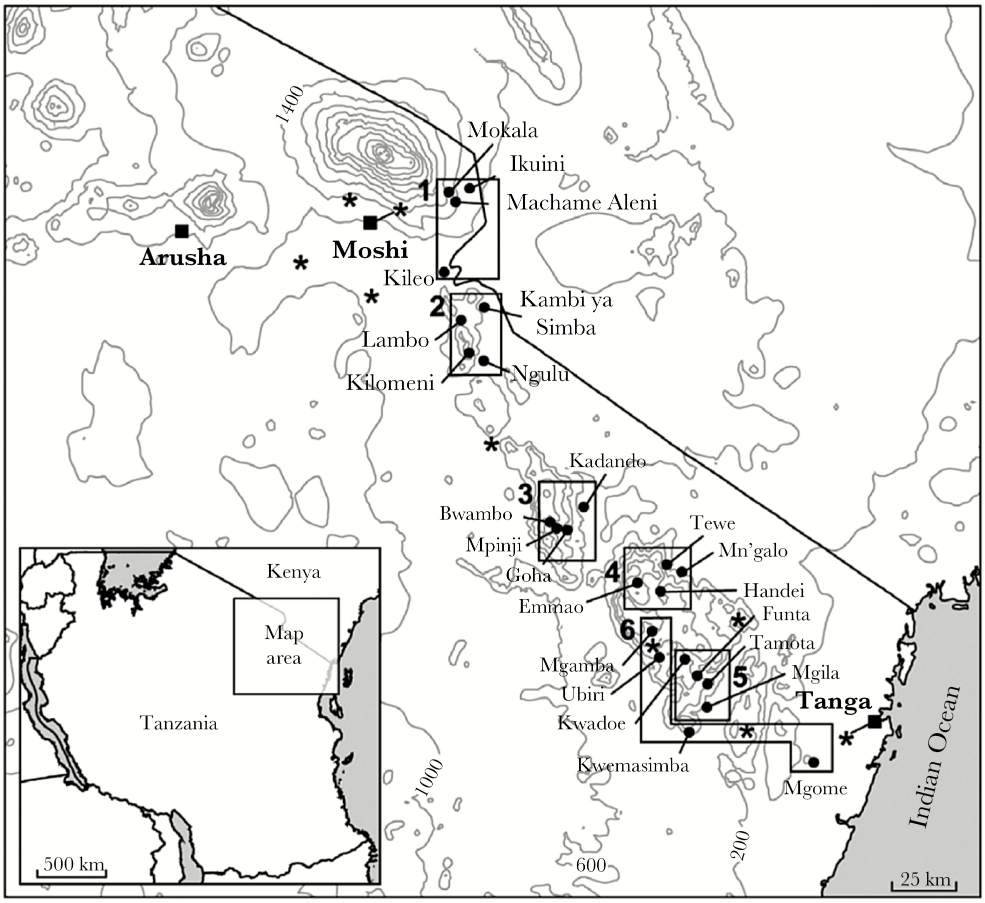

Figure 2.1: Map of study sites (black circles) grouped by respective numbered region transects: Rombo (transect 1), North Pale (transect 2), South Pale (transect 3), West Usambara 1 (transect 4), West Usambara 2 (transect 5), and West Usambara 3 (transect 6). Locations of the 8 meteorological stations are also shown (asterisks). Reprinted with permission [29].

2.2 Variables description

To study transmission intensity amongst Tanzania's different populations, several variables were selected to be analysed in this thesis. Some variables refer the demographical information collected regarding each site. For each village, its mean altitude and encompassing transect were registered. Variables that characterise the selected inhabitants were also used. Firstly, the infection outcome recorded for each individual. This variable has value 0 in absence of infection and 1 when presence of malaria parasites is found, indicating the assumed true individual status. Other binary variables selected were the gender of each individual (as female or male), and the three *P. falciparum* antigens (either present or absent), used here as serological markers to identify exposure to parasite instead of immunological protection. Since each antigen has different response levels to the presence of *P. falciparum* parasites, the comparison of their serological outcomes under the same study characteristics could help to better understand how transmission intensity varies across the different villages. Ethnicities were also recorded, with four possible ethnic groups: Wachaga, Wapare, Wasambaa, and Other. As mentioned in the previous section, the ethnic groups of each village are expected to represent the genetic variability seen in each transect. In other words, in each transect an ethnic group dominated the others in proportion. The age in years of every individual was also a recorded variable, as well as the respective age group each one belonged to, with three possible categories, 1−4, 5−14, and 15−45.

Depending on the study objective, more than one of the described variables can be classified as the desired response variable, with others being used as explanatory variables and potential risk factors. An example of response variable can be the infection or serological status of each individual. The binary outcomes of these variables can be used to calculate the number of infected or previously exposed individuals within each village, attending other particular variables that may characterise such response. All ten variables considered for the analyses are listed below.

Altitude: Mean altitude of each village, in meters;

Transect: Transect of villages (categorical variable with six possible outcomes, Rombo, North Pare, South Pare, West Usambara 1, West Usambara 2, and West Usambara 3);

Gender: Gender of each individual (binary variable with possible outcomes, Female and Male);

EthGp: Ethnic group of each individual (categorical variable with four possible outcomes, Wachaga, Wapare, Wasambaa, and Other);

AgeYears: Age of each individual, in years;

AgeGp: Age group of each individual (categorical variable with three possible outcomes, 1−4, 5−14, and 15−45);

Infection: Infection status of each individual (binary variable with possible outcomes, 0 = infection is absent, and 1 = *P. falciparum* malaria detected);

MSP1: Serological status regarding the antigen MSP1 (binary variable with possible outcomes, 0 = antigen absent in serum, and 1 = antigen detected);

MSP2: Serological status regarding the antigen MSP2 (binary variable with possible outcomes, 0 = antigen absent in serum, and 1 = antigen detected);

AMA1: Serological status regarding the antigen AMA1 (binary variable with possible outcomes, 0 = antigen absent in serum, and 1 = antigen detected).

2.3 Exploratory analysis

The decision to perform analyses using only the complete cases for the described variables led to the exclusion of three villages from the West Usambara 3 transect (the villages removed were Magamba, Ubiri and Kwemasimba), where serological samples were not collected. In this transect the remaining village, Mgome, is a coastal village with the lowest recorded mean altitude, at 179 meters. By comparison, the first transect of the Usambara mountains registers the village located at highest altitude, Emmao, at 1780 meters high. The resulting total population sample size is 5058 individuals, separated across the 21 villages (Table 2.1). Considering each village, the sample sizes ranged from 53 (Ngulu) up to 379 individuals (Handei). Even with different sample sizes it was possible to verify the attention taken by the study to maintain a similar structure for ages and genders, as these variables presented similar proportions across all sites. Similar attention was also taken for the different ethnic groups within each one of the six transects. Each transect had a dominant ethnicity. Rombo was represented by the Wachaga ethnic group, having only one village (Kileo) with Wapare as the main ethnicity, North and South Pare had mostly individuals from the Wapare ethnic group, both West Usambara 1 and 2 had Wasambaa as the main ethnic group, and the last transect, West Usambara 3, represented a mixture of different ethnicities (Other).

Table 2.1: Descriptive table with the recorded variables for each one of the 21 selected villages within six different transects. For all villages, values for mean altitude, population sample size, gender proportion, mean age, and proportion of each ethnic group are presented.

Transect	Village	Altitude, m	n	Female proportion, % (n)	Mean age, years	Ethnic group, % (n)			
						Wachaga	Wapare	Wasambaa	Other
Rombo	Mokala	1703	291	62.20 (181)	15.34	98.28 (286)	0.00 (0)	0.00 (0)	1.72 (5)
	Machame Aleni	1422	225	55.11 (124)	15.28	100.00 (225)	0.00 (0)	0.00 (0)	0.00 (0)
	Ikuini	1160	256	57.03 (146)	14.99	98.83 (253)	0.39 (1)	0.39 (1)	0.39 (1)
	Kileo	723	223	61.88 (138)	14.94	3.14 (7)	86.55 (193)	5.38 (12)	4.93 (11)
N. Pare	Kilomeni	1556	101	55.45 (56)	17.71	0.99 (1)	96.04 (97)	0.99 (1)	1.98 (2)
	Lambo	1188	131	64.89 (85)	15.43	0.00 (0)	94.66 (124)	1.53 (2)	3.82 (5)
	Ngulu	832	53	52.83 (28)	13.75	1.89 (1)	96.23 (51)	1.89 (1)	0.00 (0)
	Kambi ya Simba	746	93	50.54 (47)	15.43	6.45 (6)	81.72 (76)	0.00 (0)	11.83 (11)
S. Pare	Bwambo	1598	240	54.17 (130)	16.61	1.25 (3)	98.33 (236	0.00 (0)	0.42 (1)
	Mpinji	1445	198	63.13 (125)	14.03	0.51 (1)	93.94 (186)	2.02 (4)	3.54 (7)
	Goha	1163	337	58.75 (198)	14.53	0.00 (0)	96.44 (325)	2.97 (10)	0.59 (2)
	Kadando	528	281	59.07 (166)	16.19	1.42 (4)	69.40 (195)	18.51 (52)	10.68 (30)
W. Usamb. 1	Emmao	1780	170	61.18 (104)	16.04	0.00 (0)	15.29 (26)	60.00 (102)	24.71 (42)
	Handei	1368	379	56.46 (214)	14.21	0.00 (0)	2.11 (8)	94.20 (357)	3.69 (14)
	Tewe	999	326	61.96 (202)	15.68	0.31 (1)	3.07 (10)	93.56 (305)	3.07 (10)
	Mn'galo	389	363	58.95 (214)	15.58	0.00 (0)	1.10 (4)	89.53 (325)	9.37 (34)
W. Usamb. 2	Kwadoe	1564	296	62.16 (184)	15.14	0.68 (2)	2.03 (6)	94.59 (280)	2.70 (8)
	Funta	1240	252	66.67 (168)	15.90	0.40 (1)	0.40 (1)	96.43 (243)	2.78 (7)
	Tamota	1055	330	53.94 (178)	15.62	0.61 (2)	1.21 (4)	94.55 (312)	3.64 (12)
	Mgila	375	288	71.88 (207)	15.62	0.00 (0)	9.72 (28)	67.36 (194)	22.92 (66)
W. Usamb. 3	Mgome	179	225	54.22 (122)	15.46	0.44 (1)	5.33 (12)	9.78 (22)	84.44 (190)

The combination of characteristics from each site may have different impacts in the resulting prevalence of infection, changing the seroprevalence values as consequence (Table 2.2). The overall prevalence of infection was reported as 19.81%, with seroprevalence values for MSP1, MSP2, and AMA1 being 36.22%, 40.25%, and 46.56%, respectively. Analysing the different categories showed that *Gender* did not seem to pose an effect on neither prevalence nor seroprevalence, as both estimates did not change much from female individuals to males. Prevalence of infection across ethnic groups (variable *EthGp*), showing different registered sample sizes, suggested mixed ethnicities ($EthGp_{Other}$) suffered the most from detectable cases, presenting higher seroprevalence levels as well.

Since individuals are born almost immunologically unprotected, with only maternal antibodies inherited from the mother, the first age group ($AgeGp_{1-4}$) showed lower seroprevalence values. Contrarily, the last age group ($AgeGp_{15-45}$) had a reduced prevalence of infection, with over 50% of this older population showing presence of the various antimalarial antigens. In this particular scenario, the difference in prevalence and seroprevalence between age groups might be justified by the gradual accumulation of protective antibodies due to exposure over time. The primary risk group in these African populations is then children younger than 5 years old, who have yet to develop an efficient immune system [9]. Pregnant woman can also be a vulnerable group to malaria infections, as pregnancy affects the immune system's defences against the disease [7, 8]. This influence of age on prevalence of infection and seroprevalence can be seen as the previously described peak-shift. As malaria does not produce long-lasting immunity, in a case of transmission intensity reduction (or eradication), the developed (and accumulated) antibodies for the specific antigens could gradually decay, with the immunologically protected individuals becoming susceptible to high and symptomatic cases of infection, once again.

Table 2.2: Prevalence and seroprevalence of each one of the specific *P. falciparum* antigens, estimated for the categorised available variables. Each prevalence and seroprevalence shows the 95% confidence interval (CI), estimated using the Wald interval.

Variables	Categories	n	Prevalence, % (95% CI)	Seroprevalence, % (95% CI)		
				MSP1	MSP2	AMA1
AgeGp	1−4	1261	19.59 (17.40, 21.78)	19.83 (17.63, 22.03)	24.27 (21.90, 26.63)	30.61 (28.07, 33.15)
	5−14	1789	25.88 (23.85, 27.91)	30.58 (28.44, 32.71)	39.18 (36.92, 41.45)	45.78 (43.47, 48.09)
	15−45	2008	14.54 (13.00, 16.08)	51.54 (49.36, 53.73)	51.25 (49.06, 53.43)	57.27 (55.11, 59.43)
Gender	Female	3017	18.99 (17.59, 20.39)	38.88 (37.14, 40.62)	42.46 (40.70, 44.22)	47.99 (46.21, 49.78)
	Male	2041	21.02 (19.25, 22.79)	32.29 (30.26, 34.32)	36.99 (34.90, 39.09)	44.44 (42.28, 46.59)
EthGp	Wachaga	794	6.68 (4.94, 8.41)	20.03 (17.24, 22.81)	4.28 (2.87, 5.69)	7.68 (5.83, 9.54)
	Wapare	1583	9.85 (8.39, 11.32)	34.30 (31.96, 36.64)	39.48 (37.07, 41.89)	47.38 (44.92, 49.84)
	Wasambaa	2223	28.39 (26.51, 30.26)	38.46 (36.44, 40.48)	48.13 (46.06, 50.21)	55.29 (53.22, 57.35)
	Other	458	35.37 (30.99, 39.75)	60.04 (55.56, 64.53)	67.03 (62.73, 71.34)	68.78 (64.53, 73.02)
Transect	Rombo	995	6.73 (5.18, 8.29)	31.26 (28.38, 34.14)	4.82 (3.49, 6.16)	10.35 (8.46, 12.24)
	N. Pare	378	4.76 (2.62, 6.91)	28.31 (23.77, 32.85)	38.36 (33.46, 43.26)	54.23 (49.21, 59.26)
	S. Pare	1056	11.74 (9.80, 13.68)	29.64 (26.89, 32.39)	45.17 (42.17, 48.17)	49.91 (46.89, 52.92)
	W. Usamb. 1	1238	32.31 (29.71, 34.92)	29.08 (26.55, 31.61)	40.39 (37.65, 43.12)	65.75 (63.11, 68.39)
	W. Usamb. 2	1166	23.93 (21.48, 26.38)	46.83 (43.96, 49.69)	56.69 (53.85, 59.53)	42.97 (40.13, 45.81)
	W. Usamb. 3	225	50.67 (44.13, 57.20)	86.67 (82.22, 91.11)	91.11 (87.39, 94.83)	91.11 (87.39, 94.83)
Overall	–	5058	19.81 (18.71, 20.91)	36.22 (34.90, 37.54)	40.25 (38.90, 41.60)	46.56 (45.19, 47.93)

When analysing the prevalence for each village, the infection outcome varied from 0.89% in Machame Aleni, to 50.67% in Mgome (Table 2.3). Since villages within each transect were ordered by decreasing altitude, it was possible to see an apparent pattern for prevalence of infection. Villages located at higher altitudes seemingly had lower prevalence of infection than villages at lower altitudes. As climate changes with altitude, sites with lower temperature or humidity values (that do not suit the *Anopheles* mosquitoes) tend to present lower prevalence estimates [39]. This impact on maximum reachable prevalence of infection allows altitude to be considered a proxy for malaria transmission intensity.

The sero-epidemiological variables were related to altitude as well, with seroprevalence appearing as a direct response to infection, since the gradient of immunological responses depends on the level of exposure to the parasite. For the same village, the three antigens analysed presented different seroprevalence values, confirming their different immunogenic profiles when close to the same parasites and corroborating the idea that using more than a single antigen could be useful to better understanding the heterogeneity in malaria transmission intensity. Of all, AMA1 appeared the most sensitive to *P. falciparum*, with higher seroprevalence even at low altitude villages, when compared to the remaining two. Seroprevalence estimates for the MSP1 antigen tend to be the lowest, at times presenting similar outcomes as MSP2. All seroprevalence values appear to correlate well with prevalence of infection.

Table 2.3: Values of prevalence and seroprevalence of each one of the specific *P. falciparum* antigens, estimated for each village and their respective 95% confidence interval (CI), estimated using the Wald interval. Proportions are based on the number of individuals in each village, column n from previous Table 2.1.

Transect	Village	Prevalence, % (95% CI)	Seroprevalence, % (95% CI)		
			MSP1	MSP2	AMA1
Rombo	Mokala	5.84 (3.15, 8.54)	15.46 (11.31, 19.62)	3.09 (1.10, 5.08)	6.53 (3.69, 9.37)
	Machame Aleni	0.89 (0.00, 2.12)	24.44 (18.83, 30.06)	3.11 (0.84, 5.38)	4.00 (1.44, 6.56)
	Ikuini	12.89 (8.79, 17.00)	17.97 (13.27, 22.67)	1.95 (0.26, 3.65)	7.81 (4.53, 11.10)
	Kileo	6.73 (3.44, 10.01)	73.99 (68.23, 79.75)	12.11 (7.83, 16.39)	24.66 (19.01, 30.32)
N. Pare	Kilomeni	1.98 (0.00, 4.70)	8.91 (3.35, 14.47)	8.91 (3.35, 14.47)	23.76 (15.46, 32.06)
	Lambo	2.29 (0.00, 4.85)	19.85 (13.02, 26.68)	21.37 (14.35, 28.39)	54.20 (45.67, 62.73)
	Ngulu	3.77 (0.00, 8.90)	52.83 (39.39, 66.27)	81.13 (70.60, 91.67)	84.91 (75.27, 94.54)
	Kambi ya Simba	11.83 (5.26, 18.39)	47.31 (37.16, 57.46)	69.89 (60.57, 79.22)	69.89 (60.57, 79.22)
S. Pare	Bwambo	5.00 (2.24, 7.76)	8.33 (4.84, 11.83)	47.50 (41.18, 53.82)	33.33 (27.37, 39.30)
	Mpinji	2.53 (0.34, 4.71)	6.06 (2.74, 9.38)	55.05 (48.12, 61.98)	44.44 (37.52, 51.37)
	Goha	11.57 (8.16, 14.99)	28.49 (23.67, 33.31)	38.58 (33.38, 43.77)	53.41 (48.09, 58.74)
	Kadando	24.20 (19.19, 29.21)	65.84 (60.29, 71.38)	44.13 (38.32, 49.93)	63.70 (58.08, 69.32)
W. Usamb. 1	Emmao	2.94 (0.40, 5.48)	1.76 (0.00, 3.74)	1.18 (0.00, 2.80)	13.53 (8.39, 18.67)
	Handei	27.97 (23.45, 32.49)	18.21 (14.32, 22.09)	35.36 (30.54, 40.17)	68.07 (63.38, 72.77)
	Tewe	33.74 (28.61, 38.88)	30.06 (25.08, 35.04)	44.17 (38.78, 49.56)	64.42 (59.22, 69.61)
	Mn'galo	49.31 (44.17, 54.45)	52.34 (47.20, 57.48)	60.61 (55.58, 65.63)	88.98 (85.76, 92.20)
W. Usamb. 2	Kwadoe	7.09 (4.17, 10.02)	9.12 (5.84, 12.40)	16.89 (12.62, 21.16)	21.62 (16.93, 26.31)
	Funta	24.60 (19.29, 29.92)	61.51 (55.50, 67.52)	64.68 (58.78, 70.58)	51.19 (45.02, 57.36)
	Tamota	26.06 (21.32, 30.80)	49.09 (43.70, 54.48)	63.03 (57.82, 68.24)	32.42 (27.37, 37.47)
	Mgila	38.19 (32.58, 43.81)	70.14 (64.85, 75.42)	83.33 (79.03, 87.64)	69.79 (64.49, 75.09)
W. Usamb. 2	Mgome	50.67 (44.13, 57.20)	86.67 (82.22, 91.11)	91.11 (87.39, 94.83)	91.11 (87.39, 94.83)

Chapter 3

Statistical methodology

The statistical approaches used throughout the thesis are introduced in this chapter. Section 3.1 describes the sampling distribution. All the different sets of stochastic models applied to the data are introduced in Section 3.2, and finally, the methods used for parameter estimation, the approaches used to estimate confidence intervals, models evaluation, and selection, are described in Section 3.3.

3.1 Sampling distribution

For the data collection in the original study [27], the selected individuals were screened for the presence malaria parasites and three specific *P. falciparum* malarial antigens (MSP1, MSP2, and AMA1). Each individual was recorded as infected or not infected for malaria, and seropositive or seronegative for the different antigens. Within each village, the study design placed all sampled individuals in three distinct age groups with defined ranges $[1, 5)$, $[5, 15)$, and $[15, 46)$, each one representing a specific percentage of the recorded population (example of village structure in Table 3.1). Under this structured data, the objective of the thesis is then to infer on the number of individuals with status infected/seropositive.

Since the total number of individuals placed within each age group $g = (1, 2, 3)$ is known, one can assume that the random vector containing the number of individuals with each combination of characteristics (age $t = (T_{g_{min}}, \ldots, T_{g_{max}})$ in years and infection or seropositivity status, $j = (0, 1)$) follows a Multinomial distribution, where $T_{g_{min}}$ and $T_{g_{max}}$ represent the minimum and the maximum exact ages in age group g. Let $\tilde{X}$ be such vector. For easiness of reading we shall refer to $\tilde{X}$ as $\tilde{X}_{gtj}$ so that track of indexes is kept. Let $\tilde{x}_{gtj}$ be the observed vector and x_{gtj} represent the number of individuals in age group g, exact age t, and status j. The vector $\tilde{x}_{gtj}$ is then presented in Table 3.1 with three groups of rows ($g = (1, 2, 3)$), each with the distribution of the positive and negative cases ($j = (0, 1)$) for all the ages within the age group ($t = (T_{g_{min}}, \ldots, T_{g_{max}})$).

$$\begin{aligned} f(\tilde{x}_{gtj}|\tilde{x}_{g\cdot\cdot}, \tilde{\theta}_{gtj}) &= \prod_{g=1}^{3} \left[\frac{x_{g\cdot\cdot}!}{\prod_{t=1}^{T_g} \prod_{j=0}^{1} x_{gtj}!} \prod_{t=1}^{T_g} \prod_{j=0}^{1} (\theta_{gtj})^{x_{gtj}} \right] \\ &= \prod_{g=1}^{3} \left[\frac{x_{g\cdot\cdot}!}{\prod_{t=1}^{T_g} x_{gt0}!\, x_{gt1}!} \prod_{t=1}^{T_g} (\theta_{gt0})^{x_{gt0}} (\theta_{gt1})^{x_{gt1}} \right] , \end{aligned} \tag{3.1}$$

where x_{gtj} is the number of individuals from group g, with age t, and status outcome j, $x_{g\cdot\cdot}$ is the fixed number of sampled individuals contained in the age group g, and θ_{gtj} is the joint probability of an individual that belongs to age group g with exact age t, being identified with status $j = (0, 1)$.

The following analysis will be focused simply on a single Multinomial distribution from a unspecified age group g, with $t = (1, \ldots, T)$, as it is simpler to analyse and can be later extended. Under this assumption of a unique age group, $x_{gtj} \equiv x_{tj}$, $x_{g\cdot\cdot} \equiv x_{\cdot\cdot}$, and $\theta_{gtj} \equiv \theta_{tj}$. According to the rule of conditional probability, the joint probability of an individual of age t being identified as having status $j = (0, 1)$ can be described as

$$\theta_{t1} = \pi_t \gamma_t \ ,$$

$$\theta_{t0} = (1 - \pi_t)\gamma_t \ ,$$

where γ_t is the probability of a sampled individual having age t, and π_t the probability of an individual of age t being positive for infection or seropositive for the antigens. The sampling distribution for one village can be then decomposed as follows

$$\begin{aligned} f(\tilde{x}_{tj}|x_{\cdot\cdot}, \tilde{\theta}_{tj}) &= \frac{x_{\cdot\cdot}!}{\prod_{t=1}^{T} x_{t0}!\, x_{t1}!} \prod_{t=1}^{T} (\theta_{t0})^{x_{t0}} (\theta_{t1})^{x_{t1}} \\ &= \frac{x_{\cdot\cdot}!}{\prod_{t=1}^{T} x_{t0}!\, x_{t1}!} \prod_{t=1}^{T} [(1 - \pi_t)\gamma_t]^{x_{t0}} \ (\pi_t \gamma_t)^{x_{t1}} \\ &= \frac{x_{\cdot\cdot}!}{\prod_{t=1}^{T} x_{t0}!\, x_{t1}!} \prod_{t=1}^{T} (1 - \pi_t)^{x_{t0}} \ \pi_t^{x_{t1}} \ \gamma_t^{x_{t0}+x_{t1}} \ . \end{aligned} \tag{3.2}$$

From the equation one can identify the marginal frequency of all individuals with age t, $x_{t\cdot}$, represented by the expression $x_{t0} + x_{t1}$. This resulting marginal statistics is characterised by a Multinomial distribution, depending only on the total number of individuals within the group, $x_{\cdot\cdot}$, and parameter γ_t . Since its distribution does not depend on π_t, $x_{t\cdot}$ is an ancillary statistics for this interest parameter, being a sufficient statistics for γ_t [40].

Under these considerations, one can infer about parameter π_t in a way that the Multinomial distribution for the number of infection/seropositive outcomes in age t (x_{t1}) is simplified, without losing its information.

$$
\begin{aligned}
f\left(\tilde{x}_{t1}|\tilde{x}_{t\cdot},\tilde{\gamma}_t,\tilde{\pi}_t\right) &= \frac{f(\tilde{x}_{t1},\tilde{x}_{t\cdot})}{f(\tilde{x}_{t\cdot}|x_{\cdot\cdot},\tilde{\gamma}_t)} \\
&= \frac{\left(\frac{x_{\cdot\cdot}!}{\prod_{t=1}^{T}(x_{t\cdot}-x_{t1})!x_{t1}!}\right)\prod_{t=1}^{T}(1-\pi_t)^{x_{t\cdot}-x_{t1}}\ \pi_t^{x_{t1}}\ \gamma_t^{x_{t\cdot}}}{\left(\frac{x_{\cdot\cdot}!}{\prod_{t=1}^{T}x_{t\cdot}!}\right)\prod_{t=1}^{T}\gamma_t^{x_{t\cdot}}} \\
&= \prod_{t=1}^{T}\left(\frac{x_{t\cdot}!}{(x_{t\cdot}-x_{t1})!\,x_{t1}!}\right)(1-\pi_t)^{x_{t\cdot}-x_{t1}}\ \pi_t^{x_{t1}} \\
&= \prod_{t=1}^{T}\binom{x_{t\cdot}}{x_{t1}}(1-\pi_t)^{x_{t\cdot}-x_{t1}}\ \pi_t^{x_{t1}}\ .
\end{aligned}
\tag{3.3}
$$

The frequency of infection/seropositive cases, for each age value t in years has then a Binomial distribution. The final sampling distribution for the number of Bernoulli trials for positive or seropositive cases within each age, considering all independent villages studied, $k = (1,\ldots,K)$, can be written using the sequence of marginal frequencies for all ages $t = (1,\ldots,T)$, where T is the maximum age recorded for each village, $m_t = x_{t1}$, and $n_t = x_{t\cdot}$. Let M_{kt} represent the number of infected/seropositive individuals amongst the sampled, at village k, and with age t. Once again, to facilitate the reading, $\tilde{M} = \{M_{kt}\}$ will be represented as $\tilde{M}_{kt}$ so that track of indexes is kept. Then, $\tilde{M}_{kt}$ is a vector of size $K \times T$. Each M_{kt} follows a Binomial distribution with parameters n_{kt} for the total number of sampled individuals with age t from village k, and π_{kt} for the probability of and individual with age t, living at village k, being positive for *P. falciparum* infection or seropositive for its antigen inspected. The resulting sampling distribution formula is then

$$
f(\tilde{m}_{kt}|\tilde{n}_{kt},\tilde{\pi}_{kt}) = \prod_{k=1}^{K}\prod_{t=1}^{T}\binom{n_{kt}}{m_{kt}}\pi_{kt}^{m_{kt}}(1-\pi_{kt})^{n_{kt}-m_{kt}}\ . \tag{3.4}
$$

This sampling distribution allows to disregard the initially defined age groups and work solely with the individuals of each age t, with prevalence/seroprevalence being assessed through the proportion values for each t.

