

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Importance, size and mobility of forest-going populations for malaria elimination in Lao People's Democratic Republic

**Permalink**

<https://escholarship.org/uc/item/32c803mv>

**Author**

Rerolle, Francois

**Publication Date**

2021

Peer reviewed|Thesis/dissertation

Importance, size and mobility of forest-going populations for malaria elimination in Lao People's Democratic Republic.


by  
Francois Rerolle

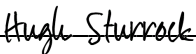
DISSERTATION  
Submitted in partial satisfaction of the requirements for degree of  
DOCTOR OF PHILOSOPHY


in  
Epidemiology and Translational Science

in the  
GRADUATE DIVISION  
of the  
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:  
  
0F4EF7533DF644A... Adam Bennett  
Chair

DocuSigned by:  
  
DocuSigned by:  
00000000000000000000000000000000... Hugh Sturrock

  
CAEE9AD7782C421... John Marshall

---

---

Committee Members



# Dedication and Acknowledgment

I would like to dedicate this dissertation work to everyone that has ever supported me. In particular, thanks to my babies for keeping me awake for so many nights and to my partner for sharing that joy with me. Special thanks to my parents, brothers and friends who have been there for me when I needed along the tortuous path to graduation.

I could not have accomplished this work without the incredible support from my dedicated dissertation committee members Adam Bennett, Hugh Sturrock and John Marshall. I also want to thank my co-authors for collaborating on my dissertation papers: Emily Dantzer, Andrew Lover, Jerry Jacobson, Paul Wesson, Jenny Smith, Toula Phimmakong, Bouasy Hongvanthong and Rattanaxay Phetsouvanh.

For their expertise and assistance, I also thank Michelle Roh, Ricardo Andrade Pacheco, Alemayehu Midekisa and Stephen Shiboski. Kudos to my PhD cohort classmates for sharing this journey and special thanks to Steven Asiimwe, my mentor, friend and confidant throughout the program. I also want to express my sincere gratitude to everyone at UCSF, staff and professors, who have created an amazing learning environment for me to learn and thrive. In particular, thanks to Maria Glymour for her inspiration, mentoring and genuine friendship and to Eva Wong-Moy for taking such a good care of me.

Last, I thank the Center for Malariology, Parasitology and Entomology (CMPE), the Lao Tropical Public Health Institute (TPHI, formerly NIOPH), the study teams and the study participants as well as the Bill & Melinda Gates Foundation for enabling this research.

# Contributions

Chapter 1 of this dissertation was published in eLife in 2021\*. The dissertation committee members supervised the research that forms the basis of this dissertation chapter and the published material is substantially the product of Francois Rerolle's period of study at UCSF and was primarily conducted and written by him. The work he completed for this published manuscript is comparable to a standard dissertation chapter.



Approved: \_\_\_\_\_ Adam Bennett, PhD, Dissertation Chair.

---

\*Rerolle, F. *et al.* Spatio-temporal associations between deforestation and malaria incidence in Lao PDR. *Elife* **10**, e56974 (2021).

# **Importance, size and mobility of forest-going populations for malaria elimination in Lao People's Democratic Republic**

François Rerolle

## **Abstract**

Malaria is a parasite infection transmitted by mosquitoes that infected 229 million cases and resulted in 409,000 deaths worldwide in 2019. In the Greater Mekong Sub-region (GMS), resistance to primary malaria treatments has emerged and is threatening to set back recent control successes. As countries ambition to eliminate malaria by 2030, infections clusters in forest-going populations that are increasingly targeted for prevention and treatment efforts by national control programs across the GMS. Yet, as pointed out by a recent review of the literature on forest-goers, a more detailed characterization of forest-going population is needed to accelerate malaria elimination in the GMS.

In chapter 1, we evaluated the association between deforestation and malaria incidence in northern and southern Lao PDR. Our approach leveraged surveillance records collected by the national program and high-resolution forest data to characterize the importance of forest-going population on malaria transmission in the GMS. Our results highlighted the challenges to transition from *Plasmodium falciparum* to *Plasmodium vivax* elimination and suggest programs may benefit from monitoring areas of on-going deforestation using remotely sensed data.