Table 3.1: Frequency table of infection status for all sampled individuals in the Bwambo village, from the South Pare transect, ordered by age in years and age group. For each age group $[1, 5)$, $[5, 15)$, and $[15, 46)$, independent and with Multinomial distribution, individuals were selected respecting the $30 : 30 : 40$ ratio. For the village sample size of 396 individuals, each age group $g = (1, 2, 3)$, has a known $n_{g\cdot\cdot}$, with $n_{1\cdot\cdot} = 92$, $n_{2\cdot\cdot} = 151$, and $n_{3\cdot\cdot} = 153$. Within each age group g, selected individuals were then registered for their age, and screened for presence/absence of malaria parasites. Each age group, has then a fixed number of frequency columns $J = 2$, and specific number of T_g rows, $T_1 = 4$, $T_2 = 10$, and $T_3 = 31$.

Age group,	Age, t	Frequency, j		$n_{g\cdot\cdot}$
g	in years	0	1	
1	1	22	1	
	2	22	0	
	3	27	1	
	4	19	0	92
2	5	13	0	
	6	13	2	
	7	13	1	
	8	9	0	
	9	13	1	
	10	11	0	
	11	14	1	
	12	19	1	
	13	14	1	
	14	13	1	151
3	15	10	1	
	16	9	0	
	17	3	0	
	18	4	0	
	19	5	0	
	20	4	0	
	21	3	0	
	22	1	0	
	23	4	0	
	24	6	0	
	25	5	0	
	26	8	0	
	27	6	0	
	28	7	0	
	29	5	0	
	30	5	1	
	31	2	0	
	32	3	0	
	33	3	0	
	34	9	0	
	35	7	0	
	36	10	0	
	37	4	0	
	38	6	0	
	39	4	0	
	40	3	1	
	41	3	1	
	42	8	0	
	43	6	0	
	44	5	0	
	45	3	0	153

3.2 Statistical models to analyse data

The first set of models used were the generalised linear models (GLMs). This extension of the linear models was applied to study the prevalence of infection amongst the sampled individuals. From the results, one can identify the principal determinants whose effects may influence transmission intensity. By studying the characteristics of different sites, this approach allows to better understand the patterns causing transmission heterogeneity. The GLMs have been used extensively in epidemiology due to their flexibility [41, 42], pairing well with standard measures such as the previously described parasite rate.

After the GLMs, the class of the reverse catalytic models (RCMs) was applied to the serological data. These specific stochastic models serve as an alternative for the study of infections, and have been used in low transmission settings [17]. Here, the RCMs were used to describe situations of infectious diseases assuming the developed antibodies do not last an individuals' life. Both modelling approaches play an important role in better understanding the different mechanisms of disease transmission rates and its effects on the analysed populations.

3.2.1 Generalised linear models to infer on infection determinants

For a more simplistic theory description, a single unspecified village will be focused. Thus, the random variable M_t describes the frequency of positively infected individuals with age t in the village, among the sampled. This Binomial distribution is included in the exponential family of distributions, with parameters n_t and π_t, describing the total number of sampled villagers with age t and the probability of a individual of that age being positive, respectively. Under this assumption one can make use of the GLMs. A GLM is characterised by three distinct components: a random component, a systematic component, and a link function.

For binary outcomes such as the infection status of an individual, the random component identifies the response variables from each individual $i = (1, \ldots, n)$ as Bernoulli trials for the infection status outcome of each individual. The systematic component defines a linear combination of the explanatory variables. Both random and systematic components are related via a link function. Usually, this function links the expected value of the response variable, $E(M_i)$, with the systematic component of the model. For a set of binary response variables, rather than directly modelling the dependence of the expected value, the link function explores how the probability of infection, $\pi_i = E(\frac{M_i}{n_i})$, can be described by the observed explanatory variables in the systematic component. Given the formula

$$g(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \ , \tag{3.5}$$

where $g(\cdot)$ is the link function that associates both random and systematic components, $\beta_0, \ldots, \beta_k$ are the unknown coefficient parameters, with $\beta_1, \ldots, \beta_k$ being the regression parameters associated with covariates $x_{i1}, \ldots, x_{ik}$, respectively, and β_0 representing the intercept as the overall effect when all the categorical explanatory variables are set to their reference level and the continuous variable is set to zero (i.e. the mean effect in the absence of covariates). The link function typically transforms the probability of range $[0,1]$ to a value in $(-\infty,+\infty)$.

There are different possible link functions. The link functions used throughout the thesis are the logistic, the probit, and the complementary log-log, described bellow. In the case of binary response variables with Binomial distribution, the most used is the logistic link function,

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \ . \tag{3.6}$$

This transformation is called the canonical link function of the Binomial model, as it is the natural parameter of the Binomial exponential family. Since $\frac{\pi_i}{1-\pi_i}$ is the odds of an individual i being infected, the logistic, or logit transformation is then the log-odds of this event.

Considering a resulting logistic GLM, the odds of success can be described by $\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}$, where $\widehat{\pi}_i$ is the estimated probability of an individual i being infected, under the model characteristics. The odds of success from this set of specific characteristics can give the odds ratio (OR) of a single level of a covariate by relating the odds of success of individual i with the odds of success of a different individual j, given as

$$\widehat{\mathrm{OR}} = \frac{\widehat{\pi}_i/(1-\widehat{\pi}_i)}{\widehat{\pi}_j/(1-\widehat{\pi}_j)} = e^{\widehat{\beta}_k} \ , \tag{3.7}$$

where $\widehat{\pi}_j$ is the probability of individual j being infected, knowing he/she has the same exact characteristics as individual i, only changing the level of one of the explanatory variables included in the model, say the one associated with parameter β_k. From this ratio, values of OR < 1 suggest that the odds of malaria infection are bigger for individual j (denominator from equation (3.7)) than the risk of the inferred individual i. Values of OR > 1 suggest the opposite, with the odds of infection for the analysed individual with defined characteristics being higher. Values close to OR $= 1$ indicate both individuals have similar odds of being infected. Based on estimates for the parameters and standard errors, the confidence intervals for the true odds ratio of each level of the covariates can be estimated by exponentiation the limits of the estimated confidence interval for the corresponding β [43].

Other known link functions used are the probit link function,

$$g(\pi_i) = \Phi^{-1}(\pi_i) \ , \tag{3.8}$$

that uses an inverse normal link function, and the complementary log-log, or cloglog link function

$$g(\pi_i) = \log[-\log(1-\pi_i)] \ . \tag{3.9}$$

The GLMs take into consideration the variables described in Chapter 2, as well as potential relationships amongst them, better recreating a rational representation of the real world situation in which the data were collected.

3.2.2 Specific models for estimating malaria transmission intensity

When modelling serological data, individuals are generally assumed to be born seronegative. Upon exposure to malaria parasites, individuals might become seropositive by producing antibodies to deal with the infection. Since malaria does not induce long lasting immunity, seropositive individuals can later revert into a seronegative state in the absence of continued exposure and recurrent infections. This process of sequential change between two serological states can be described by the reverse catalytic models (RCMs). The RCMs are mathematically formulated as a two state Markov Chain, where individuals transit between the seronegative and seropositive states (Figure 3.1). The transitions from one state to the other occur at specific age-dependent rates that allow to quantify the levels of parasite exposure and malaria transmission intensity [44].

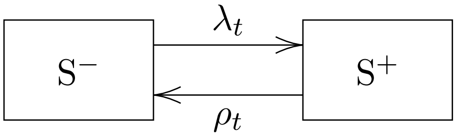

Figure 3.1: Schematic representation of the reverse catalytic model, where individuals transit between seronegative (S$^-$) and seropositive (S$^+$) states with age-dependent rates λ_t (seroconversion rate) and ρ_t (seroreversion rate).

The seroconversion rate (SCR or λ_t) is the annual average rate by which individuals with age t change from seronegative to seropositive, upon malaria infection. This rate is directly related to transmission intensity and has shown to correlate well with popular epidemiological measures of malaria transmission such as parasite and entomological inoculation rate [17, 18, 45]. Seroreversion rate (SRR or ρ_t) is the annual average rate by which seropositive individuals with age t return to the seronegative state due to antibody decay. This rate can be influenced by individual characteristics such as age or genetics [17], and be used to predict how many seropositive individuals are expected to remain after complete interruption of transmission [17].

When using RCMs to estimate malaria transmission intensity, one should assume the lower age limit to be $t = 1$ year old. While all individuals are usually assumed to be born seronegative ($\pi_0 = 0$), when dealing with responses to malaria parasites, the new born's immune system may be influenced by the presence of inherited maternal antibodies. These antibodies eventually wane over the first months of the newborn's life. By removing the first year from the analyses (with $t > 0$), the influence of the maternal antibodies in the model's estimates is expected to be reduced.

Malaria under stable transmission intensity

The simplest RCM assumes both SCR and SRR to be constant over time: $\lambda_t = \lambda$ and $\rho_t = \rho$ (Figure 3.2A). In this situation, all individuals experience equal risk of exposure at all times. The expected seroprevalence of individuals with age t explained by this model, henceforward denoted M$_0$, is given by

$$\pi_{t|\lambda,\rho} = \frac{\lambda}{\lambda+\rho}\left(1 - e^{-(\lambda+\rho)t}\right) \ , \tag{3.10}$$

where $\lambda \in \mathbb{R}_0^+$ and $\rho \in \mathbb{R}_0^+$. Both transition rates are expected to be positive, although the possibilities for $\lambda = 0$ and $\rho = 0$ are included to allow for some particular cases. A null SCR value can happen if, by any change, a population has experienced transmission interruption. In that situation any model will predict seroprevalence equal to zero. A null SRR indicates that all individuals who are exposed and become seropositive will remain so throughout their remaining life. Model M_0 is an increasing function of age that tends exponentially to a plateau given by $\frac{\lambda}{\lambda+\rho}$, when $t \to \infty$. The derivation of the equation can be found in Section A from Appendices.

While not common to use this model in situations of null SCR, it is worth mention that there are situations where SRR can become a rare event. When this happens, by considering $\rho \to 0$ the model can be rewritten as a traditional complementary log-log (or cloglog) model, with resulting formula

$$\log[-\log(1-\pi_t)] = \log\lambda + \log t \ . \tag{3.11}$$

However, this cloglog model is not usually applied in malaria research, though it has been used to study scenarios where a unique exposure to a disease develops immunisation with resulting permanent seropositive state [21].

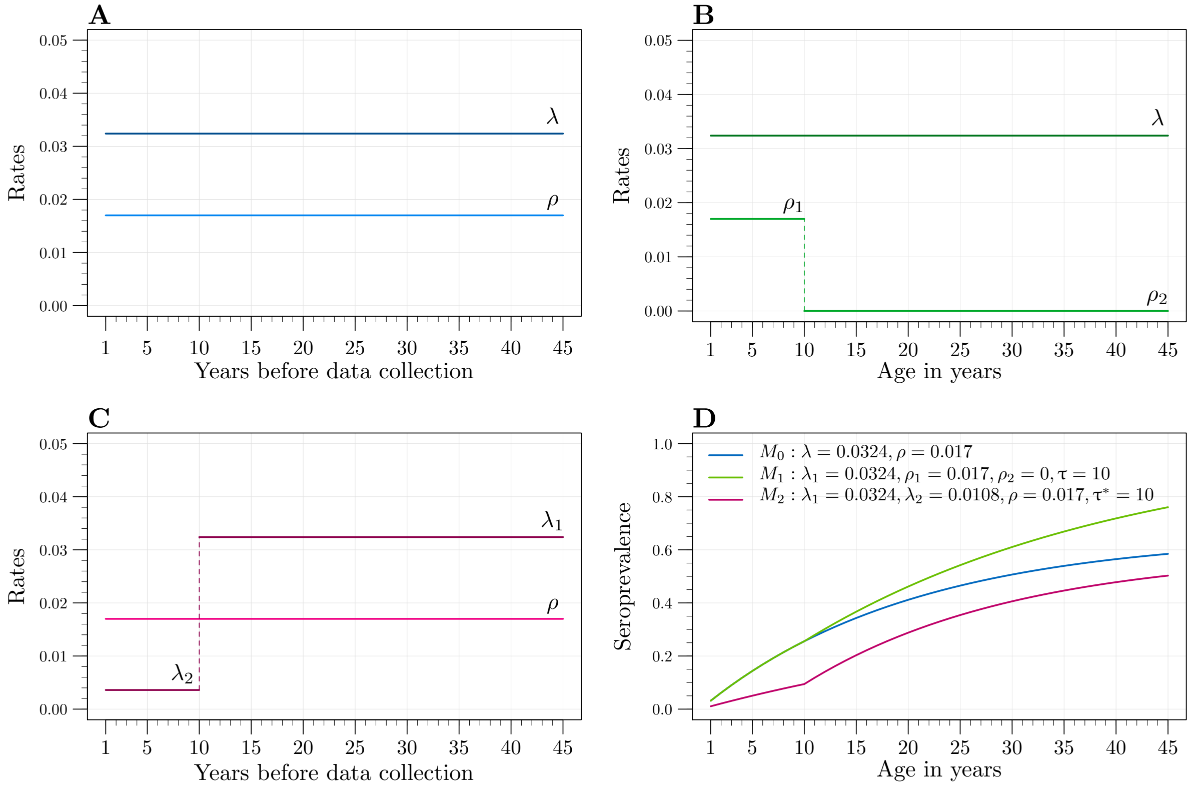

Figure 3.2: Graphical representation of the SCR and SRR for each one of the three RCMs described. (**A**) M_0 with both parameters stable and constant over time; (**B**) M_1 (more specifically model $\mathrm{M}_{1,2}$) that assumes SCR to be constant for all ages and SRR to abruptly decrease to a lower value given a change point $\tau = 10$; and (**C**) M_2 that assumes SCR to change its value after an age cutoff $\tau^* = 10$ while SRR remains constant over time. Plot (**D**) illustrates the resulting age-dependent seroprevalence calculated for each one of the models.

RCM assuming age-dependent rates to detect heterogeneity in malaria transmission intensity

From a biological point of view, the previous model can be restrictive. Even in a situation of constant transmission intensity (and thus constant SCR), exposed individuals will eventually develop specific antibodies from an early age, changing SRR over time [24]. From an epidemiological perspective, considering both rates to be fixed is also limiting. Model M_0 does not account for past actions for control and elimination against the disease that may have occurred, causing SCR to change [23, 46]. To better represent these two scenarios further mathematical models must be assessed.

In a situation of endemic populations, individuals are expected to gradually develop specific immunity over multiple episodes of infection, throughout their lives [7]. As more individuals become seropositive and remain there due to the constant transmission intensity, over time, less will revert into a seronegative state. Modelling this effect of acquired immunity should then consider a change in SRR as a function of age. Mathematically, this age-dependent SRR reduction can be simplified by modelling all individuals with an initial parameter ρ_1 that abruptly changes to a different parameter ρ_2 after a specific age cutoff τ. The change occurs while considering $\lambda_t = \lambda$ (Figure 3.2B). The age cutoff when such change occurs should vary inversely to the transmission intensity, as individuals exposed to high levels of parasite rate are expected to develop specific immunologic responses earlier in life (standard relation between SCR and expected change in SRR presented in Table 1 from Appendices). The resulting expected seroprevalence for individuals with age t is explained by model $\mathrm{M}_{1,1}$, with formula

$$\pi_{t|\lambda,\rho_1,\rho_2,\tau} = \begin{cases} \frac{\lambda}{\lambda+\rho_2}\left(1 - e^{-(\lambda+\rho_2)(t-\tau)}\right) + \frac{\lambda}{\lambda+\rho_1}\left(1 - e^{-(\lambda+\rho_1)\tau}\right) e^{-(\lambda+\rho_2)(t-\tau)} \ , & \text{if } t > \tau \\ \frac{\lambda}{\lambda+\rho_1}\left(1 - e^{-(\lambda+\rho_1)t}\right) \ , & \text{if } t \leq \tau \ , \end{cases} \tag{3.12}$$

where $\lambda \in \mathbb{R}_0^+$ is the SCR constant over time, ρ_1 the initial SRR that changes to ρ_2 given the cutoff τ. Similar to M_0, this function for seroprevalence increases exponentially.

The derivation of this model is based on a RCM presented by Sepúlveda et al. [46]. The original model was used to detect an increase in SCR from children and adolescents to adults, due to age-dependent behaviours. The model proved useful in populations where the male adults go to work on sites that are malaria transmission hotspots, as opposed to younger individuals who stay within the malaria protected housing regions. For instance, some mining populations in the Pará state, near the Brazilian Amazonia [47]. Having a similar structure, model $\mathrm{M}_{1,1}$ was created considering a change in SRR instead, assuming SCR to be constant throughout the years. However, to accurately represent the effect of acquired immunity, the model requires the restriction $\rho_1 \geq \rho_2$, where $\rho_1 \in \mathbb{R}_0^+$ (similar to ρ in M_0) and $\rho_2 \in [0, \rho_1]$.

In scenarios of endemic malaria with stable SCR, the constant exposure throughout an individual's life results in a gradual development in specific immunity. This also indicates the whole population will eventually become seropositive, with SRR reduced to nearly zero by the time an individual reaches adulthood [19]. Based on equation (3.12), this scenario can be represented by an abrupt reduction from $\rho_1 \in \mathbb{R}_0^+$, to $\rho_2 = 0$, after the age cutoff. With this model (hereafter denoted $\mathrm{M}_{1,2}$) one expects that after a certain age, all seropositive individuals will remain so, with resulting expected seroprevalence closer to 1, as age increases.

When considering effective interventions for control of endemic malaria, one expects to infer a noticeable reduction in malaria transmission intensity some time before the sampling [23]. Mathematically, this reduction can be represented by admitting SCR as a function of time that changes from λ_1 to λ_2, given a cutoff τ^*, and $\rho_t = \rho$ (Figure 3.2C) [46]. The resulting seroprevalence in this model (hereafter labelled M_2) is described by

$$\pi_{t|\lambda_1,\lambda_2,\rho,\tau^*} = \begin{cases} \frac{\lambda_2}{\lambda_2+\rho}\left(1 - e^{-(\lambda_2+\rho)\tau^*}\right) + \frac{\lambda_1}{\lambda_1+\rho}\left(1 - e^{-(\lambda_1+\rho)(t-\tau^*)}\right) e^{-(\lambda_2+\rho)\tau^*} \ , & \text{if } t > \tau^* \\ \frac{\lambda_2}{\lambda_2+\rho}\left(1 - e^{-(\lambda_2+\rho)t}\right) \ , & \text{if } t \leq \tau^* \ , \end{cases} \tag{3.13}$$

where λ_1 is the initial SCR parameter that abruptly changes to λ_2 following the age cutoff τ^*, and $\rho \in \mathbb{R}_0^+$ is the constant and stable SRR. Model M_2 is published in Sepúlveda et al. Section 2.3.1 [46]. When considering successful interventions, a reduction in malaria transmission intensity is assumed, with corresponding reduction in SCR. With this change in mind, one can expect a reduction after the cutoff through $\lambda_1 \geq \lambda_2$, where $\lambda_1 \in \mathbb{R}_0^+$ and $\lambda_2 \in [0, \lambda_1]$.

All variations of the RCMs create slightly different age-dependent seroprevalence curves with influence on the maximum reachable plateaus (Figure 3.2D). Considering the simpler model M_0 as reference for the curve analysis, the biphasic behaviour of SRR from M_1 appears to not reflect a visible effect on the expected seroprevalence as does M_2 considering variation in SCR. Model M_0 is nested within the remaining RCMs, as is model $\mathrm{M}_{1,2}$ in $\mathrm{M}_{1,1}$. This relation means the latter models can be transformed into M_0 by imposing certain parametric constrains (Figure 3.3). This facilitates model comparisons.

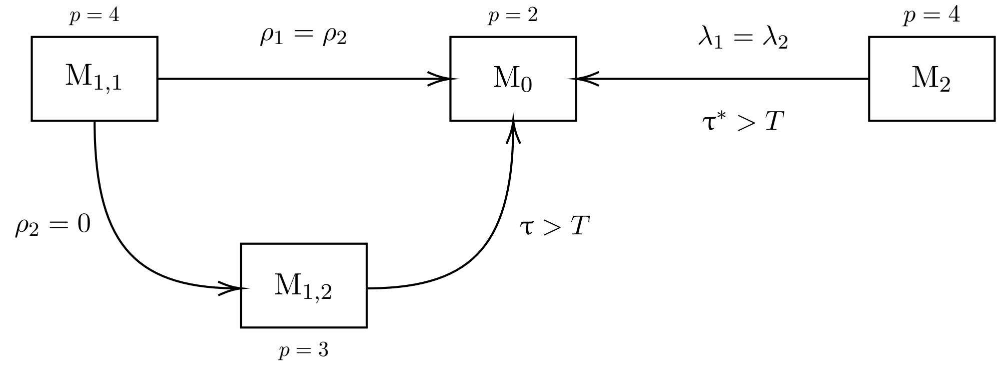

Figure 3.3: Schematic representation of the different nested RCMs and the possible parametric restrictions that allow one model to transform into another. $\mathrm{M}_{1,2}$ is equivalent to $\mathrm{M}_{1,1}$ when $\rho_2 \neq 0$. M_0 is equivalent to $\mathrm{M}_{1,1}$ when $\rho_1 \neq \rho_2$, equivalent to $\mathrm{M}_{1,2}$ when $\tau > T$, or equivalent to M_2 when $\lambda_1 \neq \lambda_2$ or $\tau^* > T$. Values of p indicate the number of parameters in each model.

3.3 Statistical inference

Throughout this thesis, analyses were performed within the frequentist framework. The method of maximum likelihood was applied to estimate the parameters of the described models. Two distinct approaches were used when estimating the parameters' confidence intervals. Comparisons between the adjusted models were performed by the Akaike's information Criterion (AIC), the Bayesian information criterion (BIC), the log-likelihood ratio test (for nested models), and the area under curve of the receiving operating characteristic (AUC-ROC). Finally, goodness-of-fit tests were applied to measure the models' adequacy.

3.3.1 Model estimation

Maximum likelihood estimation

Parameter estimation was done by maximising equation (3.4) for the sampling distribution. In this method, the maximum likelihood estimates (MLE) are the parameters values that maximise the value of the model's likelihood function based on the observed m_t. Equivalently, one can use the log-likelihood function, transforming all products into sums of the likelihood and thus facilitating the maximisation process [48]. MLE are calculated by solving the derivative of the log-transformed equation (3.4) when it is equal to zero. In theory, the MLE are the realised value of the estimators, being asymptotically unbiased and jointly normal [40]. The unknown parameters for the GLMs and RCM model M_0 can be estimated via maximum likelihood estimation, calculating the likelihood of each value that maximises the overall likelihood function.

Maximum likelihood estimates of the regression coefficients in the GLMs were calculated through a software command-defined function. The method uses the iteratively reweighted least squares (IRLS) for the maximum likelihood estimation, where through a process of weighted iteration, the best linear unbiased estimates $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are found. These estimated values are the ones which maximise the likelihood.

Profile likelihood method

For the particular case of the RCMs M_1 (both variations of the model) and M_2, with time-dependent parameters τ and τ^* defined in years, and restricted within the parametric space $\mathbb{N}^+$, the use of the simple maximum likelihood estimation is difficult to apply. Knowing that τ (although different, for the following description both cutoffs τ and τ^* will be broadly described using τ) can be a sequence of positive natural numbers, the profile likelihood method can be used, varying the natural cutoff value in order to estimate the remaining unknown parameters and its respective likelihood. The estimation is done by the following steps: (i) the cutoff parameter τ is initially fixed at 1; (ii) the remaining parameters are estimated via maximum likelihood; (iii) the corresponding log-likelihood function is calculated at these estimates and; (iv) τ is then increased by one unit of time, repeating steps (ii) and (iii) for every increment until it reaches a predefined maximum age value T. In the end, the overall maximum likelihood estimates are the ones associated with the value of τ that provide the maximum value of all the log-likelihood estimates calculated.

3.3.2 Confidence intervals

The estimation of the confidence interval for the parameters done in the GLMs was based on properties of the maximum likelihood estimators. It assumed estimated coefficients $\widehat{\boldsymbol{\beta}}$ to be asymptotically normal distributed. This approach makes use of the most broadly known form of confidence interval estimation, based on the standard error method. The standard error, $\text{se}(\widehat{\beta}) = \sqrt{V(\widehat{\beta})}$, is defined as the square root of the variance, a measure of variability of the estimate. Since the $\widehat{\beta}$ are asymptotic normal distributed, the $100(1-\alpha)\%$ confidence intervals can be estimated by calculating the lower and upper limits, given the formula

$$\left(\widehat{\beta} \pm \Phi_{\alpha/2} \times \text{se}(\widehat{\beta})\right) \ , \tag{3.14}$$

where $\pm\Phi_{\alpha/2}$ represents the lower and upper quantiles of the standard normal distribution, considering tails of size $\alpha/2$.

While working with the RCMs, in some cases the unknown parameters are expected to present estimated values close to zero. When in these situations, using the standard deviation method to calculate confidence intervals that close to the parametric space margin may not be the most efficient approach. For this thesis, the proposed alternative is to make use of the likelihood ratio statistic properties by testing null hypotheses for the acceptance or rejection of each one of the estimated parameters, within a defined critic region. Considering the case of the RCM M_0 and the estimation of the confidence interval of parameter λ. Via likelihood ratio, the null and alternative hypothesis in these situations are

$$H_0 : \lambda = \lambda_0 \ \textit{vs.} \ H_1 : \lambda \neq \lambda_0 \ ,$$

where λ is the model's parameter and λ_0 is a defined estimate. The likelihood ratio statistics is then based on

$$\text{D} = (-2) \times \left(\Lambda(\lambda_0, \rho^*) - \Lambda(\widehat{\lambda}, \widehat{\rho})\right) \overset{a(\text{H}_0)}{\rightsquigarrow} \chi^2_{(1)} \ , \tag{3.15}$$

where $\Lambda(\lambda_0, \rho^*)$ is the value of the log-likelihood function under the null hypothesis, with fixed λ_0 and estimated ρ^*, and $\Lambda(\widehat{\lambda}, \widehat{\rho})$ is the value of the log-likelihood function under the alternative hypothesis with both parameters $\widehat{\lambda}$ and $\widehat{\rho}$ equal to the MLE. This test statistic is asymptotically chi-squared distributed under the null hypothesis, with 1 degree of freedom that results from to the difference between the total number of unknown parameters of the models. The $100(1-\alpha)\%$ confidence interval for λ is then the range of all possible values of λ for which the null hypothesis is not rejected at a given critical region identified at the level of significance α. This method identifies the confidence intervals as the values for λ for which the estimated likelihood ratio test statistic is smaller or equal than the predefined critical value ($\lambda : D \leq \chi^2_{(1)}$).

3.3.3 Model comparison

Information criteria

The selection of the best fitted models was evaluated using two information criteria: the Akaike's information criterion,

$$\mathrm{AIC} = (-2) \times \Lambda_{\mathrm{model}} + 2p \ , \tag{3.16}$$

and the Bayesian information criterion,

$$\mathrm{BIC} = (-2) \times \Lambda_{\mathrm{model}} + p(\log n) \ , \tag{3.17}$$

where Λ_{model} is the log-likelihood function evaluated at the MLE for the model under consideration, p is the number of parameters, and n is the sample size. The first term of the criteria reflects the goodness of fit and the second term describes the model's complexity. The latter adds a penalty for the number of parameters p included. Under the principle of parsimony, for a set of candidate models the 'best' model is the one presenting the smallest values for AIC or BIC.

Likelihood ratio test

The Wilks' likelihood ratio test was used to compare the nested RCMs (see Figure 3.3). Under a defined null hypothesis for a parametric constrain, this test indicates the more parsimonious model based on the following test statistic

$$\mathrm{LRT} = (-2) \times (\Lambda_{\mathrm{H}_0} - \Lambda_{\mathrm{H}_1}) \overset{a(\mathrm{H}_0)}{\rightsquigarrow} \chi^2_{(\Delta p)} \ , \tag{3.18}$$

where Λ_{H_0} and Λ_{H_1} are the estimated maximum log-likelihood functions of the models that characterise the null and alternative hypothesis. Under the null hypothesis this test statistic is asymptotically chi-squared distributed, $\chi^2_{(\Delta p)}$, where Δp degrees of freedom is the difference between the total number of parameters from each one of the compared models (subtracting as $\Delta p = p_{H_1} - p_{H_0}$). For a significance level fixed at 0.05, p-values > 0.05 indicate the model defined by the null hypothesis is statistically better than the model represented by the alternative hypothesis.

Area under the receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is a standard technique used to infer about the performance of a model by measuring its predictive outcome accuracy [49]. This method explores the trade-off between sensitivity and specificity. These statistical measures are, respectively, the proportion by which a model correctly predicts a true positive or seropositive individual as a case, and the proportion by which it correctly detects a negative or seronegative individual as a non-case. Based on sensitivity and specificity, the accuracy of a model depends on how often, and with no wrong predictions, it differentiates between cases and non-cases.

The ROC curve of a model can be plotted for different cutoff points using the sensitivity values (proportion of identified true cases) in function of 1 − specificity (proportion of wrongly identified cases). The resulting area under the ROC curve (AUC), with range from 0 to 1, can be used as an index for a model's accuracy. AUC values equal to 0 indicate a poorly performing model, misidentifying every single individual of a population sample. Value of 0.5 predicts that a model is no better than a random guess. And value of 1 indicates a perfectly accurate model that is able of correctly predict the status of all individuals. When applied to several models under the same conditions, this measure for predictive accuracy can be used as an index statistics for model comparison.

3.4 Goodness-of-fit tests

To assess the goodness-of-fit of the models, several tests were performed inspecting the agreement between the observed and expected values. The Hosmer-Lemeshow goodness-of-fit statistic was used to assess the fit of the GLMs [49]. This test organises the outcomes in bin-like sub-groups $i = (1, \ldots, g)$, based on percentiles of the estimated probabilities called 'deciles of risk groups', and with formula given by

$$C_{HL}^2 = \sum_{i=1}^{g} \frac{(M_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i\,(1-\widehat{\pi}_i)} \overset{a(\mathrm{H}_0)}{\rightsquigarrow} \chi^2_{(g-1)} \ , \tag{3.19}$$

where M_i and n_i are the number of infected or seropositive individuals and the total number of individuals recorded within each decile of risk group i, respectively, $\widehat{\pi}_i$ is the prevalence or seroprevalence for individuals in the decile of risk group i. Under the null hypothesis that the model fits the data well, this test statistic is asymptotically chi-squared distributed, $\chi^2_{(g-1)}$, with $g-1$ degrees of freedom. With the intent to create balanced sample sizes across the deciles of risk groups, this test can create a somewhat arbitrarily subdivision of the observations instead of grouping observations by their respective values of variables, possibly lowering the test's power.

Assuming that more than a single goodness-of-fit test statistic would be applied, the test proposed by Noel Cressie and Timothy Read [50] was used, of formula

$$\mathrm{CR}^2 = \frac{2}{\delta(\delta+1)} \sum_{i=1}^{g} M_i \left[\left(\frac{M_i}{n_i\widehat{\pi}_i} \right)^{\delta} - 1 \right] \overset{a(\mathrm{H}_0)}{\rightsquigarrow} \chi^2_{(g-1)} \ , \tag{3.20}$$

where M_i and n_i are the number of infected or seropositive individuals and the total number of individuals within a class $i = (1, \ldots, g)$, $\widehat{\pi}_i$ is the prevalence or seroprevalence for individuals of that group, $\delta \in \mathbb{R}$ is a parameter that depending on its attributed value identifies the different goodness-of-fit tests used, and g is the total number of classes considered. By varying the values of δ on the equation, the tests here used are the Pearson's χ^2 ($\delta = 1$), the log-likelihood ratio statistic ($\delta = 0$), the Freeman-Tukey statistic ($\delta = -\frac{1}{2}$), the Neyman modified χ^2 statistic ($\delta = -1$), and the modified log-likelihood ratio statistic ($\delta = -2$). Under the null hypothesis for no significant difference between the observed and the estimated values, all test statistics are asymptotically chi-squared distributed, $\chi^2_{(g-1)}$, with $(g-1)$ degrees of freedom. For a significance level fixed at 0.05, p-values above this limit lead to the non rejection of the hypothesis of equality between the observed frequency distribution and the expected frequencies obtained by the model under testing.

3.5 Statistical software

All statistical analyses and inference tests were done in the software R, version 3.4.1. The command-defined function `glm()` was applied when calculating the GLMs' regression coeficients Their respective confidence intervals were obtained by use of the function `confint()`.

Profile likelihood method to estimate parameters from the RCMs used the `optim()` function (see the R script used to estimate parameters from RCM $\mathrm{M}_{1,1}$ in Appendices H). For each initialised value of τ, `optim()` finds the remaining estimates that maximise the likelihood function associated to the model's equation, through consecutive iterations. The development and analyses of the RCMs done in this thesis helped to develop and test the *SERO-AID* package. This R package (currently in its final stages of development) was created specifically with the intent to facilitate seroprevalence analyses, allowing the comparisons between different RCMs, such as the M_0 or M_2.

Chapter 4

Analysis of prevalence of infection using generalised linear models

Prevalence of infection can be studied through the use of the generalised linear models (GLMs). These models use the infection status of an individual as outcome of interest. The analysis of a GLM can help to explain how each covariate adds either an increase or decrease effect on the risk of infection. Throughout this chapter, the significance level for all statistical hypothesis testing remained fixed at 0.05.

4.1 Data preparation

When using GLMs, the probability of an individual from a population being identified as infected depends on the values of multiple explanatory variables, i.e., the risk factors or transmission determinants. Depending on the values when presented, the transmission determinants can influence the risk of infection, thus characterising the transmission heterogeneity recorded at different sites. To compare how malaria transmission behaves across the different surveyed sites, the best approach was to identify a baseline village from which all results could be compared to. The single village from West Usambara 3, Mgome, was the one selected. With this village as reference, its altitude – the lowest recorded – was subtracted to the values from the explanatory variable *Altitude*, allowing to interpret the null values of this quantitative variable. Starting at Mgome (0 meters), each unit value incremented in the transmission determinant *Altitude* represented a 100 meters high increase.

All explanatory variables here were categorical with the exception of *Altitude*, being fitted into the models as factors with different levels. To each factor was attributed a level corresponding to the baseline reference, from where the remaining levels were to be compared to. The reference level from *AgeGp* was the age group of individuals with ages within the interval of 1 to 4 years old, $Agegp_{1-4}$. Binary factor *Gender* had its first level characterising a female individual and the second level a male one. With Mgome as reference, the dominant ethnic group from the village, $EthGp_{Other}$, was the indicator level from variable *EthGp*, and West Usambara 3

(*Transect*$_{WU3}$) the baseline factor from variable *Transect.* All three binary variables representing the antigens *MSP1*, *MSP2*, and *AMA1*, had their reference levels corresponding to the absence of each respective antigen.

4.2 Model fitting and selection

When contemplating the potential transmission determinants in a model, their presence or absence may cause different impacts on the prevalence of malaria infection. To study how influential these variables might be, simple univariate GLMs were fitted [43]. From an analysis of deviance, variables that more effectively could change the outcome results on their own were identified (Table 4.1). Determinants *Transect* and *Altitude* appeared as the more informative variables. The logarithmic transformation of *Altitude* was firstly thought as a way to reduce variability of the data, although it turned out causing a reduced change in deviance, being less impactful than the original continuous variable. Further analysis under the logit link function showed a relationship between the response variable and *Altitude*, thus being preferred to the use of the transformed variable (Figure 4.1).

Table 4.1: Change in deviance for the variables considered for the prevalence of infection logistic GLMs (with respective degrees of freedom). P-values result from testing the significance of the model against the null model containing no explanatory variables. P-values > 0.05 indicate the constructed univariate model is not statistically different than the null model.

Risk factor	Change in deviance (d.f.)	p-value
Altitude	460.07 (1)	< 0.001
log(*Altitude*)	437.54 (1)	< 0.001
AgeGp	76.62 (2)	< 0.001
Gender	3.13 (1)	0.0767
EthGp	379.56 (3)	< 0.001
Transect	482.73 (5)	< 0.001
MSP1	156.23 (1)	< 0.001
MSP2	223.58 (1)	< 0.001
AMA1	324.75 (1)	< 0.001

Only the univariate model containing the factor *Gender*, with one degree of freedom, did not reject the null hypothesis for equality when compared to a null model without explanatory variables (p-value > 0.05). However, despite not being significant, and even though it caused the lesser impact on the deviance, *Gender* was still considered for the models' construction.

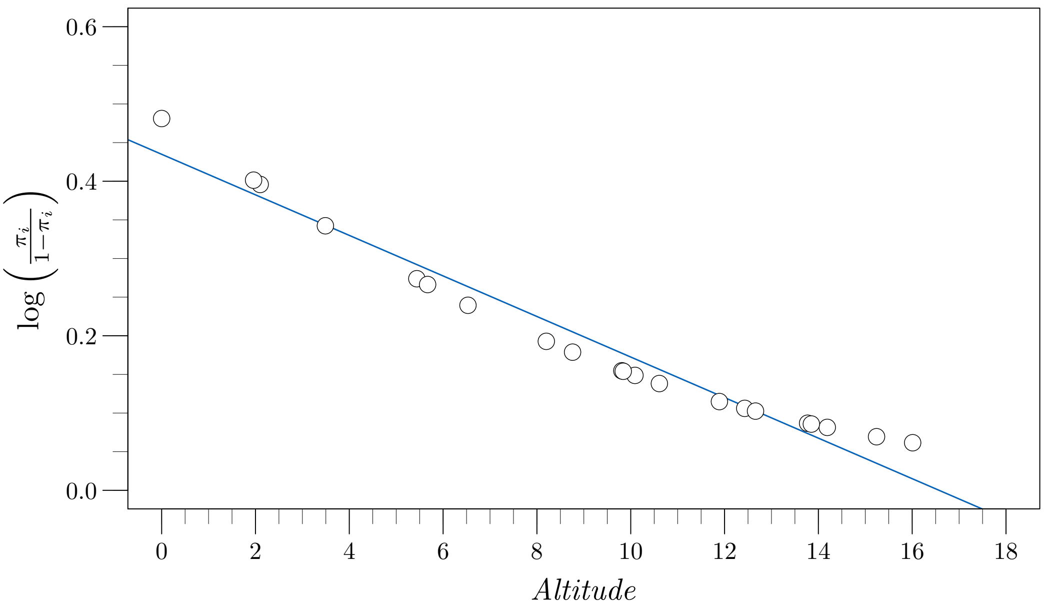

Figure 4.1: Relationship between prevalence of infection and the transmission determinant *Altitude*. Starting at 0 meters, each unit in the *Altitude* axis represents an increment in 100 meters high.

For inference purposes, the transmission determinants presented were separated into two different groups. The first group encompassed the 'demographical determinants', all variables characterising the environmental effects and individual characteristics that influenced the outcome. Risk factors *Transect*, *Altitude*, *AgeGp*, *Gender*, and *EthGp* belong in this group. A second group identified the 'exposure antigens', *MSP1*, *MSP2*, and *AMA1*. This approach allowed the creation of a first set of models and explain prevalence of infection based exclusively on demographical variables. Afterwards, the more informative systematic structure was selected and used as reference when adding the exposure factors. The addition of exposure antigens can be used to indicate the level of exposure to malaria parasites (Table 4.2).

Table 4.2: Structure of all GLMs' systematic components created, with respective degrees of freedom. Each model assumes probability of infection as response variable. Models `fit1` to `fit3` exclusively use variables from the demographical determinants. The following models add variables from the exposure antigens. The exposure transmission determinants are added in order, first individually (models `fit4` through `fit6`), and then pairing them up in two (`fit7`, `fit8`, and `fit9`). The last two models (`fit10` and `fit11`) add all three exposure factors to the `fit2` and `fit3` systematic structures.

Model	Terms fitted in model	d.f
`fit1`	*Altitude* + *AgeGp* + *Gender* + *EthGp* + *Transect*	5045
`fit2`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect*	5043
`fit3`	*Altitude* × *AgeGp* + *Gender* + *EthGp* × *Transect*	5039
`fit4`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP2*	5042
`fit5`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP1*	5042
`fit6`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *AMA1*	5042
`fit7`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP1* + *MSP2*	5041
`fit8`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP2* + *AMA1*	5041
`fit9`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP1* + *AMA1*	5041
`fit10`	*Altitude* × *AgeGp* + *Gender* + *EthGp* + *Transect* + *MSP1* + *MSP2* + *AMA1*	5040
`fit11`	*Altitude* × *AgeGp* + *Gender* + *EthGp* × *Transect* + *MSP1* + *MSP2* + *AMA1*	5036

The first structure built, `fit1`, described malaria prevalence as a function of all demographical determinants with no interactions considered. When recreating malaria prevalence over different time points, the age (or in this case the age group) of an individual is an important risk factor to consider, as it can be a categorical proxy for time of exposure. With *Altitude* being a proxy for transmission intensity, the empirical association between this variable and *AgeGp* (`fit2` and `fit3`) was expected to more accurately represent the altitude-dependent prevalence values. Under this interaction, the prevalence of infection should present higher values at age group $1-4$, decreasing to lower estimates until the higher age group $15-45$. However, the maximum prevalence reached depends on the continuous *Altitude*, as lower altitude sites have higher transmission intensities recorded. With each transect mostly represented by a single ethnic group, the structure from `fit3` recreated the association between variables *EthGp* and *Transect.*

Afterwards, the addition of exposure variables onto the models was made gradually. Starting at `fit4`, the exposure antigens were added to the previous model structure from `fit2`. First, a single exposure transmission determinant was added to the model. Factor *MSP2* was added in `fit4`, *MSP1* in `fit5`, and *AMA1* was added in `fit6`. The comparison of the three models should clarify for a better understanding of how sensitive the antigens are to the presence of infection (their immunogenicity). Using `fit2`'s demographical structure, models `fit7`, `fit8`, and `fit9`, added two exposure factors. Systematic component from `fit7` included both *MSP1* and *MSP2*, fit8 incorporated *MSP2* and *AMA1*, and `fit9` built *MSP1* and *AMA1.* These models functioned as an intermediary complement between the addition of a single antigen factor, and the inclusion of all three. The order by which the exposure variables were included was based on previous knowledge that specific antigens MSP1 and AMA1 are more impactful than MSP2. Finally, `fit10` included all exposure risk factors into `fit2`'s systematic structure. The systematic component `fit11` then extended all three exposure antigens onto the demographical systematic structure from `fit3`, with a relation between *EthGp* and *Transect.*

All systematic components were then fitted using three link functions, logistic, probit, and complementary log-log, for a total of 33 possible GLMs (Table 4.3). Information criteria (AIC and BIC) and area under the ROC curve (AUC) were the measures used to compare the produced results and select the overall best GLM. AUC reflected mostly the information gained by the addition of different variables. This criterion remained unchanged for models possessing the same systematic component structures, simply reflecting the discriminating power of each model. To compare models built with an equal structure, that only varied their respective link functions, values of AIC and BIC were used. Since the penalty parameter from the criteria remained unchanged for models with same systematic components (parameters seen in equations (3.16) and (3.17)), the lowest produced value depended only on the likelihood value given by each associated link function. The GLMs were tested for their goodness-of-fit using the Hosmer-Lemeshow test statistic with ten deciles of risk groups, and all test statistics originated from equation (3.20), considering a similar number of classes.

Under the systematic component from `fit1`, the best fitted GLM linked the outcome to its predictors via the cloglog link function. Assessing the three GLMs from `fit1`, this model presented the lowest AIC and BIC values recorded. Complementary, it also had the highest AUC value. Models adjusted with both `fit2` and `fit3` presented best estimated results when the probit link function was used. The use of an interaction between *Altitude* and *AgeGp* in these models showed an increase in information gain, reaching AUC values above 0.78. Comparing both probit models, `fit2` appeared the better option, thus discarding the $EthGp-Transect$ interaction.

All models where exposure antigens were added presented better comparative results when the logistic link function was used. The AIC and BIC values from the ordered models increased gradually, as more transmission determinants were added. Model created using the systematic components `fit10` and `fit11`, reached estimated values of AUC above 0.80. Comparing the logistic models `fit4`, `fit5`, and `fit6` suggested *AMA1* (`fit6`) to be the more informative factor, within the analysed antigens. Indeed, `fit6` described better the outcome using the same demographical structure as the other two antigens, also corroborating the values obtained for the AMA1 antigen in the analysis of deviance (Table 4.1). Factor *MSP1* (`fit5`) was the second more informative, with *MSP2* (`fit4`) being the one granting the least information to the models. Logistic models `fit7`, `fit8`, and `fit9` consolidated this information, as both GLMs with the lesser informative transmission determinant *MSP2* showed reduced information. The models' analysis made from the majority of the goodness-of-fit results suggested the logistic `fit8` and `fit9` results did not depart from the model, not rejecting the null hypothesis.

Similarly to the comparison between the two probit models `fit2` and `fit3`, the logistic models `fit10` and `fit11` were compared one against the other. Having similar values for AUC, the overall lower AIC and BIC results from the single interaction model identified it as the more parsimonious GLM built. The logistic GLM `fit10` presented good performances from the comparison methods, although its p-values from the Hosmer-Lemeshow goodness-of-fit test statistic showed some evidence for lack of fit (p-value= 0.016). The test indicated the model to be poorly adjusted to the data, contrarily to the rest. Regardless of this result, and with the main objective being to build the best possible descriptive model, the logistic `fit10` was the selected model.

Table 4.3: Results for the constructed GLMs, grouped by the trio of link functions applied on each structure. Each model presents values of information criteria (AIC and BIC), and AUC, as well as results for the goodness-of-fit test statistics: Hosmer-Lemeshow with 10 deciles of risk group (C^2_{HL}), Pearson χ^2 (X^2), log-likelihood ratio statistic (G^2), Freeman-Tukey statistic (T^2), Neyman modified χ^2 (NM^2), and modified log-likelihood ratio statistic (GM^2). A bold font indicates the best AIC and BIC for fit.

Model	Link	Comparison methods			Goodness-of-fit					
		AIC	BIC	AUC	C^2_{HL}	X^2	G^2	T^2	NM^2	GM^2
`fit1`	logit	4221.24	4306.12	0.778	0.057	0.374	0.374	0.372	0.362	0.370
	probit	4224.02	4308.89	0.778	0.054	0.382	0.650	0.378	0.365	0.186
	cloglog	4218.74	4303.62	0.778	0.097	0.443	0.438	0.433	0.417	0.428
`fit2`	logit	4190.80	4288.73	0.782	0.059	0.258	0.251	0.245	0.221	0.238
	probit	4188.80	4286.73	0.782	0.104	0.327	0.281	0.285	0.228	0.287
	cloglog	4195.65	4293.58	0.782	0.026	0.180	0.051	0.161	0.130	0.412
`fit3`	logit	4194.48	4318.52	0.783	0.028	0.132	0.138	0.138	0.127	0.136
	probit	4193.65	4317.70	0.783	0.002	0.018	0.021	0.023	0.019	0.024
	cloglog	4195.97	4320.02	0.783	0.094	0.273	0.215	0.310	0.315	0.425
`fit4`	**logit**	4145.53	4249.99	0.788	0.585	0.877	0.885	0.888	0.898	0.892
	probit	4146.00	4250.46	0.788	0.854	0.970	0.983	0.969	0.967	0.949
	cloglog	4150.52	4254.98	0.788	0.334	0.697	0.312	0.710	0.717	0.987
`fit5`	**logit**	4134.33	4238.79	0.793	0.034	0.164	0.168	0.168	0.160	0.167
	probit	4135.47	4239.93	0.793	0.008	0.053	0.033	0.036	0.019	0.039
	cloglog	4143.88	4248.34	0.793	0.054	0.234	0.035	0.237	0.228	0.808
`fit6`	**logit**	4112.50	4216.96	0.794	0.205	0.472	0.470	0.465	0.433	0.457
	probit	4114.18	4218.64	0.794	0.536	0.814	0.834	0.811	0.803	0.786
	cloglog	4118.25	4222.71	0.794	0.180	0.491	0.152	0.448	0.388	0.864
`fit7`	**logit**	4107.09	4218.08	0.795	0.013	0.066	0.044	0.032	0.007	0.022
	probit	4109.29	4220.28	0.795	0.011	0.050	0.047	0.049	0.030	0.049
	cloglog	4116.52	4227.51	0.795	0.014	0.087	0.006	0.043	0.011	0.211
`fit8`	**logit**	4087.89	4198.88	0.797	0.267	0.543	0.493	0.464	0.358	0.431
	probit	4092.61	4203.60	0.797	0.174	0.388	0.536	0.429	0.441	0.331
	cloglog	4092.59	4203.58	0.797	0.275	0.579	0.208	0.477	0.354	0.822
`fit9`	**logit**	4078.17	4189.16	0.800	0.019	0.084	0.102	0.108	0.112	0.111
	probit	4081.87	4192.86	0.800	0.004	0.021	0.030	0.034	0.036	0.037
	cloglog	4086.87	4197.85	0.800	0.037	0.157	0.024	0.154	0.137	0.605
`fit10`	**logit**	4062.81	4180.33	0.801	0.016	0.089	0.100	0.103	0.100	0.104
	probit	4068.41	4185.92	0.801	0.001	0.010	0.018	0.017	0.016	0.014
	cloglog	4070.81	4188.33	0.801	0.012	0.085	0.009	0.052	0.023	0.220
`fit11`	**logit**	4067.07	4210.70	0.802	0.097	0.276	0.260	0.247	0.186	0.230
	probit	4073.75	4217.38	0.802	0.001	0.008	0.023	0.019	0.022	0.015
	cloglog	4071.02	4214.65	0.802	0.072	0.265	0.129	0.259	0.235	0.464

4.3 Model inferences

With the majority of the goodness-of-fit test statistics suggesting no evidence against the null hypothesis of good fit, no further adjustments were made to the logistic model `fit10` (Table 4.4). Assessing the AUC of approximately 0.801, its optimal cutoff for sensitivity and specificity were calculated as 0.786 and 0.691, respectively (Figure 4.2).

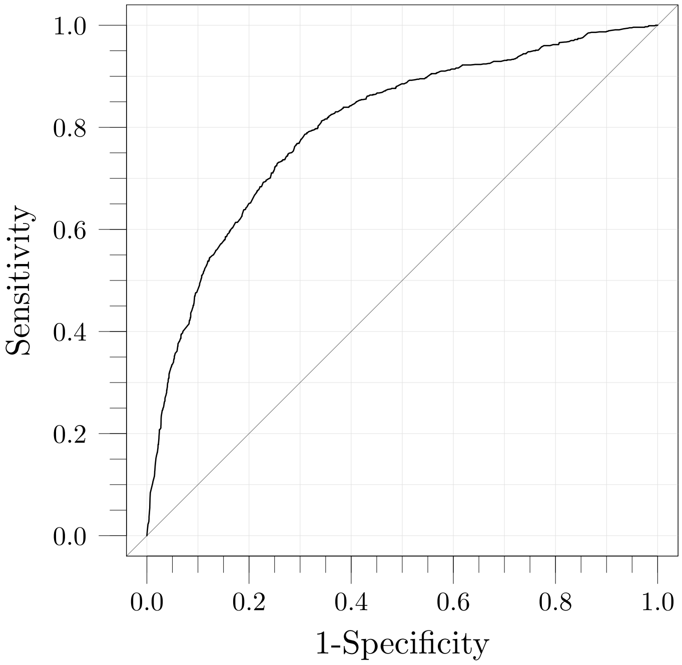

Figure 4.2: ROC curve of the logit model `fit10` (AUC = 0.801). Maximum discrimination for values of sensitivity equal to 0.786 and specificity equal to 0.691.

The logit link function facilitated further epidemiological inferences of the coefficients of the model, allowing for the interpretation of the explanatory variables in terms of odds ratio [43]. This is, comparing the odds of possible malaria infection when in relation to a similar individual that has a single variation in one of its variables. The analysis of the more significant transmission determinants and their respective odds ratios can create an association between the variations in malaria transmission intensity across the different situations that characterise the data.

Table 4.4: Adjusted logistic GLM `fit10` with respective significant variables estimates, standard error and p-value. Odds ratio are given in relation to the indicator level of each transmission determinant (the odds ratios for these levels are equal to 1.00). For reference, the baseline level from the categorical demographical determinants are here indicated: $AgeGp_{1-4}$, $Gender_{Female}$, $EthGp_{Other}$, and $Transect_{WU3}$.