In chapter 2, three population-based surveys and one rolling survey among forest-goers from a randomized controlled trial in southern Lao PDR were combined to estimate the size of forest-

going populations. Population size estimates (PSEs) were produced at three different time points and the capture-recapture methodology was used to estimate the total number of forest-goers in the study area over the study period. This study highlighted an important seasonality in malaria risk behaviors among forest-goers and illustrates population size estimation methods that can be replicated to support national control programs in the GMS.

In chapter 3, GPS logging devices were leveraged to measure and describe fine-scale mobility patterns of forest-goers recruited in a focal test and treat (FTAT) active case detection intervention conducted in southern Lao PDR. Combining clustering analyses and machine learning regressions, our results assessed the diversity within forest-going trips but did not translate into a clear segmentation of forest-goers' role in malaria transmission in the GMS.

Taken together, this work characterizes the importance, size and mobility of forest-going populations in Lao PDR. These results are key for national control programs across the GMS to assess and meet their 2030 malaria elimination goals.

# Table of contents

<b>INTRODUCTION .....</b>	<b>1</b>
<b>CHAPTER 1: SPATIO-TEMPORAL ASSOCIATIONS BETWEEN DEFORESTATION AND MALARIA INCIDENCE IN LAO PDR. ....</b>	<b>4</b>
1.1. ABSTRACT .....	5
1.2. INTRODUCTION .....	6
1.3. RESULTS .....	10
1.4. DISCUSSION .....	24
1.5. MATERIALS AND METHODS .....	29
1.6. APPENDIX 1 .....	42
<b>CHAPTER 2: POPULATION SIZE ESTIMATION OF SEASONAL FOREST-GOING POPULATIONS IN SOUTHERN LAO PDR .....</b>	<b>72</b>
2.1. ABSTRACT .....	73
2.2. INTRODUCTION .....	74
2.3. METHODS .....	76
2.4. RESULTS .....	84
2.5. DISCUSSION .....	92
2.6. APPENDIX 2 .....	96
<b>CHAPTER 3: CHARACTERIZING MOBILITY PATTERNS OF FOREST GOERS IN SOUTHERN LAO PDR USING GPS LOGGERS .....</b>	<b>116</b>
3.1. ABSTRACT .....	117



3.2.	INTRODUCTION .....	119
3.3.	METHODS .....	121
3.4.	RESULTS .....	131
3.5.	DISCUSSION .....	143
3.6.	APPENDIX 3 .....	147
	<b>REFERENCES .....</b>	<b>151</b>

# List of figures

<b>Figure 1.1</b> - Average tree crown cover (%) in 2016 and percent area that experienced forest loss between 2011 and 2016 within a 10 km radius in northern and southern Lao PDR. ....	11
<b>Figure 1.2</b> - Malaria incidence and test positivity over time .....	14
<b>Figure 1.3</b> - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km of a village in the previous 1 to 5 years in Lao PDR.....	17
<b>Figure 1.4</b> - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km of a village in the previous 1 to 5 years in southern Lao PDR, differentiated by malaria species.....	20
<b>Figure 1.5</b> - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 30 km of a village in the previous 1 to 5 years and within areas with tree crown cover density above 0%, 68% and 87% in Lao PDR.. ....	23
<b>Figure 1.6</b> - Map of study's districts.....	30
<b>Figure 1.7</b> - Forest data methods.....	35
<b>Figure 1.8</b> - Conceptual model for our analysis.....	41
<b>Figure 1.9</b> - Treatment-seeking modeling plots. ....	46
<b>Figure 1.10</b> - Residual temporal autocorrelation when malaria incidence in previous 1 and 2 months are included or not.....	49
<b>Figure 1.11</b> - Relationships between malaria incidence and the environmental covariates in the multivariable model described in equation 1.2. No uncertainty bands.....	51
<b>Figure 1.12</b> - Relationships between malaria incidence and the temporal trend in the multivariable model described in equation 1.2.....	53