Coefficient	Factor level	Parameter estimate, $\widehat{\beta}$	Standard error	p-value	Odds ratio, $e^{\widehat{\beta}}$	95% Confidence interval
(Intercept)	—	−0.863	0.218	<0.001	0.42	(0.28, 0.65)
Altitude	—	−0.148	0.019	<0.001	0.86	(0.83, 0.89)
AgeGp	5−14	0.112	0.190	0.555	1.12	(0.77, 1.63)
	15−45	−1.681	0.197	<0.001	0.19	(0.13, 0.27)
Gender	Male	0.134	0.083	0.105	1.14	(0.97, 1.34)
EthGp	Wachaga	0.619	0.320	0.053	1.86	(1.00, 3.50)
	Wapare	−0.228	0.209	0.276	0.80	(0.53, 1.20)
	Wasambaa	0.194	0.160	0.226	1.21	(0.89, 1.67)
Transect	Rombo	−0.956	0.322	0.003	0.38	(0.20, 0.72)
	N. Pare	−1.470	0.337	<0.001	0.23	(0.12, 0.44)
	S. Pare	−0.407	0.247	0.099	0.67	(0.41, 1.08)
	W. Usamb. 1	0.472	0.214	0.027	1.60	(1.05, 2.44)
	W. Usamb. 2	0.028	0.212	0.894	1.03	(0.68, 1.56)
AMA1	1	0.666	0.099	<0.001	1.95	(1.61, 2.36)
MSP1	1	0.505	0.097	<0.001	1.66	(1.37, 2.00)
MSP2	1	0.396	0.095	<0.001	1.49	(1.23, 1.79)
Altitude × *AgeGp*	5−14	0.022	0.022	0.335	1.02	(0.98, 1.07)
	15−45	0.114	0.023	<0.001	1.12	(1.07, 1.17)

In Mgome, 111 out of the 225 inhabitants were not infected. Being the village with the highest prevalence of infection at the moment of sampling, one expected this selected baseline site to present good characteristics that, by comparison, could help to describe the different prevalence amongst all other villages. The odds ratio analysis suggested that the odds of an individual developing malaria infection at the reference village's altitude (0 meters, recall Section 4.1) was significantly higher than those who inhabited any site with a higher altitude ($\widehat{\beta} = -0.148$ and p-value < 0.001). Being a proxy of malaria transmission, each 100 meters increased in altitude granted a reduction in odds of infection of approximately 14% ($\widehat{OR} = 0.86$). This reduction showed how impactful altitude is, working as a protective factor. The small confidence interval for the true odds ratio suggested a high precision of the odds of infection.

Comparing the different age groups, only individuals older than 14 years old appeared to have their odds of infection significantly reduced ($\widehat{\beta} = -1.681$ and p-value < 0.001). This reduction – $\widehat{OR} = 0.19$, the biggest estimated in the categorical transmission determinants – suggested older individuals were more prepared to live in sites of endemic transmission settings. These older inhabitants, when in similar conditions as children from the reference level $AgeGp_{1-4}$, would had their odds of being infected reduced by 73% to 87%. There was no significant difference from infants to children with ages between 5 and 14 years old. The binary risk factor *Gender*, was kept throughout the model's construction. Ultimately it did not seem to play a significant role influencing prevalence of infection ($\widehat{\beta} = 0.134$ and p-value = 0.105).

Assessing the different ethnicities, only the Wachaga ethnic group ($\widehat{\beta} = 0.619$ and p-value = 0.053) seemed to produce a borderline significant result on the response variable, with $\widehat{OR} = 1.86$. Despite the inclusion of the unity in the estimated confidence interval for the true odds ratio, this level suggested a tendency. In similar conditions, individuals from this ethnic group would had higher odds infection than someone from the ethnicities found in Mgome, $EthGp_{Other}$, at the moment of sampling. The odds of infection in individuals from the remaining ethnic groups Wapare ($\widehat{\beta} = -0.228$ and p-value = 0.276) and Wasambaa ($\widehat{\beta} = 0.194$ and p-value = 0.226) did not appear significantly different in the analysis.

The covariate *Transect* showed how malaria transmission intensity varied between the two Tanzanian regions, across the different defined transects. If compared in the same conditions, individuals from the four villages in Rombo ($\widehat{\beta} = -0.956$ and p-value = 0.003) had their odds of being infected reduced by 62% ($\widehat{OR} = 0.38$), while villages in the North Pare ($\widehat{\beta} = -1.470$ and p-value < 0.001) had their odds reduced by 77% ($\widehat{OR} = 0.23$). The latter showed the more significant and precise odds reduction. Similar to the tendency value in $EthGp_{Wachaga}$, with the inclusion of the unity within the confidence interval of the true odds ratio, villages belonging to South Pare seem to present reduced odds for malaria infection. In the Tanga region, West Usambara 1 ($\widehat{\beta} = 0.472$ and p-value = 0.027) suggested a higher odds of infection relatively to Mgome, having prevalence growing by 1.60 times. West Usambara 2 ($\widehat{\beta} = 0.028$ and p-value = 0.894) did not seem to significantly impact prevalence.

The presence of each one of exposure antigens, *AMA1*, *MSP1*, or *MSP2* ($\widehat{\beta} = 0.666$, $\widehat{\beta} = 0.505$, and, $\widehat{\beta} = 0.022$, respectively, all with p-values < 0.001), when compared to their respective control level, consistently increased the prevalence as a consequence for exposure ($\widehat{OR}_{AMA1} = 1.95$, $\widehat{OR}_{MSP1} = 1.66$, $\widehat{OR}_{MSP2} = 1.49$, respectively). This increment suggested how impactful the antigens can be. Since in normal conditions any non exposed individual is expected to not develop specific antimalarial antibodies, thus not being at risk of becoming infected, the simple detection of an antigen by itself implies a increase on the odds of infection due to exposure. Depending on the three antigens, odds of infection increased from almost 50%, when MSP2 was detected, to almost doubling the odds whenever antigen AMA1 was detected in the serum.

4.4 Summary

The identification and study of the primary malaria infection covariates in a region is an approach that can be made as soon as enough variables are collected. These analyses present some flexibility, as different and more specific transmission determinants can be included for consideration, depending on the study objectives [51]. When overseeing the regions of the Northeast Tanzania, the best combination of transmission determinants through the GLMs helped making inferences about the relative proportion of infected malaria cases across the different sites at the time of sampling. Also, through the regression coefficient estimates and estimated odds ratio, the relative transmission intensity across the different villages was compared, inferring about impactful determinants that could originate the prevalence estimated and transmission intensity heterogeneity measured.

The application of different categorical variables onto the logistic regression model `fit10` identified the more susceptible levels to infection. In the demographical determinants, both *AgeGp* and *Transect* proved to be important categorical risk factors to describe occurrence of infection.

The odds of adults becoming infected with malaria is lower since they might have developed specific immunity throughout periods of recurrent exposure and reinfections. The impact of these transmission determinants can be related to the geographical conditions of each site, modelling the environment and overall anthropomorphic characteristic of its inhabitants. Variable *Altitude* also played an important role. As a proxy for transmission intensity, a value increment in this continuous risk factor can be interpreted as a decrease in temperature and humidity, gradually reducing the number of *Anopheles* mosquitoes, *P. falciparum* transmitters.

Binary variables for the antigen responses presented elevated odds for malaria infection. All three exposure antigens worked well as stable predictors for malaria exposure, being indirectly influenced by the demographical determinants delineating the level of exposure, and presenting a direct effect in the immunological uplift (aquired immunity) over time gained by an individual [52]. With the antigens identified as good indicators for exposure to malaria parasites and overall transmission intensity, when measured at different ages, Chapter 5 focused the detection of the immunologic responses as outcome of interest. Specific stochastic models were applied, measuring transmission intensity produced by different characteristics from each village, across the sequence of different ages.

Chapter 5

Reverse catalytic models: analysis of seroprevalence

All analyses performed in this chapter used the seropositivity registered from the three specific *P. falciparum* antimalarial antigens MSP1, MSP2, and AMA1, as the outcome of interest. Focusing on these detected signals of exposure to malaria parasites brought new inference tools to the study of transmission intensity across the distinct sites. The nested reverse catalytic models (RCMs) were here applied to the data of each village. The models' results were compared using the likelihood ratio tests, identifying the statistically overall best model for each village and antigen (parametric dependencies represented in Figure 3.3).

Models considering a change in seroreversion rate (SRR), $M_{1,1}$ and the more restricted $M_{1,2}$, were proposed alongside this thesis. Their results were compared beforehand, selecting the true model for the age-dependent change in SRR. Afterwards, both models assuming changes in their transition rates were compared to M_0 with constant seroconversion rate (SCR) and SRR across all ages. The application of different RCMs allowed for inferences on how the acquired immune system behaves upon exposure to a certain level of transmission intensity. Also, by tracking different antimalarial antibodies detected at different ages, the RCMs can explain the past serological history of the populations, describing if any noticeable change occurred in recent decades.

5.1 Model results and comparison

5.1.1 Models for age-dependent SRR: $M_{1,1}$ *vs.* $M_{1,2}$

Before comparing the different RCMs alongside their sero-epidemiological implications, both models assuming change in SRR were compared (Tables 2, 3 and 4 from Appendices, for the antigens MSP1, MSP2, and AMA1, respectively). Both models represented the individual biological effects of accumulated antimalarial immunity due to exposure, under stable and constant rate of transmission across all ages. The models assumed a reduction in SRR, ρ_1 to ρ_2, given an age cutoff τ. However, while $M_{1,1}$ considered the reduction $\rho_1 \geq \rho_2$, the more restricted $M_{1,2}$ assumed that after some age τ, all individuals had transited into the seropositive state ($\rho_2 = 0$), remaining so.

When the models were fit to data from different antigens, $\widehat{\rho}_2$ estimates from model $M_{1,1}$ were consistent with zero for some villages. The likelihood ratio tests further validated the results, suggesting the model to be statistically equivalent to $M_{1,2}$ (p-value > 0.05) for a majority of the villages. When comparing both age-dependent SRR models, despite some villages being significantly better described by $M_{1,1}$, possible numeric errors registered in the estimates of this model (resulting in numerous estimates for $\widehat{\rho}_1 > 10$) led to the choosing of $M_{1,2}$ as the best model to take further into the analyses.

With $M_{1,2}$ as the selected model, estimates for $\widehat{\rho}_2$ equal to zero in every village suggested individuals persist as seropositive after some time, regardless of antigen or transmission intensity. Within each transect, SCR estimates, $\widehat{\lambda}$, seemed to decrease with altitude. Contrarily, the influence of altitude was positively related to $\widehat{\rho}_1$, as lower estimates were generally found in villages located at low and intermediate altitudes. Estimates for the cutoff parameter $\widehat{\tau}$ were consistently higher for lower altitude villages, reaching lower estimates as the villages' elevation increased. This relation implied that at lower altitude regions, with consequent higher transmission levels of endemic malaria infection, members of an exposed population tend to become seropositive and revert at lower rates throughout their life. As opposed to individuals inhabiting villages at higher altitudes, with lower transmission levels of malaria infection. On those sites, higher estimates for SRR explain the faster recovery rate due to non recurrent exposure to *P. falciparum* parasites. Also, higher altitude exposed populations can become and remain seropositive earlier in live under endemic transmissison intensity.

5.1.2 Testing change in SRR: M_0 *vs.* $M_{1,2}$

Having model $M_{1,2}$ identified, the analysis proceeded by testing whether this proposed age-dependent change in SRR was indeed significantly better to estimate malaria transmission intensity. Model $M_{1,2}$ was compared against the nested M_0 (Table 5.1 for MSP1, and from Appendices D, Tables 5 and 6 for MSP2 and AMA1, respectively). The likelihood ratio tests concluded that for most villages, M_0 was the preferred model, irrespectively of the antigen under analysis. Some villages at intermediate and high altitudes, however, presented a significant age-dependent change in its SRR. The likelihood ratio tests from the MSP1 data set identified villages Mpinji ($\widehat{\tau} = 8$) from the South Pare transect, and Kwadoe ($\widehat{\tau} = 14$) from the West Usambara 2 transect. Assessing the MSP2 antigen, Mpinji ($\widehat{\tau} = 34$) was once again identified for having a significant reduction in its SRR. This time taking place in individuals with ages between 31 and 36 years old. Village Funta ($\widehat{\tau} = 40$), from the West Usambara 2, was also identified in this data set. Finally, in the AMA1 antigen, individuals at Machame Aleni ($\widehat{\tau} = 14$), from the transect Rombo a were estimated to have had a SRR reduction to zero approximately at the age of 14, remaining seropositives for this antigen for the rest of their lifes.

Inferring about the practical epidemiological implications of choosing the simpler model over $M_{1,2}$, correlation analyses between the estimated SCR and SRR were performed (Figure 5.1). Attending the SCR correlations (plots A1, A2, and A3 from Figure 5.1), estimates from both models were strongly correlated with each other ($r^2_{\text{MSP1}} = 0.99$, $r^2_{\text{MSP2}} = 0.82$, and $r^2_{\text{AMA1}} = 0.75$). Despite this linear relation it is worth noting that by choosing model M_0 rather than $M_{1,2}$, all SCR estimates were subjected to a reduction up to approximately 11%. In the SRR analysis (plots B1, B2, and B3 from Figure 5.1), parameter $\widehat{\rho}$ from M_0 did not present a strong positive correlation to $\widehat{\rho}_1$ from $M_{1,2}$, estimated for ages lower than $\widehat{\tau}$ ($r^2_{\text{MSP1}} = 0.03$, $r^2_{\text{MSP2}} = 0.27$, and $r^2_{\text{AMA1}} = 0.57$). Model M_0 largely underestimated SRR for ages lower than the age cutoff in antigens MSP1 and MSP2, and overestimated the same rate when applied to the AMA1 antigen. After the age cutoff, the underestimation seen in MSP1 and MSP2 shifts, as model $M_{1,2}$ assumes SRR equal to zero for ages greater than $\widehat{\tau}$. This lack of correlation between the SRR estimates from both models might depend on the different levels of altitude and transmission intensity values registered at each village, influencing the parameter estimate and corresponding value of $\widehat{\tau}$. The comparison of the estimated seroprevalence curves produced by the two models showed how assuming a reduction in SRR at some age can produce an effect on the maximum reachable seroprevalence (Figure 5.2 for the estimated seroprevalence from MSP1, and Figures 1 and 3 from Appendices F and G, respectively, for estimated seroprevalence from MSP2 and AMA1).

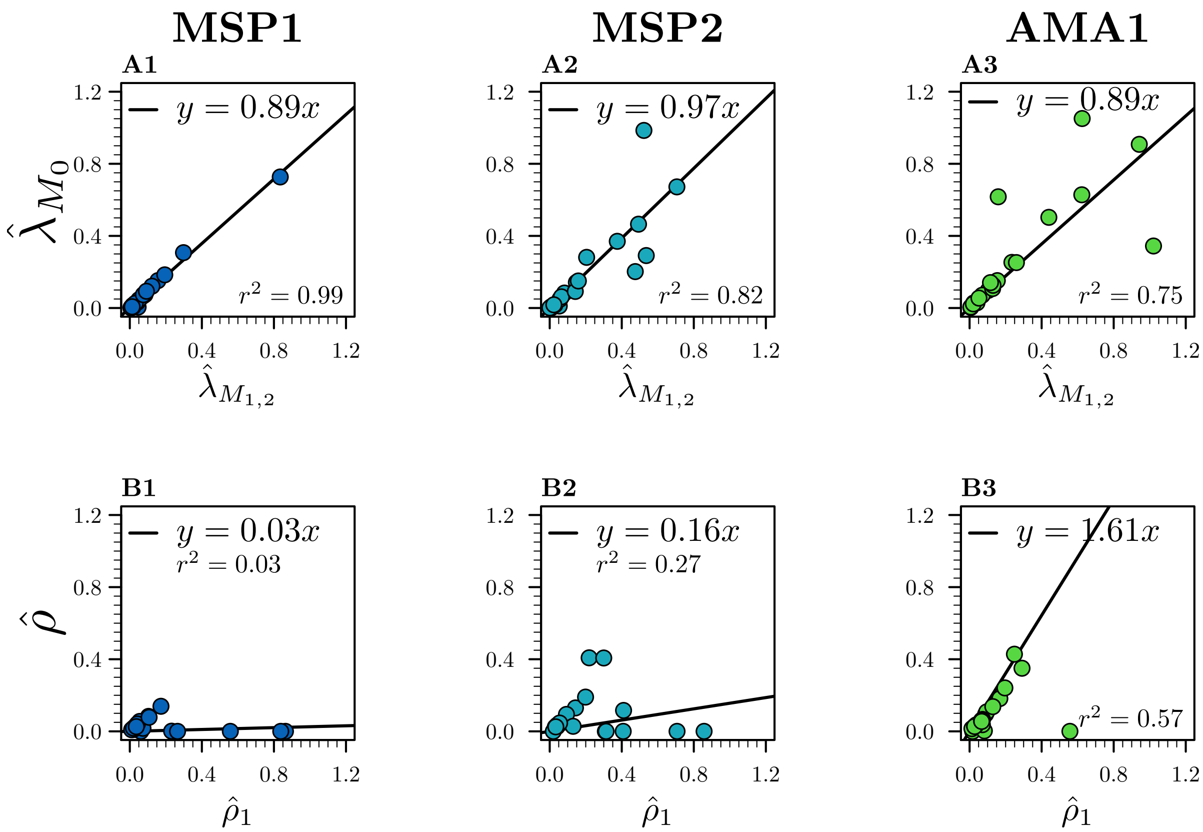

Figure 5.1: Comparison of parameter estimates from models M_0 and $M_{1,2}$. The first column, with figures **A1** and **B1**, represents analyses made for the MSP1 antigen data set, the second column, with figures **A2** and **B2**, represents analyses using the MSP2 antigen data, and third column, with figures **A3** and **B3**, represents analyses for the AMA1 antigen data set. Figures **A1**, **A2**, and **A3** in the first row, shows the 21 model M_0 SCR estimates, $\widehat{\lambda}_{M_0}$, as function of the corresponding $M_{1,2}$ estimates. Figures **B1**, **B2**, and **B3** in the second row represent model M_0 SRR estimates, $\widehat{\rho}$, as function of the $M_{1,2}$'s $\widehat{\rho}_1$, calculated before the cutoff value $\widehat{\tau}$. For the second row, villages with $\widehat{\rho}_1 > 10$ were excluded form the analysis. For each graph the line of tendency (in black) with respective formula and corresponding Pearson's correlation statistics, r^2, is shown.

Table 5.1: Comparison between models M_0 and $M_{1,2}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-MSP1 antigen in samples from the 21 villages. Model M_0 assumes a constant SCR and SRR (λ and ρ, respectively), while model $M_{1,2}$ assumes a constant SCR, λ_1 for ages $< \tau$ and $\lambda_2 = 0$ otherwise. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model M_0 with $M_{1,2}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits. This event can be mostly seen at high altitude villages.

Transect	Village	Model M_0			Model $M_{1,2}$				p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}$ (95% CI)	$\hat{\rho}_1$ (95% CI)	$\hat{\tau}$	logL	
Rombo	Mokala	0.014 (0.008, 0.030)	0.016 (0.000, 0.093)	-45.68	0.016 (0.009, >5)	0.031 (0.000, >10)	30	-45.63	0.752
	Machame Aleni	0.047 (0.024, 0.140)	0.083 (0.020, 0.372)	-56.54	0.052 (0.027, >5)	0.104 (0.031, 0.460)	37	-56.01	0.303
	Ikuini	0.014 (0.010, 0.038)	0.000 (0.000, 0.102)	-58.55	0.020 (0.011, >5)	0.061 (0.000, >10)	22	-58.16	0.377
	Kileo	0.307 (0.205, 0.499)	0.055 (0.023, 0.125)	-44.77	0.298 (0.210, >5)	0.055 (0.026, 0.110)	40	-45.61	~1.000
N. Pare	Kilomeni	0.005 (0.002, 0.038)	0.000 (0.000, 0.351)	-18.93	0.046 (0.004, >5)	0.864 (0.000, >10)	30	-17.22	0.064
	Lambo	0.015 (0.010, 0.041)	0.001 (0.000, 0.104)	-37.86	0.019 (0.010, >5)	0.232 (0.000, >10)	7	-37.60	0.471
	Ngulu	0.074 (0.046, 0.159)	0.000 (0.000, 0.058)	-13.79	0.084 (0.052, >5)	>10 (0.000, >10)	1	-13.31	0.327
	Kambi ya Simba	0.067 (0.039, 0.122)	0.010 (0.000, 0.059)	-27.45	0.074 (0.043, >5)	0.019 (0.000, >10)	36	-27.01	0.348
S. Pare	Bwambo	0.005 (0.003, 0.017)	0.000 (0.000, 0.126)	-36.14	0.017 (0.004, >5)	0.265 (0.000, >10)	27	-34.93	0.120
	Mpinji	0.005 (0.002, 0.009)	0.000 (0.000, 0.055)	-21.75	0.009 (0.005, >5)	>10 (0.134, >15)	8	-19.23	0.025
	Goha	0.032 (0.021, 0.054)	0.017 (0.000, 0.070)	-58.55	0.042 (0.025, >5)	0.073 (0.000, 0.196)	24	-57.35	0.121
	Kadando	0.152 (0.108, 0.230)	0.029 (0.011, 0.068)	-57.05	0.156 (0.115, >5)	0.032 (0.014, 0.069)	38	-56.82	0.498
W. Usamb. 1	Emmao	0.001 (0.000, 0.127)	0.000 (0.000, >10)	-10.86	0.006 (0.001, >5)	0.838 (0.000, >10)	26	-9.68	0.124
	Handei	0.044 (0.021, 0.128)	0.139 (0.040, 0.520)	-57.56	0.049 (0.025, >5)	0.173 (0.062, 0.527)	37	-56.42	0.131
	Tewe	0.026 (0.020, 0.040)	0.000 (0.000, 0.032)	-64.19	0.030 (0.023, >5)	>10 (0.000, >10)	2	-62.63	0.077
	Mn'galo	0.074 (0.056, 0.099)	0.009 (0.000, 0.026)	-67.23	0.075 (0.058, >5)	0.010 (0.000, >10)	40	-67.08	0.584
W. Usamb. 2	Kwadoe	0.007 (0.004, 0.011)	0.000 (0.000, 0.032)	-37.52	0.013 (0.007, >5)	0.558 (0.013, 2.794)	14	-35.47	0.043
	Funta	0.121 (0.088, 0.172)	0.018 (0.005, 0.045)	-47.92	0.123 (0.093, 0.169)	0.020 (0.007, 0.045)	40	-47.46	0.337
	Tamota	0.093 (0.063, 0.152)	0.041 (0.014, 0.100)	-64.58	0.092 (0.066, >5)	0.042 (0.017, 0.091)	40	-65.06	~1.000
	Mgila	0.184 (0.131, 0.286)	0.026 (0.008, 0.070)	-64.36	0.194 (0.140, >5)	0.036 (0.013, 0.091)	33	-64.45	~1.000
W. Usamb. 3	Mgome	0.727 (0.398, 4.520)	0.078 (0.023, 0.482)	-34.25	0.835 (0.457, >5)	0.107 (0.035, >10)	37	-32.48	0.060

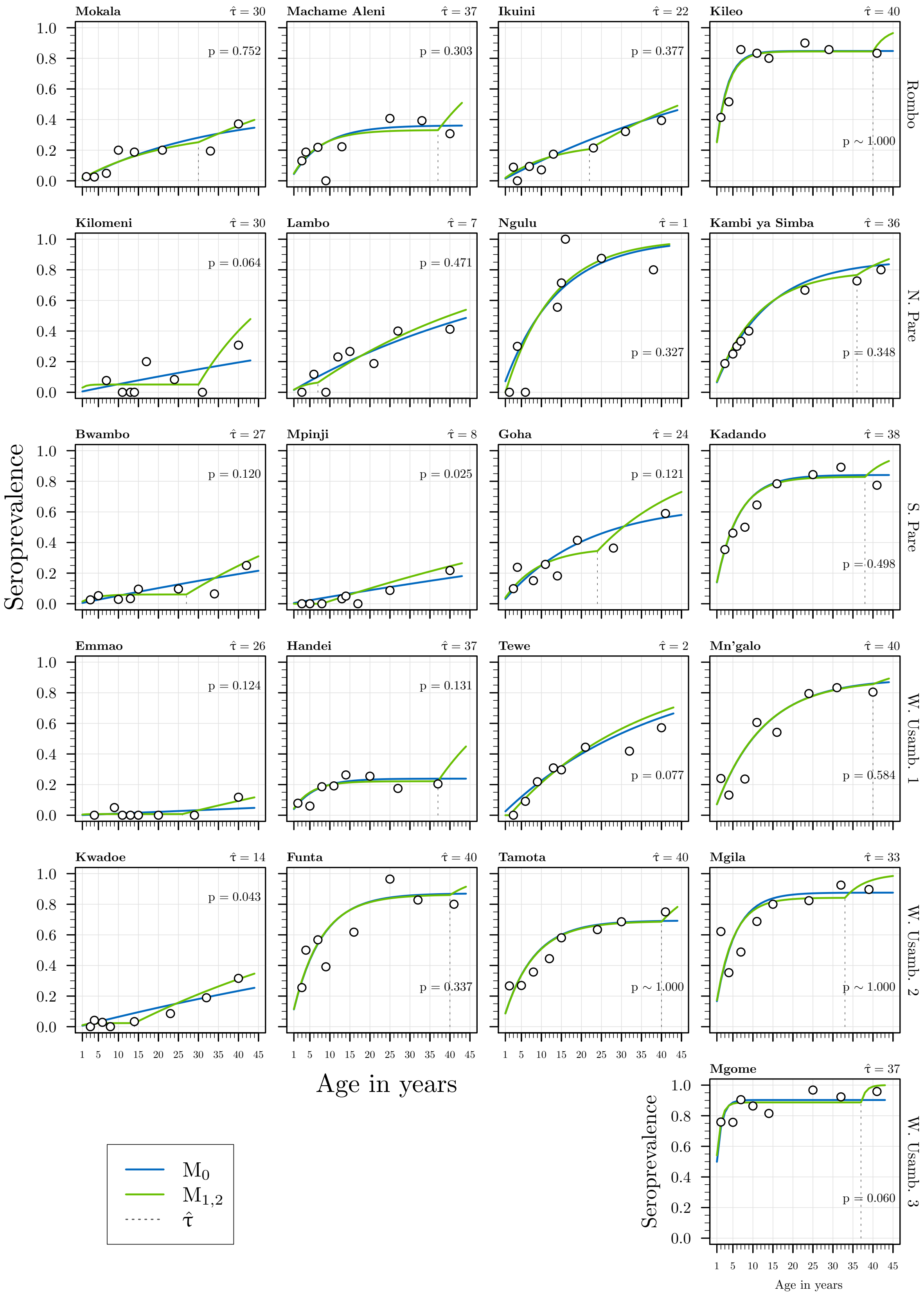

Figure 5.2: Fits for the estimated MSP1 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and $M_{1,2}$ (green lines), with the cutoff parameter of the latter signalled. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

5.1.3 Testing change in SCR: M_0 *vs.* M_2

With a possible change in SRR discarded for most villages, model M_0 was compared against M_2 to test if the SCR in each village was reasonably constant throughout the past years (Table 5.2 for MSP1, and from Appendices E, Tables 5 and 6 for MSP2 and AMA1, respectively). Model M_2 was developed assuming a change in SCR occurred some time before the sampling. Based on the frequency of antigens detected at different ages, model M_2 was able to estimate how long ago that change occurred, assuming $1 \leq \widehat{\tau}^* \leq 40$.

The estimated seroprevalence curves produced by this model suggested that a change in transmission might have occurred in various villages (Figure 5.3 for the estimated seroprevalence from MSP1, and Figures 2 and 4 from Appendices F and G, respectively, for estimated seroprevalence from MSP2 and AMA1). However, from the likelihood ratio tests, the apparently evident changes in SCR seen in model M_2 were in fact only statistically significant for a few villages. Results from the MSP1 antigen identified a significant change in SCR in Kwadoe, Tamota, Mgila (all from transect West Usambara 2), and Mgome (West Usambara 3). Estimates for Kwadoe ($\widehat{\tau}^* = 30$) indicated the SCR reduction occurred approximately between 27 and 32 years before sampling (Figure 5.4). The remaining three villages estimated the cutoff for change in SCR between one and two years ($\widehat{\tau}^* = 1$). For the MSP2 data set, villages Bwambo ($\widehat{\tau}^* = 37$) and Mpinji ($\widehat{\tau}^* = 1$) from the South Pare transect, and Funta ($\widehat{\tau}^* = 6$) and Tamota ($\widehat{\tau}^* = 1$), both from West Usambara 2, were identified. In Bwambo, the SCR reduction took place between 35 and 39 years before the survey sampling. Similar to the lower altitude villages identified using the previous antigen, the changes in Mpinji and, once again, in Tamota were estimated to approximately occur just one to two years before the data collection. And in Tamota, a the gange in SCR occurred approximately six years before the study. Likelihood ratio tests from the AMA1 antigen suggested that only individuals from the village Machame Aleni ($\widehat{\tau}^* = 20$), from the transect Rombo, were exposed to a significant reduction in SCR during the last 40 years. This change was estimated to happen approximately 20 years before sampling.

Table 5.2: Comparative analysis of the results from models M_0 and M_2, referring all 21 villages, considering MSP1 individual status as the outcomes. Model M_0 assumes constant SCR (λ) and SRR (ρ) for all ages. Model M_2 also assumes constant SRR, and a change in SCR after the cutoff parameter, τ^*. logL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using the profile likelihood method. p-value is associated with the log-likelihood ratio test comparing both models.

Transect	Village	Model M_0			Model M_2					p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}_1$ (95% CI)	$\hat{\lambda}_2$ (95% CI)	$\hat{\rho}$ (95% CI)	$\hat{\tau}^*$	logL	
Rombo	Mokala	0.014 (0.008, 0.030)	0.016 (0.000, 0.093)	-45.68	100.664 (0.006, >10)	0.012 (0.004, 0.031)	0.190 (0.000, 0.243)	8	-44.43	0.287
	Machame Aleni	0.047 (0.024, 0.140)	0.083 (0.020, 0.372)	-56.54	54.484 (0.000, >10)	0.045 (0.026, 0.085)	0.107 (0.031, 0.203)	14	-55.23	0.270
	Ikuini	0.014 (0.010, 0.038)	0.000 (0.000, 0.102)	-58.55	0.011 (0.006, 0.025)	0.051 (0.012, 0.116)	0.000 (0.000, 0.053)	1	-56.88	0.188
	Kileo	0.307 (0.205, 0.499)	0.055 (0.023, 0.125)	-44.77	236.197 (0.122, >10)	0.287 (0.184, 0.449)	0.077 (0.031, 0.127)	4	-43.83	0.391
N. Pare	Kilomeni	0.005 (0.002, 0.038)	0.000 (0.000, 0.351)	-18.93	0.027 (0.000, >10)	0.004 (0.001, 0.013)	0.003 (0.000, 0.103)	27	-17.84	0.336
	Lambo	0.015 (0.010, 0.041)	0.001 (0.000, 0.104)	-37.86	78.671 (0.000, >10)	0.010 (0.002, 0.035)	0.125 (0.000, 0.180)	10	-36.47	0.249
	Ngulu	0.074 (0.046, 0.159)	0.000 (0.000, 0.058)	-13.79	58.997 (0.071, >10)	0.038 (0.009, 0.105)	0.018 (0.000, 0.046)	13	-11.88	0.148
	Kambi ya Simba	0.067 (0.039, 0.122)	0.010 (0.000, 0.059)	-27.45	44.179 (0.000, >10)	0.068 (0.040, 0.112)	0.023 (0.000, 0.055)	17	-27.02	0.651
S. Pare	Bwambo	0.005 (0.003, 0.017)	0.000 (0.000, 0.126)	-36.14	20.786 (0.000, >10)	0.006 (0.003, 0.011)	0.039 (0.000, 0.070)	35	-33.49	0.071
	Mpinji	0.005 (0.002, 0.009)	0.000 (0.000, 0.055)	-21.75	1.962 (0.008, >10)	0.003 (0.001, 0.008)	0.058 (0.000, 0.092)	24	-18.88	0.057
	Goha	0.032 (0.021, 0.054)	0.017 (0.000, 0.070)	-58.55	0.019 (0.012, 0.036)	0.061 (0.029, 0.102)	0.000 (0.000, 0.034)	2	-56.59	0.141
	Kadando	0.152 (0.108, 0.230)	0.029 (0.011, 0.068)	-57.05	0.093 (0.046, 0.161)	0.343 (0.164, 0.575)	0.018 (0.000, 0.046)	1	-54.54	0.081
W. Usamb. 1	Emmao	0.001 (0.000, 0.127)	0.000 (0.000, >10)	-10.86	24.373 (0.000, >10)	0.001 (0.000, 0.005)	0.081 (0.000, 0.219)	30	-9.35	0.221
	Handei	0.044 (0.021, 0.128)	0.139 (0.040, 0.520)	-57.56	117.997 (0.000, >10)	0.047 (0.022, 0.100)	0.282 (0.000, 0.434)	7	-56.68	0.415
	Tewe	0.026 (0.020, 0.040)	0.000 (0.000, 0.032)	-64.19	0.048 (0.026, 0.089)	0.000 (0.000, 0.022)	0.022 (0.000, 0.065)	3	-61.72	0.085
	Mn'galo	0.074 (0.056, 0.099)	0.009 (0.000, 0.026)	-67.23	50.536 (0.167, >10)	0.071 (0.054, 0.092)	0.022 (0.014, 0.034)	15	-64.46	0.063
W. Usamb. 2	Kwadoe	0.007 (0.004, 0.011)	0.000 (0.000, 0.032)	-37.52	2.652 (0.299, >10)	0.004 (0.002, 0.008)	0.038 (0.026, 0.056)	30	-31.49	0.002
	Funta	0.121 (0.088, 0.172)	0.018 (0.005, 0.045)	-47.92	59.081 (0.000, >10)	0.116 (0.087, 0.157)	0.022 (0.011, 0.042)	13	-46.53	0.249
	Tamota	0.093 (0.063, 0.152)	0.041 (0.014, 0.100)	-64.58	0.046 (0.023, 0.084)	0.243 (0.128, 0.389)	0.016 (0.000, 0.051)	1	-60.30	0.014
	Mgila	0.184 (0.131, 0.286)	0.026 (0.008, 0.070)	-64.36	0.062 (0.038, 0.110)	0.579 (0.378, 0.807)	0.002 (0.000, 0.023)	1	-53.04	<0.001
W. Usamb. 3	Mgome	0.727 (0.398, 4.520)	0.078 (0.023, 0.482)	-34.25	0.088 (0.023, 0.432)	1.310 (0.748, 1.847)	0.010 (0.000, 0.088)	1	-30.86	0.034

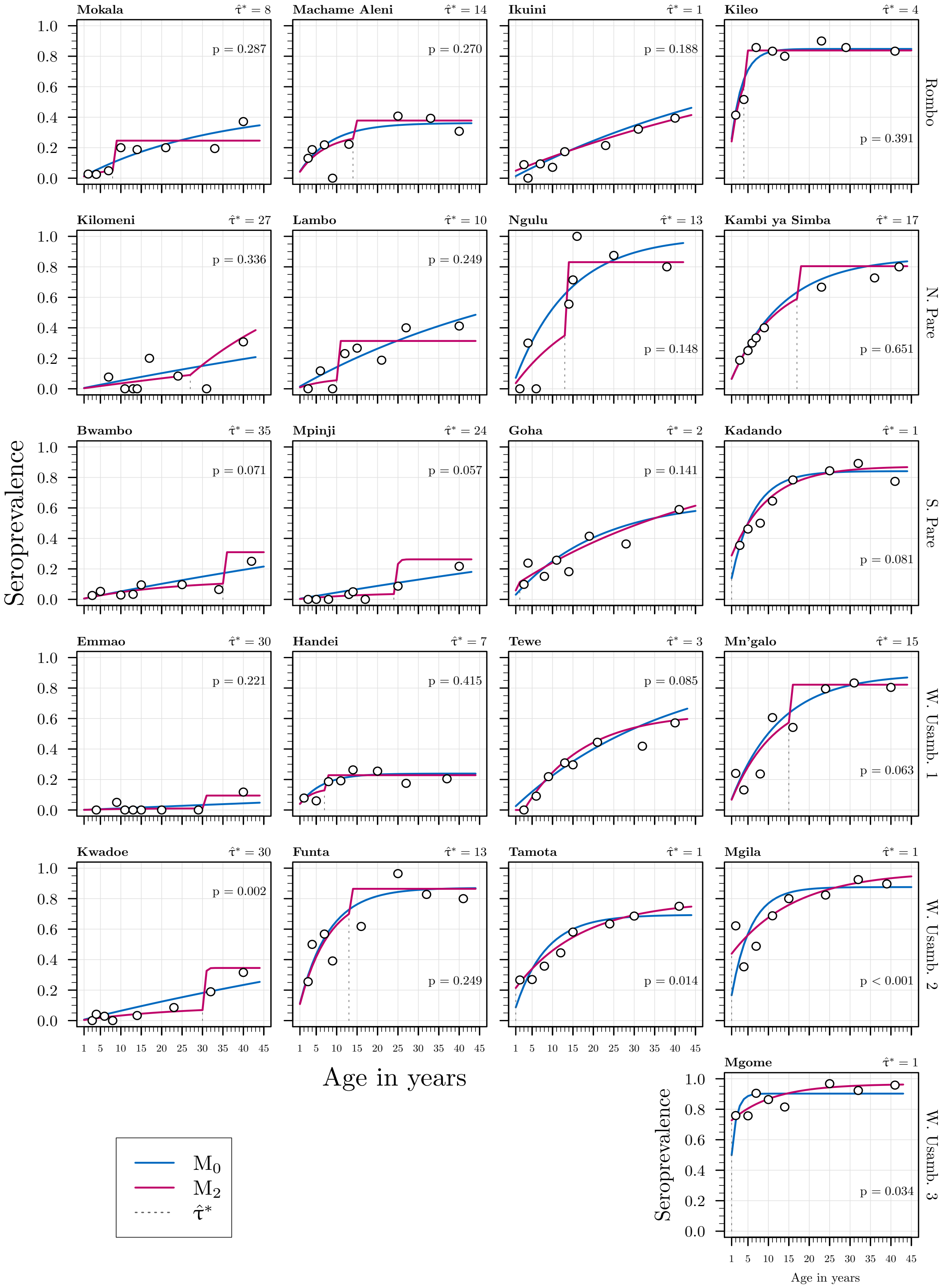

Figure 5.3: Fits for the estimated MSP1 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and M_2 (light red lines), with the cutoff parameter of the latter signalled, identifying the change in SCR happening in years before sampling. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

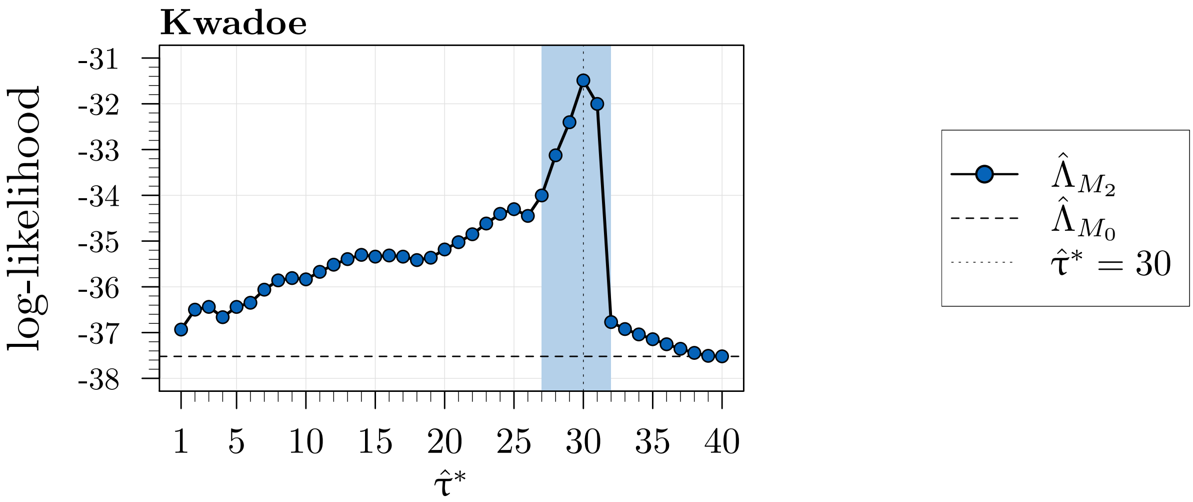

Figure 5.4: Example of the profile likelihood methodology implemented to estimate parameter values from village Kwadoe in model M_2. Parameters that maximised the log-likelihood were estimated by gradually increasing one unit in the cutoff parameter τ^* and estimate the remaining parameters λ_1, λ_2, and ρ at each level, via maximum likelihood. Estimate $\widehat{\tau}^* = 30$ (thin dashed line) returns the maximum value of log-likelihood and all the associated parameters. The Horizontal dashed line represents the maximum log-likelihood estimated using model M_0, under the same data set.

5.2 Antibody seroprevalence as indicator of malaria transmission intensity

With the more parsimonious serology-based model for each one of the studied villages identified, inferences about possible patterns in transmission intensity and how the seroprevalence evolved over time were performed. Irrespective of the malarial antigen analysed, most villages were best described considering SCR and SRR as constant transition rates across all ages (model M_0). An overall analysis on the model estimates implied a noticeable influence of altitude on the SCR, proxy for transmission intensity (Figure 5.5). As seen when describing the age-dependent model $M_{1,2}$, SCR estimates suggested an inverse relation with altitude. Similarly, the different $\widehat{\lambda}$ estimates from M_0 presented moderate correlation with altitude ($r^2_{MSP1} = 0.46$, $r^2_{MSP2} = 0.42$, and $r^2_{AMA1} = 0.48$). The analysis between the estimates produced with different antigens corroborated a conclusion made in Chapter 4, indicating AMA1 as the most immunogenic antigen of the three. The SCR estimated when using the serological data for this antigen were consistently higher across all villages, when compared to the remaining antigens. Meaning that even at high altitudes (lower transmission intensities), some levels of AMA1 antigens could be more easily detected in the individual's blood stream. The results suggested MSP1 to be the less immunogenic, confirming the results observed during the exploratory analysis (Table 2.3 from Chapter 2).

Focusing on different transects encompassing clusters of altitude-defined villages, as well as some population genetic traits, transects from the Tanga region (West Usambara 1, 2, and 3) suggested higher estimated levels of transmission intensity. More often villages from this region experienced a significant change in SCR some time before the sampling collection. West Usambara 2 and

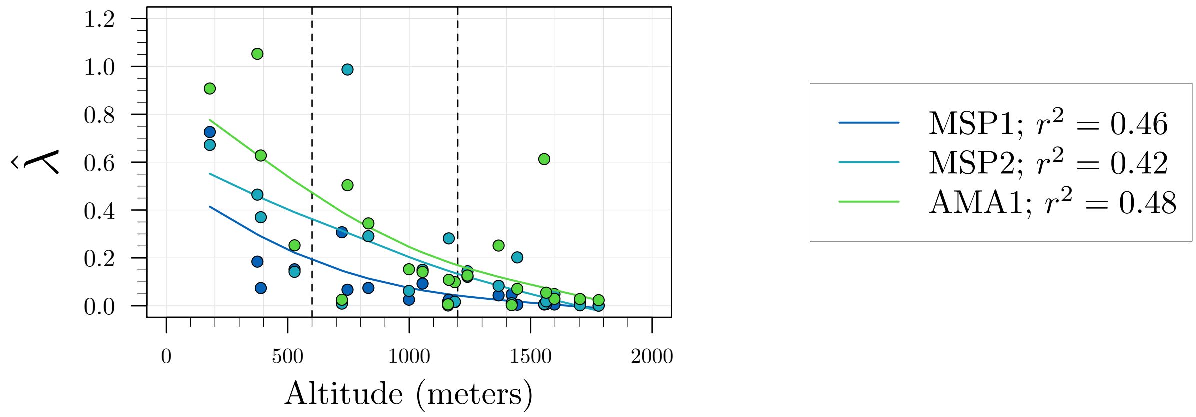

Figure 5.5: Estimated SCR for each village using the RCM M_0. Each coloured point represents the transmission intensity estimated for each antigen data (MSP1, MSP2, and AMA1), as function of the respective village's altitude. The coloured lines are but an emphatic description of the transmission's trend. Vertical dashed lines represent the defined altitude limits, distinguishing between high (>1200m), medium (600m−1200m) and low (<600m) altitude villages.

West Usambara 3 (village Mgome), had their current estimates for SCR ($\widehat{\lambda}$ or $\widehat{\lambda}_2$, depending on the selected model) generally above 0.100 in all data sets. Transects from the Kilimanjaro region (transects Rombo, North, and South Pare) suggested lower estimates for transmission intensity. Despite having similar structures in altitude, when compared to the villages from Tanga, the Kilimanjaro transects were located more inland, where a lesser humid climate is expected and could have presented an influence on the *Anopheles* mosquitoes capacity. The estimated values from Rombo relatively to the MSP1 data could suggest some affinity from individuals of the Wachaga ethnic group to produce the specific antibodies for the antigen. The transmission intensities estimated in villages from this transect were notoriously higher, presenting similar levels to villages with higher measures of prevalence of infection. However, the there was no strong discernible differences other than altitude and its effects on climate, that could indicate the reason for the different transmission intensities estimated.

5.3 Summary

Comparing different nested RCMs to test biological and epidemiological hypotheses came to show that although seemingly close to a more realistic scenario, the effects of acquired immunity (model $M_{1,2}$), or past impacts from possible control measures (model M_2), did not present an impactful statistically significance, when applied to the data. Model M_0 significantly described most of the studied villages. Some exceptions were the villages Tamota and Mgila, best described by the RCM M_2, estimating that a recent change in SCR had occurred in the past years. Model $M_{1,2}$ was also significant for some sites, describing villages located at intermediate and higher altitudes, such as Mpinji and Goha from the South Pare transect.

Correlation studies between $M_{1,2}$ and the more parsimonious M_0 suggested that neglecting to consider the age-dependent immunological uplift creates major underestimations in the SRR estimates, when considering the antigens MSP1 and MSP2. The simpler model M_0, with generally higher log-likelihoods than the remaining, more complex RCMs, is then considered the best model for the analysed data. This conclusion implies that a more generalist model will prevail over other, more specialised proposals, when the population sample sizes under analysis do not provide sufficient data that would allow better estimations of parameters from the different models. Results obtained by model M_0 demonstrated the dependency between the annual rate of seroconversion and altitude, with estimates for the AMA1 antigen being consistently higher than the others.

Chapter 6

Discussion

6.1 Project overview

The main objective of this project was to estimate and describe malaria transmission intensity, assessing the heterogeneity values from different sites. Using a benchmark cross-sectional survey data set from 21 Northeast Tanzanian villages with different endemic malaria intensities, various statistical methods were applied and tested. First, based on the recorded infections amongst individuals from the different population cohorts, generalised linear models (GLMs) were built (Chapter 4). The models were set to infer about the primary transmission determinants influencing prevalence of infection. Using different comparison methods and goodness-of-fit test statistics, the best model was selected. This structure when applied to a model described prevalence of infection through significant determinants such as altitude, transect, ethnic, gender, age groups, and the presence of antibodies for malaria antigens. Associating variables altitude and age group was also important to recreate a simplified categorical prevalence peak-shift, typically noticed when studying malaria in age-defined populations. The relative influence each demographical and exposure determinant had on prevalence of infection was assessed by the analyses of the model's odds ratios. The results corroborated what previous literature had described, relating altitude as proxy for malaria transmission intensity [27, 45], and identifying the importance of individual characteristic, such as age group [53] or ethnicity (genetic background), when inferring about risk of *P. falciparum* malaria infection. Results shown by the three exposure antigens *MSP1*, *MSP2*, and *AMA1* suggested that their presence consistently increased the odds of infection, thus evidencing their importance as immunological and hazard indicators as a consequence to the exposure to *P. falciparum* parasites. For the most immunogenic case, the sheer presence of the AMA1 antigen detected in children living in Mgome – the most vulnerable age group and the village with the highest registered prevalence, 50.67% – increased the odds of infection from 61% up to 136%. This sensitivity for the exposure antigens to indicate individuals at odds based on their serological status was then used to measure the exposure to malaria parasites.

With some of the studied villages in a state of malaria pre-elimination – potentially presenting recurrent asymptomatic cases – the use of exposure antigens could bring information to more accurately measure malaria transmission intensity. Using the serological status for the three antigens as outcome of interest, reverse catalytic models (RCMs) were applied, assuming age as a proxy for time of exposure (Chapter 5). Under different biological and epidemiological viewpoints, distinct RCMs were used. Their sero-epidemiological results were compared, characterising the transmission intensity from each village. Based on the data sets, the majority of the villages transmission intensity was better characterised by the simpler RCM M_0. Estimates from this model, assuming both seroconversion and seroreversion rates (SCR and SRR) as constant across all ages, described the relations between transmission intensity and altitude, with SCR, proxy for transmission intensity, decreasing with altitude. The results also showed how SRR was affected by the transmission intensity levels, reaching values close to zero in response to low estimates of SCR. This might be due to the fact that in low transmission intensity villages, the transition into a antigen seropositive state tends to becomes a rare event. With few seropositive individuals, the rate for seroreversion and antibody waning is expected to be close to non existent.

6.2 Epidemiological implications of the results

Recently, sero-epidemiological studies from longitudinal surveys have shown that some malaria antigens express a decrease in SRR with age [19]. These results came to contradict the overall assumption of constant SRR across all ages (assumed by the already published models M_0 and M_2). This age-dependent SRR reduction was proposed in model $\mathrm{M}_{1,2}$, with limited success in intermediate altitude villages, but mostly rejected for the simpler model M_0 regardless the antigens tested. The results suggested that although biologically plausible, there was no sufficient information granted from the data that would allow the statistical acceptance of $\mathrm{M}_{1,2}$. However, correlation tests showed underestimation of SCR from model M_0, relative to $\mathrm{M}_{1,2}$. Estimates for SRR were different depending on the antigen, although values of $\widehat{\rho}_1$ could be tendentiously higher to compensate for the sudden decrease to $\rho_2 = 0$, given $\widehat{\tau}$.

The comparison of SCR estimates showed good correlations. However, the small underestimations from M_0 (3% to 11% depending on the antigen) could become consequential in a scenario of low transmission settings, where accuracy is of most importance. Estimates for SCR are usually the focused results when assessing malaria transmission intensity. This rate can be representative of the force of infection [21], and its values have a clear transitional interpretation to more traditional measures, such as the entomological inoculation rate (Table 1 from Appendices) [18, 45]. With more areas reaching low transmission intensity, underestimating seroprevalence could produce misleading information, or even originate false sense of stability under a low transmission rate that otherwise could be taken for a somewhat more alarming event.

Model M_2 was also used to estimate heterogeneity in malaria transmission. Since its proposal, M_2 has been used as a complement to M_0 [54] and applied to estimate and monitor the effectiveness of campaigns for malaria control [23, 24]. No literature was found indicating that specific actions for control of malaria infection took place on the studied villages. Model M_2 was then applied under the hypothesis of possible historical changes in transmission intensity, being tested against M_0 for the significance of such changes. The comparison results suggested only few villages went through statistically significant SCR changes in past decades, and consequentially, their transmission intensity. Other results suggested some villages had only recently changed their exposure rate (villages where $\widehat{\tau}^{*} = 1$), consequently changing its SCR. These sites were mostly located at low altitudes, with medium to high transmission settings. For these situations, the parametric restriction proposed for model M_2 ($\lambda_1 \geq \lambda_2$) was even disregarded, with estimates for $\widehat{\lambda}_2$ presenting higher values than $\widehat{\lambda}_1$. One interpretation could be the detection of maternal antibodies that endured for more than the first year of a child's life. Other hypothesis could be that with a widespread infection across all ages in low altitude villages, children between one and two years old represented the interval where model M_2 identified the more considerable change in SCR, representing when the infants first experiment an exposure to the malaria parasites. Under this assumption of early SCR change, the generated seroprevalence curves did not present the characteristic biphasic behaviour (as seen in Figure 5.3).

The reduced number of significant changes in SCR when AMA1 antigen outcomes were used – only one village identified – have been reported in analyses done in *P. vivax* parasites [23]. Due to its high immunogenicity (population reports for more than 80% seropositive individuals by 20 years old [19]), it is possible that despite an historical change, some molecular specialised tests still detected reasonable amounts of this antigen.

6.3 Statistical and epidemiological limitations

During this project, important limitations were identified. The first was during the implementation of the simple GLMs to the data, when apparent more complex dependencies between variables were noticeable. Within each transect, villages shared not only the same ethnic group, but also the described geographical proximity when in relation to the other studied sites. When using the GLMs, these implications were disregarded with models assuming independence between the parameters of all villages. The logistic `fit10` selected in Chapter 4, performed well when inferring the odds of malaria infection, as well as justifying the heterogeneity measured. The model was used for its descriptive abilities, being refrained from use as a broad predictive tool due to this limitation that could, eventually, be amended by focusing on the dependencies between variables. These relations could possibly have been taken into account through use of generalised linear multilevel models (GLMMs). The GLMMs delineate different hierarchical levels where the villages' systematic structures would be nested within a random factor level that could be defined by the transect categorical variable, *Transect*. However, the development of these more models was out of the scope of this thesis.

The RCMs also presented limitations as they are infinite population models. Despite producing informative estimates, the application of the models to the limited sample sizes recorded in each village might not have been enough to statistically discern between each proposed model. Furthermore, the data in which the RCMs were applied to had a specific age structure with children between 1 and 4 years being oversampled in relation to other ages, because the original study used the effect of defined age groups in its survey [18]. Possibly, by increasing the sample size of the study, model $M_{1,2}$ would have a justification to be applied and produce more consistent significant results. When applying this model, some villages certainly presented limited information to accurately estimate SRR (Tables 5.1 and Tables 5 and 6 from Appendices D). These situations resulted in non expected estimated values of $\widehat{\rho}_1$ and respective confidence intervals, showing the difficulty of estimating the parameter when using samples form cross-sectional surveys. The difficulty can also create some uncertainties when estimating the change point parameter $\widehat{\tau}$ in the profile likelihood method.

6.4 Further extensions

If the intention was to measure malaria transmission intensity based only on models predicting prevalence of infection through the defined variables – instead of seroprevalence via the RCMs – some specifications or generalisations to the used linear model could be applied. The generalised estimating equations (GEEs), being an extension of the GLMs, were developed specially to analyse discrete clustered data with correlated dependencies influencing the outcome. Similar to the GLMs, the GEEs return responses that can be viewed as directly related to the more traditional tools to measure malaria transmission. This set of models has increasingly been used to study public health longitudinal surveys with multiple cohorts [55, 56]. Applying the GEEs to the Northeastern Tanzania data, clusters for transect and geographically closer villages could be created, adding an estimated correlation matrix relating the outcomes. The downside of this method could be the sample sizes of each related village. The sheer number of individuals would create large correlation matrices, limiting the correlation structures given by the GEEs to perform the estimates. If focused on a single village under more precise data, the GEEs could potentially further generate correlation between household families.