<b>Figure 1.13</b> - Adjusted relationship between deforestation and malaria incidence. ....	57
<b>Figure 1.14</b> - Distribution of average tree crown cover density within 1, 10 and 30 km of villages. ....	60
<b>Figure 1.15</b> - Distribution of percent area within 1, 10 and 30 km of villages that experienced forest loss between 2011 and 2016. ....	61
<b>Figure 1.16</b> - Distribution and time series of environmental covariates at study's villages. ....	62
<b>Figure 1.17</b> - Additional figures from malaria registries: malaria infections. ....	63
<b>Figure 1.18</b> - Distributions of socio-economical variables of all patients recorded in the malaria registries. ....	64
<b>Figure 1.19</b> - Additional figures from malaria registries: matched vs unmatched SES variables. ....	65
<b>Figure 1.20</b> - Distribution of travel time (in hours) from surveyed households to closest health facilities. ....	66
<b>Figure 1.21</b> - Relationships between malaria incidence and the environmental covariates in the multivariable model described in equation 1.2. ....	67
<b>Figure 1.22</b> - Raw scatterplot between monthly village malaria incidence rate and the percent area within 30 km of villages that experienced forest loss in the previous 1, 3 and 5 years. ....	68
<b>Figure 1.23</b> - Time series of deforestation, forest cover and malaria incidence, averaged over study's villages and for varying buffer radius around villages (1, 10 and 30 km). ....	69
<b>Figure 1.24</b> - Time series of deforestation and forest cover for a few randomly sampled study's villages. ....	70
<b>Figure 1.25</b> - Adjusted relationship between deforestation and species-specific malaria incidence in southern Lao PDR. ....	71

<b>Figure 2.1</b> - Study timeline and study area. ....	77
<b>Figure 2.2</b> - Demographics of FTAT HRP. ....	86
<b>Figure 2.3</b> - Seasonality of FTAT HRP. ....	87
<b>Figure 2.4</b> - Turnover of FTAT HRP. ....	88
<b>Figure 2.5</b> - Venn Diagram of the capture history data .....	90
<b>Figure 2.6</b> - Precipitation time series. ....	100
<b>Figure 2.7</b> - FTAT HRP enrollment. ....	101
<b>Figure 2.8</b> - Diagnostic test for heterogeneity. ....	113
<b>Figure 3.1</b> – Study timeline and study area. ....	122
<b>Figure 3.2</b> - Trajectories for GPS loggers collected during the first cycle in Moonlapamok district for PNs and forest-goers .....	127
<b>Figure 3.3</b> - Time series plot of when GPS loggers were on and collected GPS coordinates. ..	134
<b>Figure 3.4</b> - Plot of how the ICC for mobility patterns variables in Table 3.1 vary with the number of clusters selected. ....	136
<b>Figure 3.5</b> - Bi-plots of the clustering structure in the feature space. ....	137
<b>Figure 3.6</b> - SHAP importance plot. ....	141
<b>Figure 3.7</b> - SHAP dependence plot for the two main continuous predictors of high-risk trips. ....	142
<b>Figure 3.8</b> - Venn diagram for the raw probability of engaging in high-risk forest trips among the 8 strata of forest-goers defined by the three main predictors identified in the regression analysis .....	142
<b>Figure 3.9</b> - Dendrogram from the hierarchical clustering algorithm. ....	150

# List of tables

<b>Table 1.1</b> - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km of a village in the previous 1 to 5 years in northern and southern Lao PDR.....	16
<b>Table 1.2</b> - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km of a village in the previous 1 to 5 years in southern Lao PDR, differentiated by malaria species.....	19
<b>Table 1.3</b> - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 30 km of a village in the previous 1 to 5 years and within areas with tree crown cover density above 0%, 68% and 87% in Lao PDR.....	22
<b>Table 1.4</b> - Data used to parameterize the transition matrix with the travel speed between any 2 adjacent pixels of the map.....	43
<b>Table 1.5</b> - IRR associated with a 0.1% increase in forest loss. Sensitivity analysis: village population unadjusted for probability of seeking treatment. ....	47
<b>Table 1.6</b> - IRR [95% CI] associated with a 1% increase in average tree crown density.....	55
<b>Table 1.7</b> - AIC fit of univariate models when including each of the seven monthly climatic variation one at a time as unique covariate in equation 1.2.....	59
<b>Table 2.1</b> - HRP eligibility criteria.....	81
<b>Table 2.2</b> - Results for the population-based household survey method for population size estimation of HRP individuals.....	89
<b>Table 2.3</b> - Capture-recapture PSE results using log-linear models and assuming closed population. ....	91
<b>Table 2.4</b> - Identification of HRP individuals in baseline survey.....	96

<b>Table 2.5</b> - Identification of HRP individuals in MTAT survey.....	96
<b>Table 2.6</b> - Identification of HRP individuals in endline survey. ....	96
<b>Table 2.7</b> - Capture-recapture Mt PSE using 2, 3 or 4 of the survey lists available. ....	97
<b>Table 2.8</b> - Capture-recapture PSE for various models considered. ....	97
<b>Table 2.9</b> - Capture-recapture PSE for Mt models with additional interaction terms between surveys. ....	98
<b>Table 2.10</b> - Meta-analysis to estimate proportion of population older than 15. ....	99
<b>Table 2.11</b> - Results for the population-based survey method for population size estimation of HRP individuals. Sensitivity Analyses.....	103
<b>Table 2.12</b> - Parametrization for the expected number of individuals with certain capture histories for various three source capture-recapture models.....	112
<b>Table 2.13</b> - Results for the population-based survey method for population size estimation of agriculture-related HRP individuals. ....	115
<b>Table 2.14</b> - Results for the population-based survey method for population size estimation of forest-related HRP individuals. ....	115
<b>Table 3.1</b> - Mobility patterns variables computed for each of the outdoor trips.....	128
<b>Table 3.2</b> - Comparison between forest-goers that carried a GPS logger and those that did not in terms of their answers to FTAT variables. ....	132
<b>Table 3.3</b> - GPS logger self-reported utilization from retrieval questionnaire.....	135
<b>Table 3.4</b> - Distribution of input mobility patterns parameters for each cluster .....	139

# Introduction

## *Malaria situation in the GMS*

Malaria is a parasitic disease transmitted by *Anopheles* mosquitoes. Worldwide, there were 229 million malaria cases leading to 409,000 deaths in 2019 with most of the burden concentrated in Africa (94%)<sup>1</sup>. In the Greater Mekong Subregion (GMS) - Cambodia, Lao People's Democratic Republic (Lao PDR), Myanmar, Thailand, Vietnam and China's Yunnan Province – the reported number of malaria cases reached historical lows in 2019 with 239,000 cases, representing a 97% reduction from 2000<sup>1</sup>. Much of this progress has been attributed to heightened funding and better access to testing and treatments<sup>2</sup>.

In 2019, the GMS concentrates only 0.1% of all malaria cases but is known as the epicenter of antimalarial drug resistance and remains a key region for the world fight against malaria<sup>1,2</sup>.

Resistance to chloroquine, sulphadoxine-pyrimethane and mefloquine all emerged in the GMS<sup>3</sup>, spreading to Africa in the 70s and eventually causing catastrophic surges in malaria mortality and morbidity<sup>4</sup>. In 2006, *Plasmodium Falciparum* (*Pf*) malaria parasites resistant to artemisinin-based combination therapies (ACTs), the current most effective treatment, were discovered along the Thai-Cambodia border<sup>5,6</sup>, then emerging throughout the GMS<sup>7</sup>, notably in Southern Lao, the Thai-Myanmar border regions<sup>8</sup> and Cambodia<sup>9</sup>.

Elimination of *Pf* malaria is increasingly accepted as a necessary strategy to face the challenges of artemisinin resistance<sup>10,11</sup> and national governments across the GMS have adopted the

ambitious goal of elimination of *Pf* by 2025 and *Plasmodium Vivax (Pv)* and all human malaria species by 2030<sup>1,7,12</sup>.

### ***Forest malaria in the GMS***

The most efficient and widespread vectors of malaria parasites in the GMS, *Anopheles dirus* and *Anopheles minimus*, are forest mosquitoes<sup>13,14</sup>, and the malaria ecosystem in the GMS has been labelled as “forest malaria”<sup>15</sup>. *An. dirus* for instance needs shade and humidity for breeding, and therefore thrives in forested areas. In addition, *An. dirus* and *An. minimus* are highly anthropophilic, exophagic and early bitters, leading to residual outdoor transmission<sup>16</sup>. Forest malaria has been extensively described in different countries of southeast Asia where many studies have reported forest activity as a strong risk factor for malaria<sup>17–22</sup>.