A side project of this thesis is still under development, with the intent of extend the understanding of the statistical power from the models assuming age-dependent change in SRR (models $\mathrm{M}_{1,1}$ and $\mathrm{M}_{1,2}$). Due to the difficulty of rejecting M_0 for $\mathrm{M}_{1,2}$ in the populations assessed in this thesis, the project will sample different simulated populations with different levels of initially defined parameters λ, ρ_1, ρ_2, and τ. The parameter values are chosen in order to directly relate to different possible values of the EIR measure (Table 1 from Appendices). A thousand populations with different sizes (1000, 5000, 10 000, 25 000, 50 000, and 100 000 individuals) will be simulated and analysed using models $\mathrm{M}_{1,1}$ and $\mathrm{M}_{1,2}$, calculating the true SCR. Then, model M_0 will be instantiated with the true SCR onto the same simulated populations, estimating new values for SCR and SRR. The results from the different models will be compared through likelihood ratio tests, with the proportion of rejections of M_0 for $\mathrm{M}_{1,1}$ or $\mathrm{M}_{1,2}$ at each simulated data, indicating the power of the age-dependent SRR model. The results from this project could increase the knowledge on the estimation of SRR, proving this rate to be of importance when performing sero-epidemiological studies.

6.5 Conclusion

The research done to study malaria have many fronts. Various approaches can be applied, all with the main objective of eradicating malaria infections from burdened sites. The production of efficient vaccines is still under developing, however, actions such as distribution of treated mosquito nets and campaigns to directly control the mosquito populations have produced great results, while improving the life conditions of the inhabitants living on those affected sites.

Statistical models, such as spatial, temporal or stochastic models, are important tools to efficiently and quantitatively deal with malaria and its dynamics. They allow for approaching the field of biology in a more controlled way, being widely applied in epidemiology and public health sciences. Sero-epidemiology is in this scenario an innovative tool, presenting advantages in the implementation methods on the field and adapting to the recent decreases in malaria transmission intensity. The RCMs used and compared throughout this thesis were able to estimate the annual seroconversion rate, as well as the seroreversion rate across various sites. These rates could bring new light onto the disease dynamics modelled by different variables such as altitude, age or genetics. In the analyses, the simpler and more broadly used RCM was not rejected when tested against the others. However, one might suggest that in alternative scenarios, models such as the age-dependent SCR model may be more suited. The continuous innovation of these techniques could only help to further explore the possibilities to positively control one of the most important diseases humanity has faced.

References

[1] Vitoria, M., Granich, R., Gilks, C., Gunneberg, C., Hosseini, M., Were, W., Raviglione, M., and De Cock, K. (2009) The Global Fight Against HIV/AIDS, Tuberculosis, and Malaria: Current Status and Future Perspectives. *American Journal of Clinical Pathology,* **131**(6), 844–848.

[2] Geneva: World Health Organization World malaria report 2017.

[3] Warrell, D. and Gilles, H. (2002) Essential malariology, CRC Press, 4th edition.

[4] Ross, R. (1897) Observations on a condition necessary to the transformation of the malaria crescent. *British Medical Journal,* **1**(1883), 251–255.

[5] Geneva: World Health Organization (2017) A framework for malaria elimination.

[6] World Health Organization (2015) Guidelines for the treatment of malaria, World Health Organization, 3rd edition.

[7] Perlmann, P. and Troye-Blomberg, M. (2002) Malaria immunology, Vol. 80, Karger Medical and Scientific Publishers, 2nd edition.

[8] Carter, R. and Mendis, K. (2002) Evolutionary and historical aspects of the burden of malaria. *Clinical Microbiology Reviews,* **15**(4), 564–594.

[9] Snow, R. and Marsh, K. (2002) The consequences of reducing transmission of *Plasmodium falciparum* in Africa. *Advances in Parasitology,* **52**, 235–264.

[10] Kitua, A., Ogundahunsi, O., Lines, J., and Mgone, C. (2011) Conquering malaria: Enhancing the impact of effective interventions towards elimination in the diverse and changing epidemiology. *Journal of Global Infectious Diseases,* **3**(2), 161–165.

[11] Pigott, D., Atun, R., Moyes, C., Hay, S., and Gething, P. (2012) Funding for malaria control 2006–2010: a comprehensive global assessment. *Malaria Journal,* **11**(1), 246.

[12] World Health Organization (2016) WHO malaria terminology.

[13] Doolan, D. (2002) Malaria methods and protocols, Vol. 72, Springer Science & Business Media, 1st edition.

[14] Geneva: World Health Organization (2015) Global technical strategy for malaria 2016-2030.

[15] Cameron, E., Battle, K., Bhatt, S., Weiss, D., Bisanzio, D., Mappin, B., Dalrymple, U., Hay, S., Smith, D., Griffin, J., Wenger, E., Eckhoff, P., Smith, T., Penny, M., and Gething, P. (2015) Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nature Communications,* **6**, 8170.

[16] O'Meara, W., Collins, W., and McKenzie, F. (2007) Parasite prevalence: a static measure of dynamic infections. *The American Journal of Tropical Medicine and Hygiene,* **77**(2), 246–249.

[17] Corran, P., Coleman, P., Riley, E., and Drakeley, C. (2007) Serology: a robust indicator of malaria transmission intensity?. *Trends in Parasitology,* **23**(12), 575–582.

[18] Drakeley, C., Corran, P., Coleman, P., Tongren, J., McDonald, S., Carneiro, I., Malima, R., Lusingu, J., Manjurano, A., Nkya, W., Lemnge, M., Cox, J., Reyburn, H., and Riley, E. (2005) Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure. *Proceedings of the National Academy of Sciences,* **102**(14), 5108–5113.

[19] Ondigo, B., Hodges, J., Ireland, K., Magak, N., Lanar, D., Dutta, S., Narum, D., Park, G., Ofulla, A., and John, C. (2014) Estimation of recent and long-term malaria transmission in a population by antibody testing to multiple *Plasmodium falciparum* antigens. *The Journal of Infectious Diseases,* **210**(7), 1123–1132.

[20] van den Hoogen, L., Griffin, J., Cook, J., Sepúlveda, N., Corran, P., Conway, D., Milligan, P., Affara, M., Allen, S., Proietti, C., Ceesay, S., Targett, G., D'Alessandro, U., Greenwood, B., Riley, E., and Drakeley, C. (2015) Serology describes a profile of declining malaria transmission in Farafenni, The Gambia. *Malaria Journal,* **14**(1), 416.

[21] Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., and Beutels, P. (2012) Modelling infectious disease parameters based on serological and social contact data: a modern statistical perspective, Vol. 63, Springer Science & Business Media, 1st edition.

[22] Nkumama, I., O'Meara, W., and Osier, F. (2017) Changes in malaria epidemiology in Africa and new challenges for elimination. *Trends in Parasitology,* **33**(2), 128–140.

[23] Cook, J., Reid, H., Iavro, J., Kuwahata, M., Taleo, G., Clements, A., McCarthy, J., Vallely, A., and Drakeley, C. (2010) Using serological measures to monitor changes in malaria transmission in Vanuatu. *Malaria Journal,* **9**(1), 169.

[24] Cook, J., Kleinschmidt, I., Schwabe, C., Nseng, G., Bousema, T., Corran, P., Riley, E., and Drakeley, C. (2011) Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, equatorial Guinea. *PLoS ONE,* **6**(9), e25137.

[25] Hay, S., Smith, D., and Snow, R. (2008) Measuring malaria endemicity from intense to interrupted transmission. *The Lancet Infectious Diseases,* **8**(6), 369–378.

[26] Bruce-Chwatt, L., Draper, C., and Konfortion, P. (1973) Seroepidemiological evidence of eradication of malaria from Mauritius. *The Lancet,* **302**(7828), 547–551.

[27] Drakeley, C., Carneiro, I., Reyburn, H., Malima, R., Lusingu, J., Cox, J., Theander, T., Nkya, W., Lemnge, M., and Riley, E. (2005) Altitude-dependent and -independent variations in *Plasmodium falciparum* prevalence in northeastern Tanzania. *The Journal of Infectious Diseases,* **191**(10), 1589–1598.

[28] Enevold, A., Alifrangis, M., Sanchez, J., Carneiro, I., Roper, C., Børsting, C., Lusingu, J., Vestergaard, L., Lemnge, M., Morling, N., Riley, E., and Drakeley, C. (2007) Associations between $\alpha+$-thalassemia and *Plasmodium falciparum* malarial infection in northeastern Tanzania. *The Journal of Infectious Diseases,* **196**(3), 451–459.

[29] Sepúlveda, N., Manjurano, A., Campino, S., Lemnge, M., Lusingu, J., Olomi, R., Rockett, K., Hubbart, C., Jeffreys, A., Rowlands, K., Clark, T., Riley, E., and Drakeley, C. (2017) Malaria Host Candidate Genes Validated by Association With Current, Recent, and Historical Measures of Transmission Intensity. *The Journal of Infectious Diseases,* **216**(1), 45–54.

[30] Bosomprah, S. (2014) A mathematical model of seropositivity to malaria antigen, allowing seropositivity to be prolonged by exposure. *Malaria Journal,* **13**(1), 12.

[31] Marsh, K., Forster, D., Waruiru, C., Mwangi, I., Winstanley, M., Marsh, V., Newton, C., Winstanley, P., Warn, P., Peshu, N., Pasvol, G., and Snow, R. (1995) Indicators of life-threatening malaria in African children. *New England Journal of Medicine,* **332**(21), 1399–1404.

[32] Reddy, S., Anders, R., Beeson, J., Färnert, A., Kironde, F., Berenzon, S., Wahlgren, M., Linse, S., and Persson, K. (2012) High affinity antibodies to *Plasmodium falciparum* merozoite antigens are associated with protection from malaria. *PLoS ONE,* **7**(2), e32242.

[33] Wong, J., Hamel, M., Drakeley, C., Kariuki, S., Shi, Y. P., Lal, A., Nahlen, B., Bloland, P., Lindblade, K., Were, V., Otieno, K., Otieno, P., Odero, C., Slutsker, L., Vulule, J., and Gimnig, J. (2014) Serological markers for monitoring historical changes in malaria transmission intensity in a highly endemic region of Western Kenya, 1994–2009. *Malaria Journal,* **13**(1), 451.

[34] Bousema, T., Youssef, R., Cook, J., Cox, J., Alegana, V., Amran, J., Noor, A., Snow, R., and Drakeley, C. (2010) Serologic markers for detecting malaria in areas of low endemicity, Somalia, 2008. *Emerging Infectious Diseases,* **16**(3), 392–399.

[35] Box, G., Hunter, S., and Hunter, W. (2005) Statistics for experimenters: design, innovation, and discovery, Wiley-Interscience New York, 2nd edition.

[36] Sicuri, E., Vieta, A., Lindner, L., Constenla, D., and Sauboin, C. (2013) The economic costs of malaria in children in three sub-Saharan countries: Ghana, Tanzania and Kenya. *Malaria Journal,* **12**(1), 307.

[37] The World Bank, (2017) Combined Project Information Document/Integrated Safeguards Data Sheet. (A/HRC/27/37).

[38] World Heath Organization (2013) An Epidemiological Profile of Malaria and its Control in Mainland Tanzania. Report funded by Roll Back Malaria and Department for International Development–UK, July 2013.

[39] Lindsay, S. and Birley, M. (1996) Climate change and malaria transmission. *Annals of Tropical Medicine and Parasitology,* **90**(5), 573–588.

[40] Casella, G. and Berger, R. (2002) Statistical inference, Duxbury Pacific Grove, CA, 2nd edition.

[41] Nelder, J. and Wedderburn, R. (1972) Generalized linear models.. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* **135**(3), 370–384.

[42] Andersson, H. and Britton, T. (2012) Stochastic epidemic models and their statistical analysis, Vol. 151, Springer Science & Business Media, 1st edition.

[43] Collet, D. (2003) Modelling Binary Data, Chapman & Hall, 2nd edition.

[44] Muench, H. (1959) Catalytic Models in Epidemiology, Harvard University Press, Cambridge, Mass, USA.

[45] Bødker, R., Akida, J., Shayo, D., Kisinza, W., Msangeni, H., Pedersen, E., and Lindsay, S. (2003) Relationship between altitude and intensity of malaria transmission in the Usambara Mountains, Tanzania. *Journal of Medical Entomology,* **40**(5), 706–717.

[46] Sepúlveda, N., Stresman, G., White, M., and Drakeley, C. (2015) Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of Immunology Research,* **2015**, 1–21.

[47] Cunha, M., Silva, E., Sepúlveda, N., Costa, S., Saboia, T., Guerreiro, J., Póvoa, M., Corran, P., Riley, E., and Drakeley, C. (2014) Serologically defined variations in malaria endemicity in Pará state, Brazil. *PLoS ONE,* **9**(11), e113357.

[48] Williams, B. and Dye, C. (1994) Maximum likelihood for parasitologists. *Parasitology Today,* **10**(12), 489–493.

[49] Hosmer Jr, D., Lemeshow, S., and Sturdivant, R. (2000) Applied logistic regression, Vol. 398, John Wiley & Sons, 2nd edition.

[50] Cressie, N. and Read, T. (1984) Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological),* **46**(3), 440–464.

[51] Binka, F., Maude, G., Gyapong, M., Ross, D., and Smith, P. (1995) Risk factors for child mortality in northern Ghana: a case-control study. *International Journal of Epidemiology,* **24**(1), 127–135.

[52] Shelton, J., Corran, P., Risley, P., Silva, N., Hubbart, C., Jeffreys, A., Rowlands, K., Craik, R., Cornelius, V., Hensmann, M., et al. (2015) Genetic determinants of anti-malarial acquired immunity in a large multi-centre study. *Malaria Journal,* **14**(1), 333.

[53] Carneiro, I., Roca-Feltrer, A., Griffin, J., Smith, L., Tanner, M., Schellenberg, J., Greenwood, B., and Schellenberg, D. (2010) Age patterns of malaria vary with severity, transmission intensity and seasonality in sub-Saharan Africa: a systematic review and pooled analysis. *PLoS ONE,* **5**(2), e8988.

[54] Dewasurendra, R., Dias, J., Sepúlveda, N., Gunawardena, G., Chandrasekharan, N., Drakeley, C., and Karunaweera, N. (2017) Effectiveness of a serological tool to predict malaria transmission intensity in an elimination setting. *BMC Infectious Diseases,* **17**(1), 49.

[55] Hanley, J., Negassa, A., Michael, E., and Forrester, J. (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology,* **157**(4), 364–375.

[56] Hubbard, A., Ahern, J., Fleischer, N., Van der Laan, M., Satariano, S., Jewell, N., Bruckner, T., and Satariano, W. (2010) To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology,* **21**(4), 467–474.

[57] Yman, V., White, M., Rono, J., Arcà, B., Osier, F., Troye-Blomberg, M., Boström, S., Ronca, R., Rooth, I., and Färnert, A. (2016) Antibody acquisition models: a new tool for serological surveillance of malaria transmission intensity. *Scientific Reports,* **6**, 19472.

Appendices

A Model M_0 derivation

Knowing seronegative individuals become seropositive at rate λ_t and seropositive individuals revert at rate ρ_t, the proportion of seropositive individuals in a cohort P is defined by the differential equation [57]

$$\frac{dP}{dt} = \lambda_t(1-P) - \rho_t P \ . \tag{1}$$

Considering model M_0, with constant transmission rates, $\lambda_t = \lambda$ and $\rho_t = \rho$, this equation can be solved to estimate the proportion of individuals of age t at each cross-section.

$$\begin{aligned} \int \frac{1}{\lambda(1-P) - \rho P}\, dP &= \int dt \Leftrightarrow \\ \Leftrightarrow \int \frac{1}{\lambda - P(\lambda+\rho)}\, dP &= \int dt \Leftrightarrow \\ \Leftrightarrow -\frac{1}{\lambda+\rho}\int \frac{-(\lambda+\rho)}{\lambda - P(\lambda+\rho)}\, dP &= \int dt \Leftrightarrow \\ \Leftrightarrow -\frac{1}{\lambda+\rho}\ln(\lambda - P(\lambda+\rho)) &= t + c \Leftrightarrow \\ \Leftrightarrow \ \ln(\lambda - P(\lambda+\rho)) &= -\,(\lambda+\rho)(t+c) \Leftrightarrow \\ \Leftrightarrow \ P &= \frac{\lambda - e^{\{-(\lambda+\rho)t-(\lambda+\rho)c\}}}{\lambda+\rho} \ . \end{aligned} \tag{2}$$

Since a true seropositive state could only be achieved by being exposed to the malaria parasites (not accounting for the maternal acquired antibodies), this model assumes that no individual is seropositive at the moment of birth, i.e. $P(0) = 0$, thus

$$\begin{aligned} \frac{\lambda - e^{\{-(\lambda+\rho)0-(\lambda+\rho)c\}}}{\lambda+\rho} &= 0 \Leftrightarrow \\ \frac{\lambda - e^{\{-(\lambda+\rho)c\}}}{\lambda+\rho} &= 0 \Leftrightarrow \\ \Leftrightarrow e^{\{-(\lambda+\rho)c\}} &= \lambda \Leftrightarrow \\ \Leftrightarrow c &= -\frac{\ln(\lambda)}{\lambda+\rho} \ , \end{aligned} \tag{3}$$

that can be directly applied onto previous equation (2),

$$\begin{aligned} P &= \frac{\lambda - e^{\{-(\lambda+\rho)t+\ln(\lambda)\}}}{\lambda+\rho} \\ &= \frac{\lambda - e^{\{-(\lambda+\rho)t\}}\lambda}{\lambda+\rho} \\ &= \frac{\lambda}{\lambda+\rho}\left(1 - e^{\{-(\lambda+\rho)t\}}\right) \ . \end{aligned} \tag{4}$$

B Relationship between SCR, EIR, and the cutoff for SRR reducion

Table 1: Relationship between SCR and the age in years at which SRR is expected to reduce from ρ_1 to ρ_2 (with $\rho_1 \geq \rho_2$). As previously described, estimates for SCR have a somewhat direct translation to the entomological inoculation rate measure (EIR), that identifies the number of infective bites received per person in a year, in a human population.

EIR	SCR, λ	Age of SRR reduction, $\hat{\tau}$
100	0.2900	3, 5
10	0.0969	5, 10
1	0.0324	5, 10
0.1	0.0108	10, 15, 20
0.01	0.0036	10, 15, 20

C $\mathbf{M}_{1,2}$ *vs.* $\mathbf{M}_{1,1}$

Table 2: Comparison between the two age-dependent SRR models $M_{1,2}$ and $M_{1,1}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-MSP1 antigen in samples from the 21 villages. Both models assume constant SCR (λ) in the population. Model $M_{1,1}$ assumes constant SRR (ρ_1) for ages below τ, and a lower rate (ρ_2) after cut-off. Model $M_{1,2}$ assumes $\rho_2 = 0$ after τ. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model $M_{1,2}$ with $M_{1,1}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits.

Transect	Village	Model $M_{1,2}$				Model $M_{1,1}$					p-value
		$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\tau}$	logL	$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\tau}$	logL	
Rombo	Mokala	0.016 (0.009, >5)	0.031 (0.000, >10)	30	-45.63	8.795 (0.010, >10)	>10 (0.000, >15)	29.815 (0.000, >10)	8	-44.39	0.115
	Machame Aleni	0.052 (0.027, >5)	0.104 (0.031, 0.460)	37	-56.01	0.187 (0.049, >10)	0.972 (0.135, >10)	0.427 (0.000, >10)	12	-53.91	0.040
	Ikuini	0.020 (0.011, >5)	0.061 (0.000, >10)	22	-58.16	0.548 (0.016, >10)	5.683 (0.018, >10)	1.358 (0.000, >10)	13	-56.29	0.053
	Kileo	0.298 (0.210, >5)	0.055 (0.026, 0.110)	40	-45.61	15.343 (0.278, >10)	17.536 (0.045, >15)	3.491 (0.000, 0.248)	4	-42.99	0.022
N. Pare	Kilomeni	0.046 (0.004, >5)	0.864 (0.000, >10)	30	-17.22	0.046 (0.004, >10)	0.864 (0.000, >10)	0.000 (0.000, 0.098)	30	-17.22	$\sim$1.000
	Lambo	0.019 (0.010, >5)	0.232 (0.000, >10)	7	-37.60	4.885 (0.013, >10)	>10 (0.000, >15)	11.526 (0.000, >10)	10	-36.15	0.089
	Ngulu	0.084 (0.052, >5)	>10 (0.000, >10)	1	-13.31	8.845 (0.470, >10)	>10 (1.510, >15)	1.824 (0.000, >10)	13	-11.70	0.073
	Kambi ya Simba	0.074 (0.043, >5)	0.019 (0.000, >10)	36	-27.01	0.074 (0.043, 0.136)	0.019 (0.000, 0.095)	0.000 (0.000, >10)	36	-27.01	$\sim$1.000
S. Pare	Bwambo	0.017 (0.004, >5)	0.265 (0.000, >10)	27	-34.93	4.216 (0.034, >10)	>10 (0.472, >15)	10.06 (>10, >15)	35	-33.45	0.085
	Mpinji	0.009 (0.005, >5)	>10 (0.134, >15)	8	-19.23	0.009 (0.005, 0.022)	>10 (>10, >15)	0.000 (0.000, >10)	8	-19.23	$\sim$1.000
	Goha	0.042 (0.025, >5)	0.073 (0.000, 0.196)	24	-57.35	0.066 (0.028, 0.288)	0.233 (0.011, 1.438)	0.062 (0.000, >10)	13	-56.70	0.254
	Kadando	0.156 (0.115, >5)	0.032 (0.014, 0.069)	38	-56.82	0.381 (0.182, 1.093)	0.436 (0.104, 1.545)	0.089 (0.000, >10)	8	-53.36	0.009
W. Usamb. 1	Emmao	0.006 (0.001, >5)	0.838 (0.000, >10)	26	-9.68	1.616 (0.002, >10)	>10 (0.110, >15)	15.481 (>10, >15)	31	-8.93	0.221
	Handei	0.049 (0.025, >5)	0.173 (0.062, 0.527)	37	-56.42	0.268 (0.049, >10)	3.655 (0.332, >10)	1.215 (0.000, >10)	5	-55.00	0.092
	Tewe	0.030 (0.023, >5)	>10 (0.000, >10)	2	-62.63	0.044 (0.027, 0.074)	>10 (>10, >15)	0.026 (0.000, >10)	3	-61.72	0.177
	Mn'galo	0.075 (0.058, >5)	0.010 (0.000, >10)	40	-67.08	0.160 (0.101, 0.271)	0.740 (0.252, 1.631)	0.036 (0.000, >10)	6	-61.53	0.001
W. Usamb. 2	Kwadoe	0.013 (0.007, >5)	0.558 (0.013, 2.794)	14	-35.47	0.932 (0.153, >10)	30.758 (4.129, >10)	1.863 (0.000, >10)	30	-33.00	0.026
	Funta	0.123 (0.093, 0.169)	0.020 (0.007, 0.045)	40	-47.46	0.244 (0.125, 0.592)	0.209 (0.037, 0.717)	0.041 (0.000, >10)	12	-45.27	0.036
	Tamota	0.092 (0.066, >5)	0.042 (0.017, 0.091)	40	-65.06	0.295 (0.112, >10)	0.558 (0.112, >10)	0.174 (0.000, >10)	11	-60.61	0.003
	Mgila	0.194 (0.140, >5)	0.036 (0.013, 0.091)	33	-64.45	16.502 (0.891, >10)	15.795 (0.785, >10)	2.796 (0.000, >10)	10	-50.62	<0.001
W. Usamb. 3	Mgome	0.835 (0.457, >5)	0.107 (0.035, >10)	37	-32.48	1.527 (0.582, >10)	0.335 (0.076, >10)	0.090 (0.000, >10)	13	-30.48	0.046

Table 3: Comparison between the two age-dependent SRR models $\mathrm{M}_{1,2}$ and $\mathrm{M}_{1,1}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-MSP2 antigen in samples from the 21 villages. Both models assume constant SCR (λ) in the population. Model $\mathrm{M}_{1,1}$ assumes constant SRR (ρ_1) for ages below τ, and a lower rate (ρ_2) after cut-off. Model $\mathrm{M}_{1,2}$ assumes $\rho_2 = 0$ after τ. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model $\mathrm{M}_{1,2}$ with $\mathrm{M}_{1,1}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits.

Transect	Village	Model $\mathrm{M}_{1,2}$				Model $\mathrm{M}_{1,1}$					p-value
		$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\tau}$	logL	$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\tau}$	logL	
Rombo	Mokala	0.004 (0.001, >5)	0.309 (0.000, >10)	16	-19.95	1.340 (0.001, >10)	>10 (0.000, >15)	>10 (0.000, 39.595)	19	-19.66	0.446
	Machame Aleni	0.053 (0.002, >5)	1.830 (0.014, 15.071)	40	-16.16	0.053 (0.002, >10)	1.826 (0.014, >10)	0.000 (0.000, >10)	40	-16.16	∼1.000
	Ikuini	0.009 (0.001, >5)	0.857 (0.000, >10)	31	-13.68	3.888 (0.001, >10)	>10 (0.000, >10)	0.000 (0.000, 5.196)	39	-13.62	0.729
	Kileo	0.014 (0.007, >5)	0.408 (0.000, >10)	9	-41.64	0.055 (0.009, >10)	1.916 (0.012, >10)	0.164 (0.000, >10)	11	-40.61	0.151
N. Pare	Kilomeni	0.008 (0.003, >5)	0.041 (0.000, >10)	39	-18.19	40.763 (0.004, >10)	>10 (0.000, >15)	0.000 (0.000, 0.6835)	11	-16.63	0.077
	Lambo	0.029 (0.015, >5)	0.707 (0.000, >10)	9	-33.77	0.167 (0.046, >10)	3.514 (0.491, >10)	0.217 (0.000, >10)	13	-32.12	0.069
	Ngulu	0.536 (0.190, >5)	0.313 (0.000, >10)	6	-5.33	0.536 (0.190, >10)	0.314 (0.000, >10)	0.000 (0.000, 0.157)	6	-5.33	∼1.000
	Kambi ya Simba	0.523 (0.214, >5)	0.219 (0.051, 0.720)	40	-30.68	1.602 (0.317, >10)	0.923 (0.118, >10)	0.668 (0.000, >10)	13	-28.92	0.061
S. Pare	Bwambo	0.057 (0.044, >5)	>10 (0.000, >10)	2	-57.15	0.072 (0.049, 0.111)	>10 (0.000, >15)	0.013 (0.000, >10)	2	-56.41	0.224
	Mpinji	0.475 (0.172, >5)	0.410 (0.103, 1.295)	34	-47.04	11.38 (0.534, >10)	15.982 (0.654, >10)	5.736 (0.000, >10)	13	-44.07	0.015
	Goha	0.205 (0.120, >5)	0.300 (0.152, 0.632)	40	-74.54	6.778 (0.303, >10)	20.784 (0.718, >10)	>10 (0.000, >15)	3	-70.64	0.005
	Kadando	0.148 (0.089, >5)	0.143 (0.068, 0.332)	40	-60.86	0.148 (0.089, 0.297)	0.143 (0.068, 0.342)	0.000 (0.000, 82.982)	40	-60.86	∼1.000
W. Usamb. 1	Emmao	0.001 (0.000, >5)	>10 (0.000, >10)	9	-7.49	0.657 (0.000, >10)	>10 (0.000, >15)	>10 (0.000, 7.545)	15	-6.38	0.136
	Handei	0.082 (0.056, >5)	0.093 (0.048, 0.175)	40	-72.43	0.758 (0.076, >10)	4.405 (0.082, >10)	1.186 (0.080, 0.620)	6	-70.21	0.035
	Tewe	0.065 (0.046, >5)	0.031 (0.008, 0.067)	35	-65.11	0.090 (0.058, 0.139)	3.493 (0.196, >10)	0.044 (0.000, >10)	2	-63.06	0.043
	Mn'galo	0.375 (0.227, >5)	0.199 (0.100, 0.422)	40	-65.67	0.566 (0.291, >10)	0.447 (0.171, >10)	0.460 (0.000, >10)	12	-63.58	0.041
W. Usamb. 2	Kwadoe	0.023 (0.012, >5)	0.130 (0.000, >10)	21	-51.64	0.044 (0.016, 0.459)	0.301 (0.033, 4.32)	0.079 (0.000, >10)	20	-50.28	0.099
	Funta	0.142 (0.108, >5)	0.020 (0.008, 0.041)	40	-47.8	0.294 (0.119, >10)	0.425 (0.011, >10)	0.049 (0.000, >10)	6	-45.90	0.051
	Tamota	0.159 (0.112, >5)	0.041 (0.016, 0.096)	39	-62.93	0.268 (0.166, 0.535)	0.169 (0.069, 0.449)	0.048 (0.000, >10)	17	-57.48	0.001
	Mgila	0.493 (0.348, 0.789)	0.055 (0.025, 0.126)	37	-40.73	0.493 (0.348, 0.789)	0.055 (0.025, 0.126)	0.000 (0.000, 0.000)	37	-40.73	∼1.000
W. Usamb. 3	Mgome	0.706 (0.476, 1.151)	0.034 (0.009, 0.093)	33	-21.19	0.706 (0.476, 1.151)	0.034 (0.009, 0.093)	0.000 (0.000, >10)	33	-21.19	∼1.000

Table 4: Comparison between the two age-dependent SRR models $\mathrm{M}_{1,2}$ and $\mathrm{M}_{1,1}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-AMA1 antigen in samples from the 21 villages. Both models assume constant SCR (λ) in the population. Model $\mathrm{M}_{1,1}$ assumes constant SRR (ρ_1) for ages below τ, and a lower rate (ρ_2) after cut-off. Model $\mathrm{M}_{1,2}$ assumes $\rho_2 = 0$ after τ. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model $\mathrm{M}_{1,2}$ with $\mathrm{M}_{1,1}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits.

Transect	Village	Model $\mathrm{M}_{1,2}$				Model $\mathrm{M}_{1,1}$					p-value
		$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\tau}$	logL	$\hat{\lambda}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\tau}$	logL	
Rombo	Mokala	0.023 (0.007, >5)	0.291 (0.046, 1.477)	38	-34.36	3.531 (0.008, >10)	>10 (0.060, >10)	>10 (0.000, >10)	1	-33.78	0.281
	Machame Aleni	0.007 (0.003, >5)	>10 (0.257, >10)	14	-18.01	>10 (0.007, >10)	>10 (1.619, >10)	>10 (0.000, >10)	20	-17.45	0.290
	Ikuini	0.007 (0.004, >5)	0.017 (0.000, >10)	39	-35.10	2.640 (0.004, >10)	>10 (0.000, >10)	>10 (0.000, 0.308)	16	-33.67	0.091
	Kileo	0.025 (0.016, >5)	0.013 (0.000, >10)	40	-47.97	0.042 (0.017, 0.117)	0.421 (0.000, 2.476)	0.045 (0.000, >10)	6	-47.44	0.303
N. Pare	Kilomeni	0.159 (0.041, >5)	0.612 (0.116, 1.743)	35	-33.31	10.779 (0.118, >10)	>10 (0.416, >10)	7.642 (0.000, >10)	36	-31.79	0.081
	Lambo	0.104 (0.062, >5)	0.042 (0.006, 0.168)	39	-42.43	0.342 (0.082, >10)	0.591 (0.022, >10)	0.182 (0.000, >10)	10	-41.05	0.097
	Ngulu	1.021 (0.282, >5)	0.557 (0.000, >10)	6	-5.73	1.020 (0.283, >10)	0.557 (0.000, >10)	0.000 (0.000, 0.031)	6	-5.73	~1.000
	Kambi ya Simba	0.440 (0.196, >5)	0.168 (0.040, 0.518)	40	-30.55	0.633 (0.251, >10)	0.317 (0.072, >10)	0.217 (0.000, >10)	18	-29.23	0.104
S. Pare	Bwambo	0.041 (0.024, >5)	0.082 (0.000, >10)	18	-58.93	0.076 (0.026, 0.361)	0.220 (0.000, 1.528)	0.033 (0.000, >10)	18	-58.18	0.221
	Mpinji	0.070 (0.048, >5)	0.028 (0.003, 0.084)	40	-48.89	0.170 (0.069, 0.428)	0.374 (0.040, 1.321)	0.081 (0.000, >10)	11	-46.50	0.029
	Goha	0.128 (0.090, >5)	0.071 (0.031, 0.139)	28	-58.91	0.134 (0.092, 0.214)	0.075 (0.032, 0.160)	0.008 (0.000, >10)	28	-58.65	0.471
	Kadando	0.235 (0.166, >5)	0.089 (0.051, 0.159)	40	-65.47	0.562 (0.230, >10)	1.140 (0.089, >10)	0.267 (0.000, >10)	3	-61.77	0.007
W. Usamb. 1	Emmao	0.021 (0.009, >5)	0.095 (0.004, 0.604)	40	-31.37	0.133 (0.013, >10)	2.223 (0.038, >10)	0.800 (0.000, >10)	10	-29.78	0.075
	Handei	0.260 (0.188, >5)	0.072 (0.039, 0.132)	37	-62.64	0.260 (0.188, 0.382)	0.072 (0.039, 0.132)	0.000 (0.000, >10)	37	-62.64	~1.000
	Tewe	0.154 (0.116, >5)	0.037 (0.018, 0.069)	40	-56.94	0.493 (0.222, >10)	0.753 (0.192, >10)	0.138 (0.000, >10)	8	-52.38	0.003
	Mn'galo	0.623 (0.475, >5)	0.028 (0.013, 0.054)	40	-33.58	1.650 (0.602, >10)	0.868 (0.029, >10)	0.074 (0.044, 0.250)	3	-31.35	0.035
W. Usamb. 2	Kwadoe	0.052 (0.031, >5)	0.129 (0.059, 0.298)	40	-67.91	0.341 (0.103, >10)	2.844 (0.651, >10)	1.043 (0.000, >10)	7	-62.46	0.001
	Funta	0.130 (0.090, >5)	0.066 (0.034, 0.129)	40	-53.90	0.355 (0.118, >10)	1.456 (0.054, >10)	0.215 (0.052, 0.190)	4	-52.63	0.111
	Tamota	0.116 (0.066, >5)	0.196 (0.091, 0.438)	40	-65.96	0.245 (0.090, >10)	0.648 (0.139, >10)	0.885 (0.000, 1.370)	10	-63.38	0.023
	Mgila	0.625 (0.384, >5)	0.249 (0.132, 0.468)	40	-60.95	1.301 (0.572, >10)	0.695 (0.240, >10)	1.246 (0.000, >10)	8	-55.81	0.001
W. Usamb. 3	Mgome	0.942 (0.606, 1.721)	0.067 (0.026, 0.173)	35	-26.70	15.64 (0.809, >10)	7.037 (0.044, >10)	1.072 (0.000, 0.536)	2	-26.35	0.403

D $\mathbf{M}_0$ *vs.* $\mathbf{M}_{1,2}$

Table 5: Comparison between models M_0 and $M_{1,2}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-MSP2 antigen in samples from the 21 villages. Model M_0 assumes a constant SCR and SRR (λ and ρ, respectively), while model $M_{1,2}$ assumes a constant SCR, λ_1 for ages $< \tau$ and $\lambda_2 = 0$ otherwise. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model M_0 with $M_{1,2}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits. This event can be mostly seen at high altitude villages.

Transect	Village	Model M_0			Model $M_{1,2}$				p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}$ (95% CI)	$\hat{\rho}_1$ (95% CI)	$\hat{\tau}$	logL	
Rombo	Mokala	0.002 (0.001, 0.016)	0.000 (0.000, 0.432)	-20.54	0.004 (0.001, >5)	0.309 (0.000, >10)	16	-19.95	0.277
	Machame Aleni	0.011 (0.001, 0.176)	0.273 (0.000, >10)	-16.9	0.053 (0.002, >5)	1.830 (0.014, 15.071)	40	-16.16	0.224
	Ikuini	0.001 (0.000, 0.080)	0.000 (0.000, >10)	-14.57	0.009 (0.001, >5)	0.857 (0.000, >10)	31	-13.68	0.182
	Kileo	0.009 (0.006, 0.021)	0.000 (0.000, 0.079)	-42.61	0.014 (0.007, >5)	0.408 (0.000, >10)	9	-41.64	0.164
N. Pare	Kilomeni	0.007 (0.002, 0.418)	0.026 (0.000, >10)	-18.27	0.008 (0.003, >5)	0.041 (0.000, >10)	39	-18.19	0.689
	Lambo	0.017 (0.011, 0.034)	0.000 (0.000, 0.059)	-35.59	0.029 (0.015, >5)	0.707 (0.000, >10)	9	-33.77	0.056
	Ngulu	0.291 (0.165, 0.513)	0.000 (0.000, 0.030)	-5.84	0.536 (0.190, >5)	0.313 (0.000, >10)	6	-5.33	0.313
	Kambi ya Simba	0.985 (0.202, 10.491)	0.408 (0.041, 1.886)	-30.44	0.523 (0.214, >5)	0.219 (0.051, 0.720)	40	-30.68	~1.000
S. Pare	Bwambo	0.049 (0.039, 0.073)	0.000 (0.000, 0.023)	-58.59	0.057 (0.044, >5)	>10 (0.000, >10)	2	-57.15	0.090
	Mpinji	0.202 (0.084, 4.197)	0.116 (0.016, 2.792)	-52.75	0.475 (0.172, >5)	0.410 (0.103, 1.295)	34	-47.04	0.001
	Goha	0.281 (0.123, 2.166)	0.407 (0.150, 5.514)	-73.67	0.205 (0.120, >5)	0.300 (0.152, 0.632)	40	-74.54	~1.000
	Kadando	0.142 (0.081, 0.312)	0.130 (0.056, 0.347)	-61.58	0.148 (0.089, >5)	0.143 (0.068, 0.332)	40	-60.86	0.230
W. Usamb. 1	Emmao	0.001 (0.000, 0.099)	0.000 (0.000, >10)	-7.69	0.001 (0.000, >5)	>10 (0.000, >10)	9	-7.49	0.527
	Handei	0.083 (0.054, 0.142)	0.094 (0.044, 0.198)	-72.04	0.082 (0.056, >5)	0.093 (0.048, 0.175)	40	-72.43	~1.000
	Tewe	0.062 (0.043, 0.092)	0.024 (0.003, 0.059)	-65.43	0.065 (0.046, >5)	0.031 (0.008, 0.067)	35	-65.11	0.424
	Mn'galo	0.370 (0.208, 0.828)	0.190 (0.086, 0.497)	-66.52	0.375 (0.227, >5)	0.199 (0.100, 0.422)	40	-65.67	0.192
W. Usamb. 2	Kwadoe	0.018 (0.010, 0.040)	0.028 (0.000, 0.121)	-52.51	0.023 (0.012, >5)	0.130 (0.000, >10)	21	-51.64	0.187
	Funta	0.092 (0.072, 0.339)	0.000 (0.000, 0.123)	-56.8	0.142 (0.108, >5)	0.020 (0.008, 0.041)	40	-47.8	<0.001
	Tamota	0.150 (0.104, 0.246)	0.034 (0.011, 0.087)	-63.54	0.159 (0.112, >5)	0.041 (0.016, 0.096)	39	-62.93	0.269
	Mgila	0.465 (0.322, 0.760)	0.046 (0.019, 0.110)	-42.36	0.493 (0.348, 0.789)	0.055 (0.025, 0.126)	37	-40.73	0.071
W. Usamb. 3	Mgome	0.672 (0.440, 1.161)	0.025 (0.006, 0.080)	-22.33	0.706 (0.476, 1.151)	0.034 (0.009, 0.093)	33	-21.19	0.131

Table 6: Comparison between models M_0 and $M_{1,2}$ using the likelihood ratio test. Data used from the immune responses to *P. falciparum*-AMA1 antigen in samples from the 21 villages. Model M_0 assumes a constant SCR and SRR (λ and ρ, respectively), while model $M_{1,2}$ assumes a constant SCR, λ_1 for ages $< \tau$ and $\lambda_2 = 0$ otherwise. LogL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using profile likelihood method. P-value is associated with the log-likelihood ratio test comparing the nested model M_0 with $M_{1,2}$. Estimated 95% confidence intervals including >10 suggest the model did not have sufficient information to accurately estimate the lower and upper limits. This event can be mostly seen at high altitude villages.

Transect	Village	Model M_0			Model $M_{1,2}$				p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}$ (95% CI)	$\hat{\rho}_1$ (95% CI)	$\hat{\tau}$	logL	
Rombo	Mokala	0.028 (0.006, 0.297)	0.350 (0.034, >10)	-34.14	0.023 (0.007, >5)	0.291 (0.046, 1.477)	38	-34.36	~1.000
	Machame Aleni	0.003 (0.001, 0.007)	0.000 (0.000, 0.080)	-20.88	0.007 (0.003, >5)	>10 (0.257, >10)	14	-18.01	0.017
	Ikuini	0.006 (0.003, 0.019)	0.000 (0.000, 0.146)	-35.20	0.007 (0.004, >5)	0.017 (0.000, >10)	39	-35.10	0.655
	Kileo	0.025 (0.015, 0.048)	0.015 (0.000, 0.076)	-47.87	0.025 (0.016, >5)	0.013 (0.000, >10)	40	-47.97	~1.000
N. Pare	Kilomeni	0.617 (0.014, 1.401)	1.971 (0.000, 15.101)	-34.79	0.159 (0.041, >5)	0.612 (0.116, 1.743)	35	-33.31	0.085
	Lambo	0.098 (0.056, 0.232)	0.036 (0.002, 0.156)	-42.74	0.104 (0.062, >5)	0.042 (0.006, 0.168)	39	-42.43	0.431
	Ngulu	0.344 (0.192, 0.622)	0.000 (0.000, 0.043)	-7.36	1.021 (0.282, >5)	0.557 (0.000, >10)	6	-5.73	0.071
	Kambi ya Simba	0.503 (0.179, 8.564)	0.182 (0.030, 1.582)	-30.50	0.440 (0.196, >5)	0.168 (0.040, 0.518)	40	-30.55	~1.000
S. Pare	Bwambo	0.029 (0.021, 0.048)	0.003 (0.000, 0.038)	-59.84	0.041 (0.024, >5)	0.082 (0.000, >10)	18	-58.93	0.177
	Mpinji	0.071 (0.046, 0.125)	0.028 (0.001, 0.095)	-48.70	0.070 (0.048, >5)	0.028 (0.003, 0.084)	40	-48.89	~1.000
	Goha	0.108 (0.074, 0.171)	0.037 (0.010, 0.090)	-62.31	0.128 (0.090, >5)	0.071 (0.031, 0.139)	28	-58.91	0.009
	Kadando	0.253 (0.166, 0.431)	0.095 (0.049, 0.199)	-63.66	0.235 (0.166, >5)	0.089 (0.051, 0.159)	40	-65.47	~1.000
W. Usamb. 1	Emmao	0.023 (0.008, 0.587)	0.108 (0.000, 24.434)	-31.07	0.021 (0.009, >5)	0.095 (0.004, 0.604)	40	-31.37	~1.000
	Handei	0.252 (0.176, 0.390)	0.066 (0.032, 0.129)	-63.84	0.260 (0.188, >5)	0.072 (0.039, 0.132)	37	-62.64	0.121
	Tewe	0.152 (0.112, 0.221)	0.036 (0.016, 0.072)	-57.00	0.154 (0.116, >5)	0.037 (0.018, 0.069)	40	-56.94	0.729
	Mn'galo	0.628 (0.460, 0.874)	0.027 (0.011, 0.058)	-33.29	0.623 (0.475, >5)	0.028 (0.013, 0.054)	40	-33.58	~1.000
W. Usamb. 2	Kwadoe	0.055 (0.030, 0.144)	0.137 (0.055, 0.470)	-67.18	0.052 (0.031, >5)	0.129 (0.059, 0.298)	40	-67.91	~1.000
	Funta	0.126 (0.084, 0.200)	0.061 (0.028, 0.128)	-54.48	0.130 (0.090, >5)	0.066 (0.034, 0.129)	40	-53.90	0.281
	Tamota	0.141 (0.067, 0.584)	0.241 (0.091, 1.184)	-64.56	0.116 (0.066, >5)	0.196 (0.091, 0.438)	40	-65.96	~1.000
	Mgila	1.051 (0.466, 7.859)	0.428 (0.163, 1.472)	-57.03	0.625 (0.384, >5)	0.249 (0.132, 0.468)	40	-60.95	~1.000
W. Usamb. 3	Mgome	0.908 (0.556, 1.751)	0.055 (0.019, 0.157)	-28.15	0.942 (0.606, 1.721)	0.067 (0.026, 0.173)	35	-26.70	0.089

E $\mathbf{M}_0$ *vs.* $\mathbf{M}_2$

Table 7: Comparative analysis of the results from models M_0 and M_2, referring all 21 villages, considering MSP2 individual status as the outcomes. Model M_0 assumes constant SCR (λ) and SRR (ρ) for all ages. Model M_2 also assumes constant SRR, and a change in SCR after the cutoff parameter, τ^*. logL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using the profile likelihood method. p-value is associated with the log-likelihood ratio test comparing both models.

Transect	Village	Model M_0			Model M_2					p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}_1$ (95% CI)	$\hat{\lambda}_2$ (95% CI)	$\hat{\rho}$ (95% CI)	$\hat{\tau}^*$	logL	
Rombo	Mokala	0.002 (0.001, 0.016)	0.000 (0.000, 0.432)	-20.54	12.901 (0.000, >10)	0.003 (0.001, 0.007)	0.062 (0.000, >10)	36	-19.43	0.330
	Machame Aleni	0.011 (0.001, 0.176)	0.273 (0.000, >10)	-16.9	140.477 (0.000, >10)	0.010 (0.000, >5)	0.598 (0.000, >5)	6	-16.49	0.664
	Ikuini	0.001 (0.000, 0.080)	0.000 (0.000, >10)	-14.57	9.110 (0.000, >10)	0.002 (0.000, 0.004)	0.049 (0.000, >10)	39	-13.13	0.237
	Kileo	0.009 (0.006, 0.021)	0.000 (0.000, 0.079)	-42.61	1.955 (0.004, >5)	0.010 (0.004, 0.020)	0.073 (0.000, >10)	20	-41.27	0.262
N. Pare	Kilomeni	0.007 (0.002, 0.418)	0.026 (0.000, >10)	-18.27	70.665 (0.001, >10)	0.000 (0.000, 0.025)	0.189 (0.000, >10)	11	-16.63	0.194
	Lambo	0.017 (0.011, 0.034)	0.000 (0.000, 0.059)	-35.59	0.288 (0.032, >5)	0.009 (0.002, 0.024)	0.054 (0.005, >10)	13	-32.66	0.053
	Ngulu	0.291 (0.165, 0.513)	0.000 (0.000, 0.030)	-5.84	117.965 (0.068, >10)	0.251 (0.124, 0.451)	0.000 (0.000, >10)	7	-5.34	0.607
	Kambi ya Simba	0.985 (0.202, 10.491)	0.408 (0.041, 1.886)	-30.44	0.049 (0.000, >5)	0.912 (0.102, >5)	0.043 (0.000, >5)	1	-29.68	0.468
S. Pare	Bwambo	0.049 (0.039, 0.073)	0.000 (0.000, 0.023)	-58.59	19.450 (0.177, >10)	0.045 (0.036, 0.057)	0.000 (0.000, >10)	37	-55.25	0.035
	Mpinji	0.202 (0.084, 4.197)	0.116 (0.016, 2.792)	-52.75	0.033 (0.017, 0.066)	0.431 (0.258, 0.643)	0.000 (0.000, 0.038)	1	-45.25	0.001
	Goha	0.281 (0.123, 2.166)	0.407 (0.150, 5.514)	-73.67	0.055 (0.000, >5)	0.350 (0.172, >5)	0.165 (0.000, >5)	1	-72.32	0.259
	Kadando	0.142 (0.081, 0.312)	0.130 (0.056, 0.347)	-61.58	140.776 (0.000, >10)	0.144 (0.086, 0.261)	0.185 (0.000, >10)	6	-60.85	0.482
W. Usamb. 1	Emmao	0.001 (0.000, 0.099)	0.000 (0.000, >10)	-7.69	50.467 (0.000, >10)	0.000 (0.000, 0.014)	0.231 (0.000, >10)	15	-6.43	0.284
	Handei	0.083 (0.054, 0.142)	0.094 (0.044, 0.198)	-72.04	1.547 (0.080, >5)	0.072 (0.041, 0.125)	0.166 (0.083, >10)	6	-70.14	0.150
	Tewe	0.062 (0.043, 0.092)	0.024 (0.003, 0.059)	-65.43	0.418 (0.090, >5)	0.046 (0.027, 0.075)	0.059 (0.028, >10)	8	-62.79	0.071
	Mn'galo	0.370 (0.208, 0.828)	0.190 (0.086, 0.497)	-66.52	0.052 (0.006, 0.360)	0.367 (0.256, 0.519)	0.052 (0.000, 0.222)	2	-64.51	0.134
W. Usamb. 2	Kwadoe	0.018 (0.010, 0.040)	0.028 (0.000, 0.121)	-52.51	27.229 (0.002, >10)	0.019 (0.012, 0.032)	0.056 (0.009, >10)	27	-50.88	0.196
	Funta	0.092 (0.072, 0.339)	0.000 (0.000, 0.123)	-56.80	0.287 (0.081, >5)	0.128 (0.085, 0.182)	0.026 (0.009, >10)	6	-46.19	<0.001
	Tamota	0.150 (0.104, 0.246)	0.034 (0.011, 0.087)	-63.54	0.076 (0.040, 0.135)	0.347 (0.187, 0.544)	0.013 (0.000, 0.041)	1	-59.86	0.025
	Mgila	0.465 (0.322, 0.760)	0.046 (0.019, 0.110)	-42.36	0.318 (0.148, 0.657)	0.662 (0.336, 1.097)	0.039 (0.017, 0.086)	1	-41.56	0.449
W. Usamb. 3	Mgome	0.672 (0.440, 1.161)	0.025 (0.006, 0.080)	-22.33	0.303 (0.003, >5)	0.719 (0.477, 1.070)	0.020 (0.000, 0.065)	3	-21.94	0.677

Table 8: Comparative analysis of the results from models M_0 and M_2, referring all 21 villages, considering AMA1 individual status as the outcomes. Model M_0 assumes constant SCR (λ) and SRR (ρ) for all ages. Model M_2 also assumes constant SRR, and a change in SCR after the cutoff parameter, τ^*. logL refers to the log-likelihood function evaluated at the respective maximum likelihood estimates using the profile likelihood method. p-value is associated with the log-likelihood ratio test comparing both models.