As countries of the GMS work towards elimination, malaria clusters in forest-going populations that are increasingly targeted for prevention and treatment efforts by national control programs across the GMS<sup>23,24</sup>. Forest-goers are forest-fringe inhabitants who live in rural communities close to the forest and make frequent overnight trips to the forest to hunt or collect wood<sup>25</sup>. In addition to their increased exposure in the forest, forest-goers are at higher risk for malaria<sup>26</sup> because of poor adherence to protective measures against mosquitoes such as insecticide-treated bed nets (ITNs) or long-lasting insecticidal hammocks (LLIHs)<sup>24,27</sup> and delayed and inadequate access to treatment<sup>23</sup>.



## ***Objectives and significance of the dissertation***

The overall objectives of this dissertations are to provide a more detailed characterization of forest-going population which, as pointed out by a recent review of the literature on forest-goers<sup>28</sup>, is much needed to tailor interventions and accelerate malaria elimination in the GMS.

In chapter 1, we confirmed and characterized the importance of forest-going population on malaria transmission in the GMS by evaluating the association between deforestation and malaria incidence in northern and southern Lao PDR, two distinct environment with varying levels of endemicity and species composition. In chapter 2, we leveraged village and forest-based survey data from a randomized controlled trial in southern Lao PDR to estimate the size of forest-going populations. In chapter 3, we used GPS logging devices to measure and characterize fine-scale mobility patterns of forest-goers in southern Lao PDR.

In combination, this dissertation characterizes the importance, size and mobility of forest-going populations for malaria elimination in Lao PDR. These results are key for national control programs across the GMS to assess and meet their 2030 malaria elimination goals.

# **Chapter 1: Spatio-temporal associations between deforestation and malaria incidence in Lao PDR.**

François Rerolle, Emily Dantzer, Andrew A. Lover, John M. Marshall,  
Bouasy Hongvanthong, Hugh Sturrock, Adam Bennett

## 1.1. Abstract

As countries in the Greater Mekong Sub-region (GMS) increasingly focus their malaria control and elimination efforts on reducing forest-related transmission, greater understanding of the relationship between deforestation and malaria incidence will be essential for programs to assess and meet their 2030 elimination goals. Leveraging village-level health facility surveillance data and forest cover data in a spatio-temporal modeling framework, we found evidence that deforestation is associated with short-term increases, but long-term decreases in confirmed malaria case incidence in Lao People's Democratic Republic (Lao PDR). We identified strong associations with deforestation measured within 30 km of villages but not with deforestation in the near (10 km) and immediate (1 km) vicinity. Results appear driven by deforestation in densely forested areas and were more pronounced for infections with *Plasmodium falciparum* (*P. falciparum*) than for *Plasmodium vivax* (*P. vivax*). These findings highlight the influence of forest activities on malaria transmission in the GMS.

## 1.2. Introduction

Engaging in forest activities, such as logging, hunting or spending the night in the forest, is considered a primary risk factor for malaria infection in the Greater Mekong Sub-region (GMS)<sup>17–22</sup> with recent outbreaks attributed to deforestation activities<sup>29</sup>. This is most likely the result of increased human exposure to the forest dwelling *Anopheles dirus* and *Anopheles minimus*, the most efficient and widespread malaria vectors in the GMS<sup>13,14</sup>. However, deforestation may also alter this “forest malaria”<sup>15</sup> ecosystem and eventually reduce malaria incidence, as is generally accepted to be the case in Southeast Asia<sup>30</sup>. Several previous studies have assessed the relationship between deforestation and malaria, but the majority focused outside of the GMS, most notably in the Amazonian forest<sup>31–35</sup> where the evidence has been met with conflicting interpretations<sup>36</sup>. As national malaria programs across the GMS target forest-going populations for prevention and treatment efforts<sup>23,24</sup>, improved understanding of the relationship between deforestation and malaria is critical for programs to assess and meet national 2030 malaria elimination goals<sup>7,12</sup>.

In the Amazon, the “frontier malaria” hypothesis<sup>37</sup> posits that malaria temporarily increases following deforestation efforts to open a human settlement area in the forest. Subsequently, after approximately 6-8 years, settlements become more urbanized and isolated from the surrounding forest, and less suitable for malaria vectors, resulting in reduced malaria transmission<sup>38</sup>. Recent work has challenged this hypothesis, however, and found that some older settlements were also likely to have high malaria incidence<sup>39</sup>, highlighting the importance of assessing the relationship between deforestation and malaria at different spatio-temporal scales<sup>40</sup>.

A recent review of the literature on deforestation and malaria in the Amazon<sup>36</sup> recommended the integration of multiple socio-economic, demographic and ecological mechanisms to disentangle the relationship between deforestation and malaria. The complexity of land-use changes driving deforestation such as urbanization, agriculture, irrigation or resource mining can alter the environment in different ways. For example, deforestation in the Amazon has been shown to increase mosquitoes' larval habitat<sup>41</sup> through the creation of areas with abundant sunlight and pooling water, resulting in increased human biting activity<sup>42</sup>. Alternatively, immigration and rapid population movements, stirring human-vector interactions, are other mechanisms affecting malaria transmission in frontier areas<sup>43</sup>. A modeling study<sup>44</sup> showed that the temporal pattern of increased incidence followed by a decrease can vary depending on ecological and socio-economic parameters in frontier areas.

The importance of addressing complex confounding structures influencing the relationship between deforestation and malaria was also highlighted by Bauhoff et al.<sup>45</sup>. Variables such as temperature<sup>46,47</sup>, precipitation<sup>48,49</sup> or seasonality<sup>50</sup> are known environmental predictors of malaria, although the spatio-temporal scale of those effects often varies across different areas<sup>51</sup>. Furthermore, remote areas may experience higher malaria rates because of poor access to public health services, but also have denser forest cover or lower deforestation rates<sup>52</sup>. Finally, forest-going populations in the GMS are also at higher risk for malaria<sup>26</sup> due to poor adherence to protective measures against mosquitoes such as insecticide-treated bed nets (ITNs) or long-lasting insecticidal hammocks (LLIHs)<sup>24,27</sup> and inadequate access to treatment<sup>23</sup>.

Bauhoff et al.<sup>45</sup> identified only 10 empirical studies that assessed the relationship between deforestation and malaria with appropriate adjustments for confounding. Of these, seven reported a positive association<sup>31,32,35,53–56</sup>, two did not find any associations<sup>45,57</sup>, and one disputed study found a negative association<sup>33,34,58</sup>. Most recently, a study found deforestation to increase malaria risk and malaria to decrease deforestation activities in the Amazon, using an instrumental variable analysis to disentangle any reverse causality loop<sup>59</sup>. However, only half of the above-mentioned studies used high-resolution forest data, with most studies using spatially aggregated data and exploring only a limited range of spatial and temporal scales. Only three of these studies were conducted in Southeast Asia<sup>54–56</sup>, and none in the GMS. Importantly, all three found that malaria increases after deforestation, but all had limitations. The two studies in Indonesia<sup>54,55</sup> used coarsely aggregated forest data and potentially biased self-reported malaria data. The third study, in Malaysia<sup>56</sup>, focused on a specific and geographically confined malaria parasite, *Plasmodium knowlesi*, whose primary host is the long-tailed macaque and whose presence in the GMS, where *P. falciparum* and *P. vivax* dominate, is limited.

In this analysis, we examined the relationship between deforestation and malaria incidence by combining high-resolution forest coverage data<sup>60</sup> and monthly malaria incidence data from 2013 to 2016 from two separate regions in the GMS: northern Lao People's Democratic Republic (PDR) with very low malaria transmission and southern Lao PDR where *P. falciparum* and *P. vivax* are seasonal. By conducting the analysis at the village level, we were able to explore a wide range of spatial scales (1, 10 and 30 km around villages) that might be relevant in characterizing the relationship between deforestation and malaria. In addition, we leveraged the longitudinal nature of both the incidence data collected and the forest data produced from annual

remote sensing imagery<sup>60</sup> to explore the most relevant temporal scales. Finally, we considered alternative definitions of deforestation, restricted to areas with at least certain levels of forest cover, to investigate the type of deforestation driving the relationship with malaria.