Transect	Village	Model M_0			Model M_2					p-value
		$\hat{\lambda}$ (95% CI)	$\hat{\rho}$ (95% CI)	logL	$\hat{\lambda}_1$ (95% CI)	$\hat{\lambda}_2$ (95% CI)	$\hat{\rho}$ (95% CI)	$\hat{\tau}^*$	logL	
Rombo	Mokala	0.028 (0.006, 0.297)	0.350 (0.034, >10)	-34.14	8.711 (0.000, >5)	0.000 (0.000, >5)	2.437 (0.000, >5)	1	-33.78	0.698
	Machame Aleni	0.003 (0.001, 0.007)	0.000 (0.000, 0.080)	-20.88	35.913 (0.006, >5)	0.000 (0.000, 0.003)	0.105 (0.000, >10)	20	-17.45	0.032
	Ikuini	0.006 (0.003, 0.019)	0.000 (0.000, 0.146)	-35.20	47.109 (0.000, >5)	0.007 (0.002, 0.016)	0.121 (0.000, >10)	16	-33.33	0.154
	Kileo	0.025 (0.015, 0.048)	0.015 (0.000, 0.076)	-47.87	29.052 (0.000, >5)	0.028 (0.017, 0.045)	0.055 (0.000, >10)	19	-46.72	0.317
N. Pare	Kilomeni	0.617 (0.014, 1.401)	1.971 (0.000, 15.101)	-34.79	0.000 (0.000, >5)	0.403 (0.018, >5)	1.212 (0.000, >5)	2	-34.77	0.980
	Lambo	0.098 (0.056, 0.232)	0.036 (0.002, 0.156)	-42.74	0.047 (0.016, 0.144)	0.177 (0.071, 0.322)	0.012 (0.000, 0.085)	2	-41.52	0.295
	Ngulu	0.344 (0.192, 0.622)	0.000 (0.000, 0.043)	-7.36	0.237 (0.072, 0.596)	0.468 (0.189, 0.908)	0.000 (0.000, 0.030)	2	-6.95	0.664
	Kambi ya Simba	0.503 (0.179, 8.564)	0.182 (0.030, 1.582)	-30.50	1.627 (0.008, >5)	0.000 (0.000, 1.263)	0.204 (0.000, 0.683)	1	-29.41	0.336
S. Pare	Bwambo	0.029 (0.021, 0.048)	0.003 (0.000, 0.038)	-59.84	27.276 (0.000, >5)	0.032 (0.022, 0.045)	0.024 (0.000, >10)	27	-58.11	0.177
	Mpinji	0.071 (0.046, 0.125)	0.028 (0.001, 0.095)	-48.70	47.168 (0.000, >5)	0.070 (0.048, 0.106)	0.040 (0.003, >10)	16	-47.33	0.254
	Goha	0.108 (0.074, 0.171)	0.037 (0.010, 0.090)	-62.31	0.036 (0.021, 0.110)	0.129 (0.092, 0.171)	0.000 (0.000, 0.046)	4	-60.30	0.134
	Kadando	0.253 (0.166, 0.431)	0.095 (0.049, 0.199)	-63.66	117.812 (0.000, >5)	0.231 (0.154, 0.365)	0.099 (0.000, >10)	7	-62.73	0.395
W. Usamb. 1	Emmao	0.023 (0.008, 0.587)	0.108 (0.000, 24.434)	-31.07	70.783 (0.000, >5)	0.020 (0.005, 0.100)	0.194 (0.000, >10)	11	-30.04	0.357
	Handei	0.252 (0.176, 0.390)	0.066 (0.032, 0.129)	-63.84	0.168 (0.072, 0.420)	0.290 (0.188, 0.420)	0.051 (0.020, 0.108)	2	-63.30	0.583
	Tewe	0.152 (0.112, 0.221)	0.036 (0.016, 0.072)	-57.00	1.770 (0.138, >5)	0.137 (0.098, 0.194)	0.042 (0.024, >10)	10	-55.19	0.164
	Mn'galo	0.628 (0.460, 0.874)	0.027 (0.011, 0.058)	-33.29	2.340 (0.470, >5)	0.556 (0.398, 0.772)	0.028 (0.014, 0.054)	3	-31.93	0.257
W. Usamb. 2	Kwadoe	0.055 (0.030, 0.144)	0.137 (0.055, 0.470)	-67.18	88.508 (0.000, >5)	0.046 (0.026, 0.083)	0.178 (0.127, >10)	9	-64.56	0.073
	Funta	0.126 (0.084, 0.200)	0.061 (0.028, 0.128)	-54.48	0.430 (0.112, >5)	0.090 (0.047, 0.159)	0.093 (0.046, 0.174)	4	-52.74	0.176
	Tamota	0.141 (0.067, 0.584)	0.241 (0.091, 1.184)	-64.56	0.050 (0.004, 0.561)	0.200 (0.074, 0.362)	0.116 (0.000, 0.672)	1	-63.91	0.522
	Mgila	1.051 (0.466, 7.859)	0.428 (0.163, 1.472)	-57.03	0.017 (0.000, >5)	0.614 (0.452, >5)	0.052 (0.000, >5)	2	-56.66	0.691
W. Usamb. 3	Mgome	0.908 (0.556, 1.751)	0.055 (0.019, 0.157)	-28.15	0.031 (0.000, >5)	0.808 (0.617, 1.302)	0.000 (0.000, 0.105)	3	-27.14	0.364

F MSP2 estimated seroprevalence

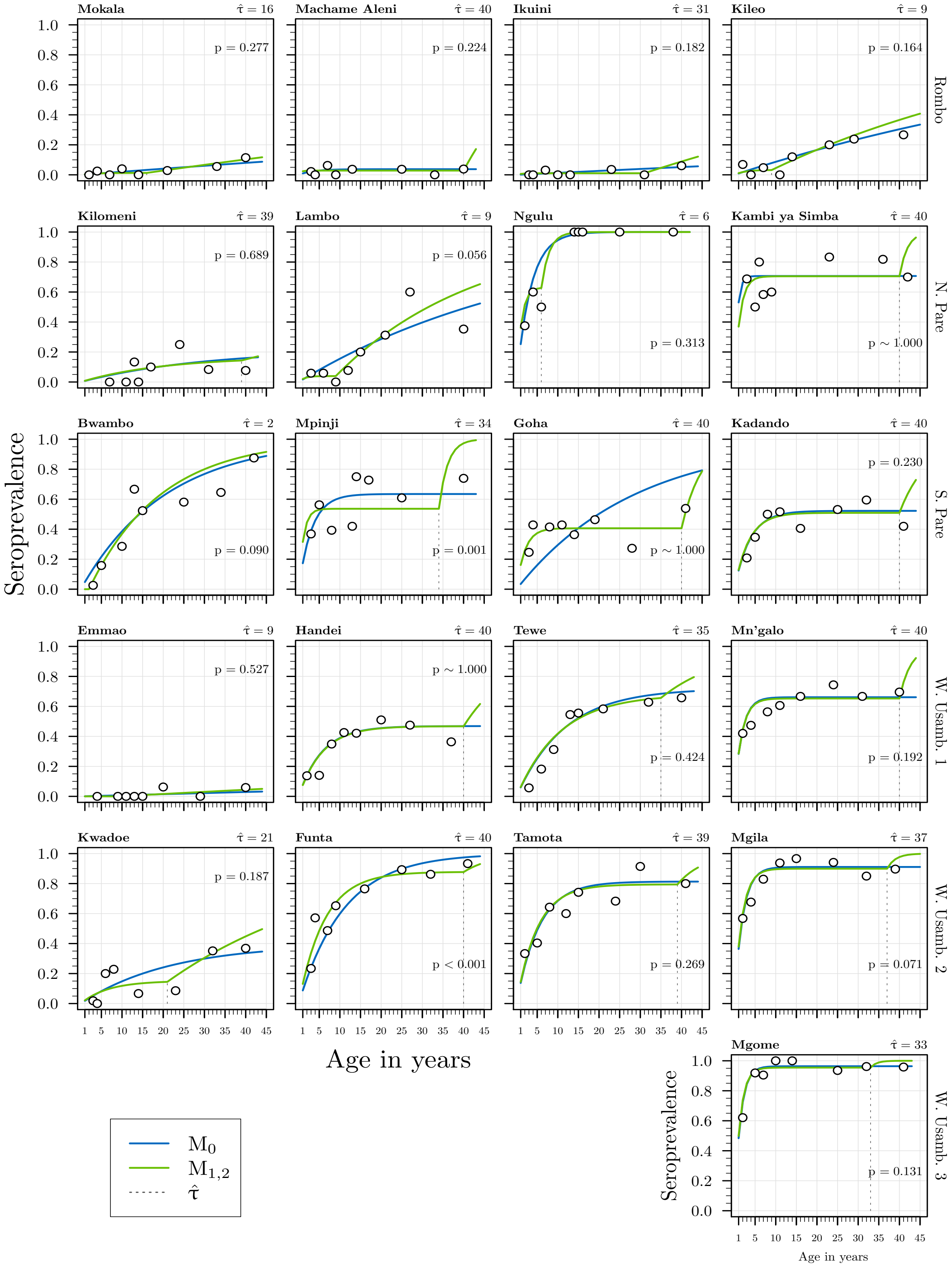

Figure 1: Fits for the estimated MSP2 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and $M_{1,2}$ (green lines), with the cutoff parameter of the latter signalled. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

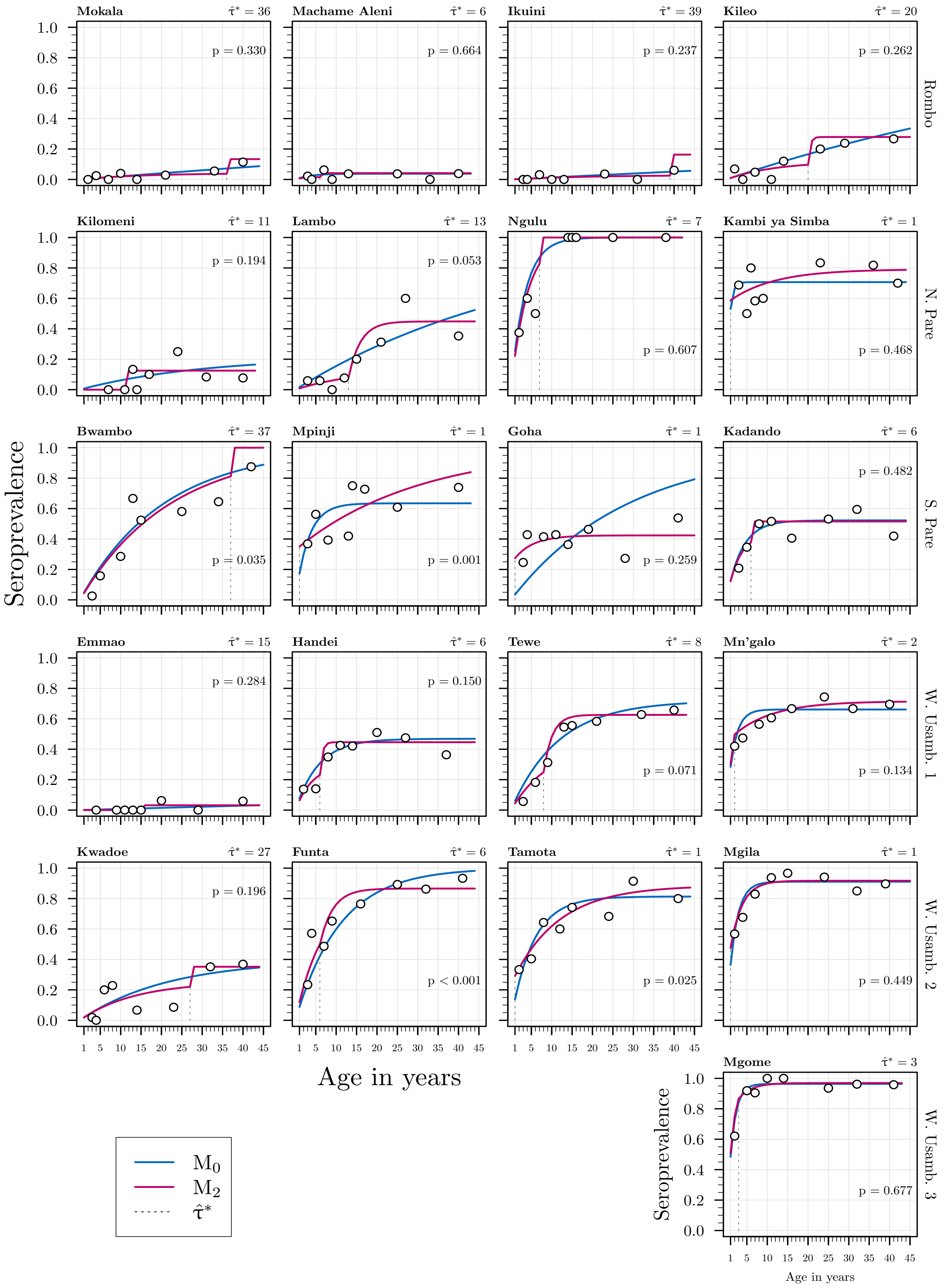

Figure 2: Fits for the estimated MSP2 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and M_2 (light red lines), with the cutoff parameter of the latter signalled, identifying the change in SCR happening in years before sampling. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

G AMA1 estimated seroprevalence

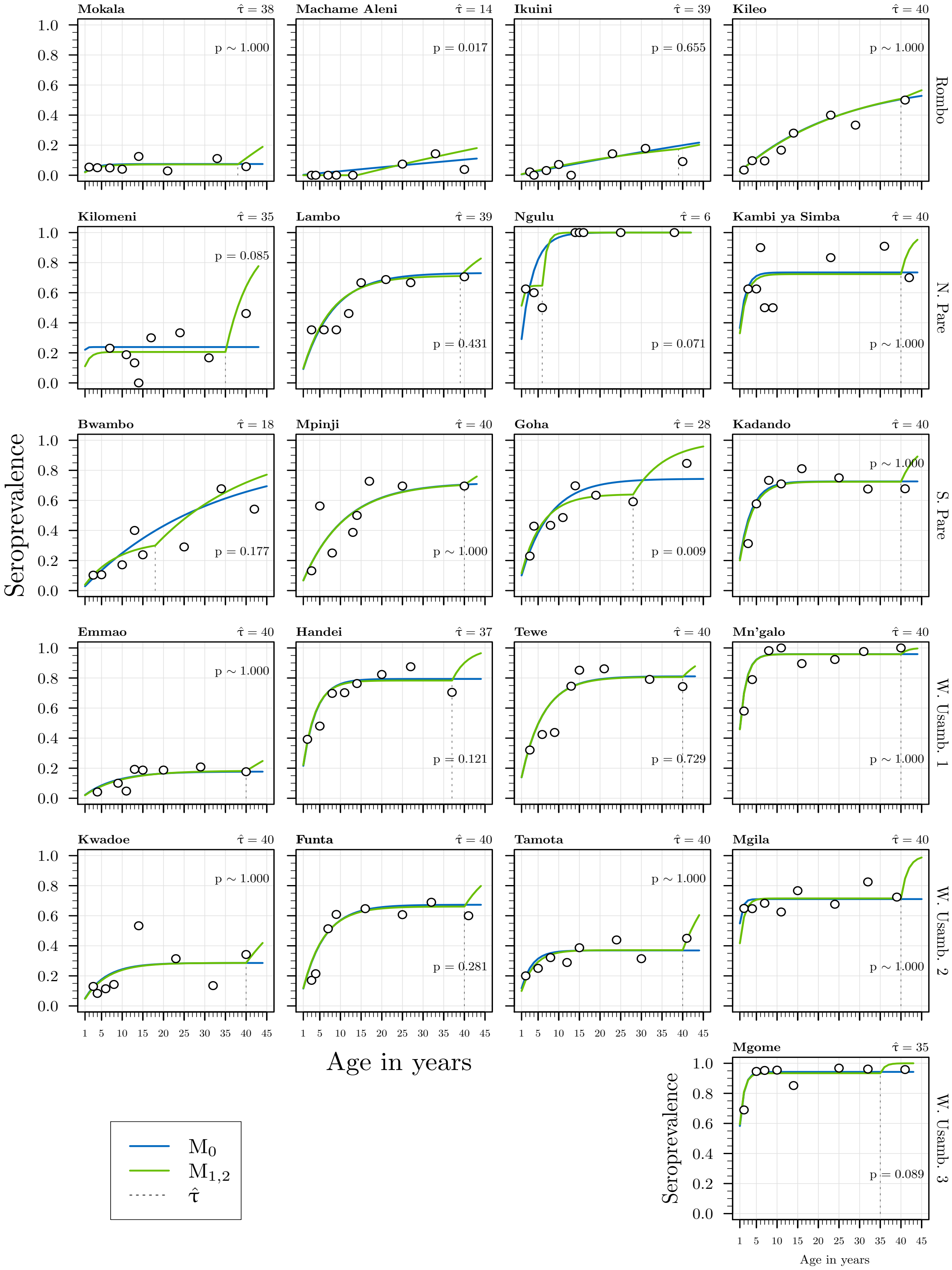

Figure 3: Fits for the estimated AMA1 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and $M_{1,2}$ (green lines), with the cutoff parameter of the latter signalled. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

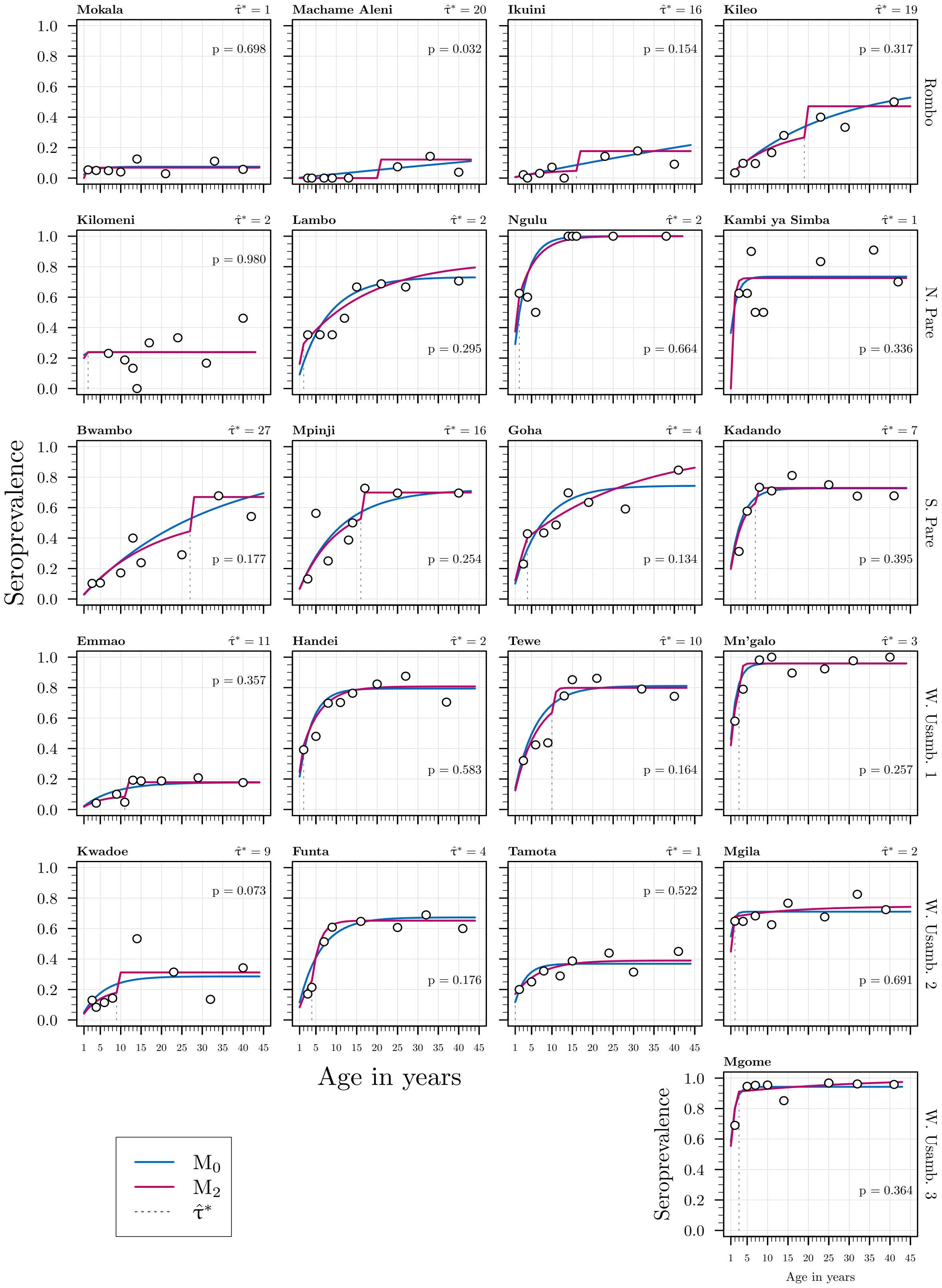

Figure 4: Fits for the estimated AMA1 antigen seroprevalence for the 21 assessed villages, using models M_0 (blue lines) and M_2 (light red lines), with the cutoff parameter of the latter signalled, identifying the change in SCR happening in years before sampling. Each row of graphs represents data from the transects (identified on the right hand side), where villages are ordered by decreasing altitude (and increasing malaria incidence). In the different plots, the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into similar bins. P-values from the resulting likelihood ratio tests are identified.

H R functions used to estimate transition rates from model $M_{1,1}$

```r
################################-
########## MODEL M11 ##########
################################-

# 1) RCM model for restriction 1 -----
rcm.M11.equation <- function(t, par, tau) {

  lambda <- exp(par[1])
  rho1 <- exp(par[2])
  p <- exp(par[3])/(1+exp(par[3]))
  rho2 <- rho1*p

  theta1 <- lambda/(lambda+rho1)
  theta2 <- lambda/(lambda+rho2)

  if(t<=tau){

    x1 <- (1-exp(-(lambda+rho1)*t))
    prob <- theta1*x1

  } else {

    x1 <- (1-exp(-(lambda+rho1)*tau))
    x2 <- (1-exp(-(lambda+rho2)*(t-tau)))
    prob <- theta2*x2 + theta1*x1*(1-x2)
  }

  return(prob)
}


# 2) Likelihood function -----
loglikelihood.M11 <- function(pos, age, par, tau) {

  # seropos/seroneg by age -----
  tabela <- table(age,pos)

  # group m|n|t -----
  tabela.binom <- cbind(tabela[,2], rowSums(tabela), as.numeric(rownames(tabela)))

  likelihood <- apply(tabela.binom, 1, function(x,par,tau) dbinom(x=x[1], size=x[2], prob=rcm.M11.equation(t=x[3],par,tau),log=F), par=par,tau=tau)

  return(sum(log(likelihood)))
}


# 3) Profile MLE for the parameters -----
mle.M11.estimates <- function(pos, age, time.int, n.start, par) {

  best.loglik <- (-1)*(10^6) # low loglikelihood initial value

  if(time.int[1] == time.int[2]) time.int <- time.int[1]

  output <- c()

  cat('\n1) Profile likelihood\n')

  for(tau in time.int) { # vary \tau -----

    # cat('tau=',tau,'\n',sep='')

    register <- c()

    for(i in 1:n.start) { # repeat n.start times each estimate of \tau -----

      # estimativas iniciais aleatorias -----
      par.ini <- par + runif(3,-0.5,0.5)

      # profile mle -----
      sol <- optim(par=par.ini, loglikelihood.M11, pos=pos, tau=tau, age=age, control=list(fnscale=(-1),maxit=1E+6))

      lambda <- exp(sol$par[1])
      rho1 <- exp(sol$par[2])
      p.rho2 <- exp(sol$par[3])
      rho2 <- rho1*p.rho2

      loglik.total <- sol$value

      register <- rbind(register, c(tau, loglik.total, lambda, rho1, rho2, sol$convergence))

      # select the best estimate for \tau -----
      aux <- which.max(register[,2]) # select the biggest loglikelihood value

      register <- register[aux,] # only that register
    }

    output <- rbind(output,register)

    if(register[2] > best.loglik) {

      output.vf <- register

      best.loglik <- register[2]
    }
  }

  colnames(output) <- c(’tau’, ’loglik’, ’lambda’, ’rho1’, ’rho2’, ’convergence’)

  return(output)
}


# 4) CI -----
rcm.M11.confidence.interval <- function(age, pos, par, tau, loglik.1){

  param.aux <- log(par)

  loglik1 <- loglikelihood.M11(age=age, pos=pos, par=param.aux, tau=tau)

  output <- c()

  for(j in 1:3) {
    if(j==1)cat(’\n2) Confidence interval for lambda’)
    if(j==2)cat(’\n3) Confidence interval for rho1’)
    if(j==3)cat(’\n4) Confidence interval for rho2’)


    param1 <- param.aux[j]
    param <- param.aux[-j]

    p <- qchisq(0.95, 1)

    loglik.0 <- loglik1 - p/2
```

```r

f2 <- function(param1, param, tau, age, pos, loglik.0, j){
    rcm.M11.estimates.one.par.fixed(age=age, pos=pos, param=param, param1=param1, tau=tau, j=j)-loglik.0
  }


##### CALCULATING LOWER BOUND #####

if(param1 < (-10)) {

  lower.bound <- exp(param1)

} else {

  sol <- rcm.M11.estimates.one.par.fixed(age=age, pos=pos, param=param, param1=-10, tau=tau, j=j)

  # cat('\nloglik(MLE) =',round(loglik1,2))
  # cat('\nCutoff(loglik) =',round(loglik.0,2))
  # cat('\nsol =',round(sol,2))

  if(sol > loglik.0){

    lower.bound <- exp(-100)

  } else {

    # print(param1)
    # print(param)
    # print(j)
    # print(rcm.M11.estimates.one.par.fixed(age=age, pos=pos, param=param, param1=param1, tau=tau, j=j))

    sol <- tryCatch(uniroot(f2,c(-10,param1), age=age, pos=pos, loglik.0=loglik.0, j=j, param=param, tau=tau),error=function(e){
      sol<-list(root=NA)
      return(sol)
    })

    lower.bound <- exp(sol$root)
  }
}


##### CALCULATING UPPER BOUND #####

sol <- rcm.M11.estimates.one.par.fixed(age=age, pos=pos, param=param, param1=20, tau=tau, j=j)

if(sol > loglik.0){

  upper.bound <- exp(100)

} else {

  sol <- tryCatch(uniroot(f2,c(param1,5), age=age, pos=pos, loglik.0=loglik.0, j=j, param=param, tau=tau), error=function(e){
    sol<-list(root=NA)
    return(sol)
  })

  upper.bound <- exp(sol$root)
}
```

```r

    cat('\nLower bound =',round(lower.bound,4))
    cat('\nUpper bound =',round(upper.bound,4),'\n')

    output <- rbind(output, c(lower.bound, upper.bound))
  }

  return(output)
}


# 5.1) Fix ONE parameter and estimate the remaining ones -----
rcm.M11.estimates.one.par.fixed <- function(age, pos, param, param1, tau, j) {

  if(j==1) fit <- optim(par=c(runif(1,-3,-1), runif(1,-3,-1)), fn = loglikelihood.M11.one.par.fixed, age=age, pos=pos, param1=param1, tau=tau, j=j, control=list(fnscale=-1,pgtol=1E-10))
  if(j==2) fit <- optim(par=c(runif(1,-3,-1), runif(1,-3,-1)), fn = loglikelihood.M11.one.par.fixed, age=age, pos=pos, param1=param1, tau=tau, j=j, control=list(fnscale=-1,pgtol=1E-10))
  if(j==3) fit <- optim(par=c(runif(1,-3,-1), log(1/runif(1,0,1))), fn = loglikelihood.M11.one.par.fixed, age=age, pos=pos, param1=param1, tau=tau, j=j, control=list(fnscale=-1,pgtol=1E-10))

  return(fit$value)
}

# 5.2) -----
loglikelihood.M11.one.par.fixed <- function(age, pos, param, param1, tau ,j) {

  if(j==1){
    all.param <- c(param1,param)
    sol <- loglikelihood.M11(pos=pos, age=age, par=all.param, tau=tau)
  }

  if(j==2){
    all.param <- c(param[1],param1,param[2])
    sol <- loglikelihood.M11(pos=pos, age=age, par=all.param, tau=tau)
  }

  if(j==3){
    all.param <- c(param,param1)
    sol <- loglikelihood.M11.Q(pos=pos, age=age, par=all.param, tau=tau)
  }

  return(sol)
}


# 5.3) -----
rcm.M11.equation.Q <- function(t, par, tau) {

  lambda <- exp(par[1])
  rho2 <- exp(par[3])
  q <- exp(par[2])
  rho1 <- rho2*q
  # p <- 1/q

  theta1 <- lambda/(lambda+rho1)
  theta2 <- lambda/(lambda+rho2)

  if(t<=tau){

    x1 <- (1-exp(-(lambda+rho1)*t))
    prob <- theta1*x1

  } else {

    x1 <- (1-exp(-(lambda+rho1)*tau))
    x2 <- (1-exp(-(lambda+rho2)*(t-tau)))
    prob <- theta1*x1+theta2*x2*(1-x1)
  }

  return(prob)
}

# 5.4) -----
loglikelihood.M11.Q <- function(pos, age, par, tau) {

  # seropos/seroneg by age -----
  tabela <- table(age,pos)

  # group m|n|t -----
  tabela.binom <- cbind(tabela[,2], rowSums(tabela), as.numeric(rownames(tabela)))

  likelihood <- apply(tabela.binom, 1, function(x,par,tau) dbinom(x=x[1], size=x[2], prob
    =rcm.M11.equation.Q(t=x[3],par,tau),log=F), par=par,tau=tau)

  return(sum(log(likelihood)))
}


# 6) FINAL FUNCTION -----
M11.analysis <- function(age, pos, time.int=c(1:40), n.start=1, par=log(runif(3,0,1))){

  # Get Profile likelihood -----
  mle <- mle.M11.estimates(pos=pos, age=age, time.int=time.int, n.start=n.start, par=par)

  # Get parameter estimates -----
  aux <- which.max(mle[,2])

  estimates <- mle[aux,]

  print(estimates)

  tau <- as.integer(estimates[1])
  loglikelihood.total <- estimates[[2]]
  lambda <- estimates[[3]]
  rho1 <- estimates[[4]]
  rho2 <- estimates[[5]]
  convergence <- as.integer(estimates[6])
  p.rho2 <- rho2/rho1

  ##### confidence intervals #####

  conf.int <- rcm.M11.confidence.interval(age=age ,pos=pos, par=c(lambda,rho1,p.rho2),
    tau=tau, loglik.1=loglikelihood.total)

  parameters <- cbind(c(lambda,rho1,rho2),conf.int)
  colnames(parameters) <- c(’estimates’, ’lower’, ’upper’)
  rownames(parameters) <- c(’lambda’, ’rho1’, ’rho2’)

  fitted.values <- c(1:max(age))

  exp.seroprev <- sapply(1:max(age),rcm.M11.equation, par=c(log(lambda),log(rho1), log(
    rho2)), tau=tau)

  fitted.values<-cbind(fitted.values,exp.seroprev)

  # print(conf.int)

  output <- list(loglikelihood.total=loglikelihood.total, estimates=parameters, tau=tau,
    df=4, expected.seroprevalence=fitted.values, proflik=mle, model='M1,1')

  return(output)
}
```

jtm_functions_M11.R