To date, no prior studies have quantified the relationship between deforestation and malaria incidence in the GMS. Understanding this relationship is especially important in the GMS, where forest-going activities are a main source of income generation<sup>61</sup> and malaria clusters in forest-going populations<sup>7,24</sup>. To assess the relationship between deforestation and malaria incidence, we modeled the monthly village-level malaria incidence in two regions of Lao PDR using health facility surveillance data and evaluated the most relevant spatio-temporal scale.

## **1.3. Results**

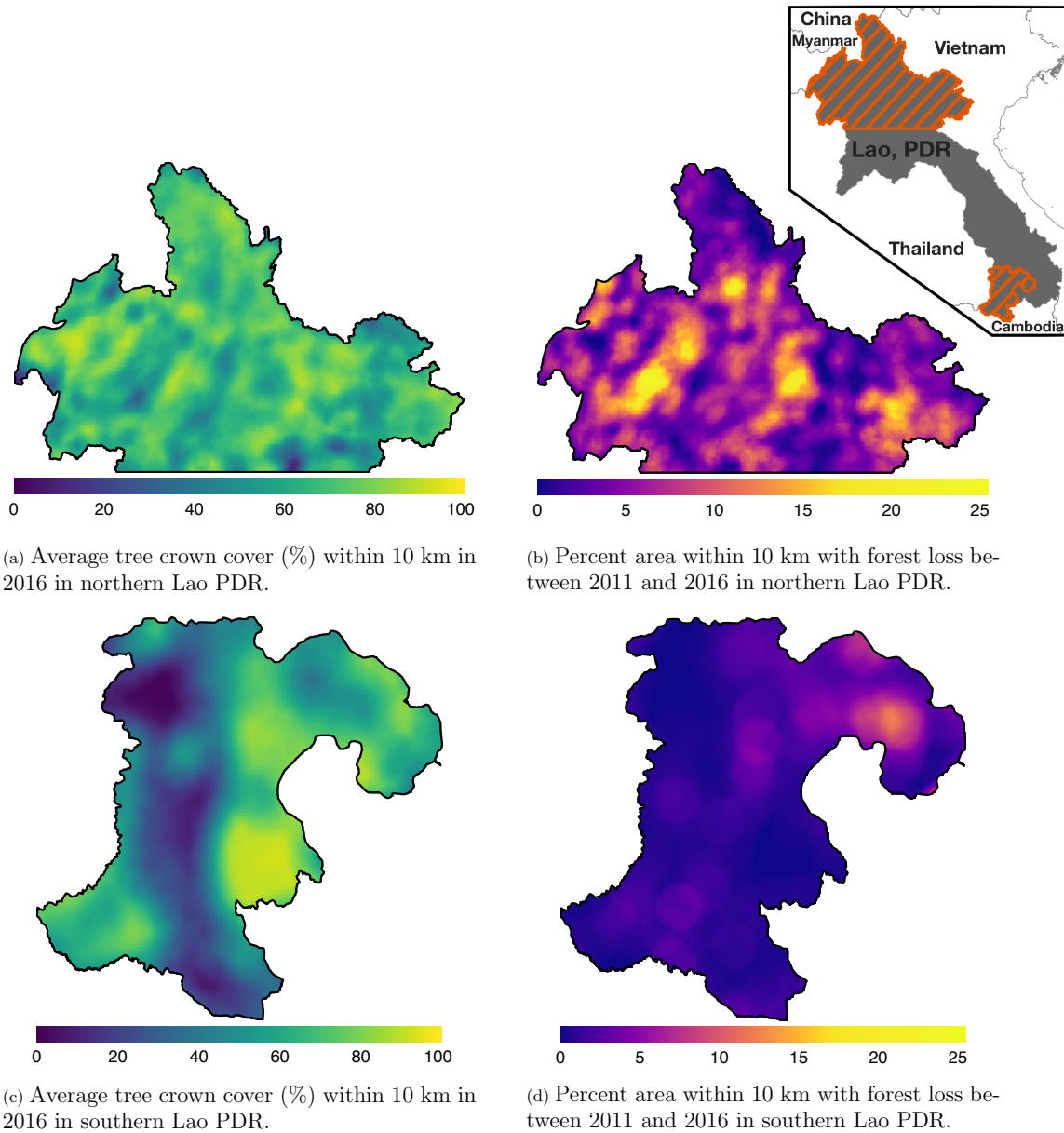
### **1.3.1. Forest and environmental data**

Figure 1.1 shows the average tree crown cover within 10 km for the year 2016 and the percent area within 10 km that experienced forest loss between 2011 and 2016 in two regions of southern and northern Lao PDR. Overall, the forest cover was denser in the north than in the south and deforestation over this period was higher in the north than in the south. Figures 1.14 and 1.15 in the appendix show the distribution of forest and deforestation variables as the temporal scales and spatial scales around study villages were varied. For example, the cumulative percent area within 30km of a village that experienced forest loss between 2011 and 2016 ranged from 0 to 10% in the north, whereas it rarely exceeded 2.5% in the south. Deforestation rate in 2015 within 10 km of a village was of about 1% in the south and 2.5% in the north. The average tree crown cover increased with increasing buffer radius around villages (1, 10 and 30 km). However, the relationship with the percent area that experienced forest loss was less clear because both the area that experienced forest loss (numerator) and the area around villages (denominator) increased. Figures 1.23 and 1.24 in the appendix show the raw time series of forest cover and percent area with forest loss.

The 284 villages in the north were overall less populated (mean 2015 population size: 498, IQR: [241; 548]) than the 207 villages in the south (2015 mean: 1095, IQR: [584; 1384]), but with some highly populated outliers. As expected, altitude differed substantially between villages of the mountainous northern region (mean: 557m, IQR: [378; 679]) and the lowlands of the south (mean: 120m, IQR: [98; 130]). Although both regions exhibited similar seasonal trends in precipitation and temperature, with a rainy season spanning from April to October, villages in the



south experienced higher monthly precipitation and temperature than in the north over the study period (Fig. 1.16 in appendix).



**Figure 1.1** - Average tree crown cover (%) in 2016 (left) and percent area that experienced forest loss between 2011 and 2016 (right) within a 10 km radius in northern (top) and southern (bottom) Lao PDR. See Methods for details on forest and deforestation metrics. Upper right indent maps northern and southern (Champasak province) Lao PDR regions.

### **1.3.2. Treatment-seeking data**

For villages with an estimated travel time of 0 hours to the closest health facility (same 300m<sup>2</sup> pixel), the predicted probability of seeking treatment for fever was 0.87 (95% CI: [0.79; 0.92]) in the north and 0.78 (95% CI: [0.63; 0.89]) in the south. A 1 hour increase in travel time to the closest health facility was associated with a similar 0.79 (95% CI: [0.55; 1.13]) reduction in the odds of seeking treatment in the north and 0.76 (95% CI: [0.43; 1.34]) in the south, almost reaching statistical significance when pooling data from both regions: 0.77 (95% CI: [0.56; 1.04]). See detailed results in Appendix 1 – S1.2.

### **1.3.3. Malaria case data**

#### ***1.3.3.1. Malaria infections***

63,040 patient records were abstracted from the malaria registries of all public health facilities in 4 southern districts between October 2013 and October 2016 and 1,754 from all health facilities in 4 northern districts between January 2013 and December 2016.

In the south, 91.2% of the patients in the registries were tested for malaria, of which 78.1% were tested by RDT and 26.2% by microscopy. Overall test positivity was 33.2% for any infection, 16.4% for *P. falciparum* and 18.2% for *P. vivax*. Monthly incidence peaked to about 6 cases per 1000 people in the 2014 rainy season, eventually decreasing to below 1 case per 1000 in 2016. Incidence and test positivity were similar between *P. falciparum* and *P. vivax* in the south (Fig. 1.2).

In the north, 92.1% of the patients in the registries were tested for malaria, of which 96.3% were tested by RDT and 9.6% by microscopy. Overall test positivity was 23.8% for any infection, 2.8% for *P. falciparum* and 22.5% for *P. vivax*. Monthly malaria incidence in the north was very low, never exceeding 0.3 per 1000 people. Most infections in the north were *P. vivax* cases with only a few seasonal *P. falciparum* cases (Fig. 1.2).

In the appendix, Figure 1.17 shows the number of patients and cases recorded per month in health facility malaria registries as well as how the smoothed test positivity rates varied over time.

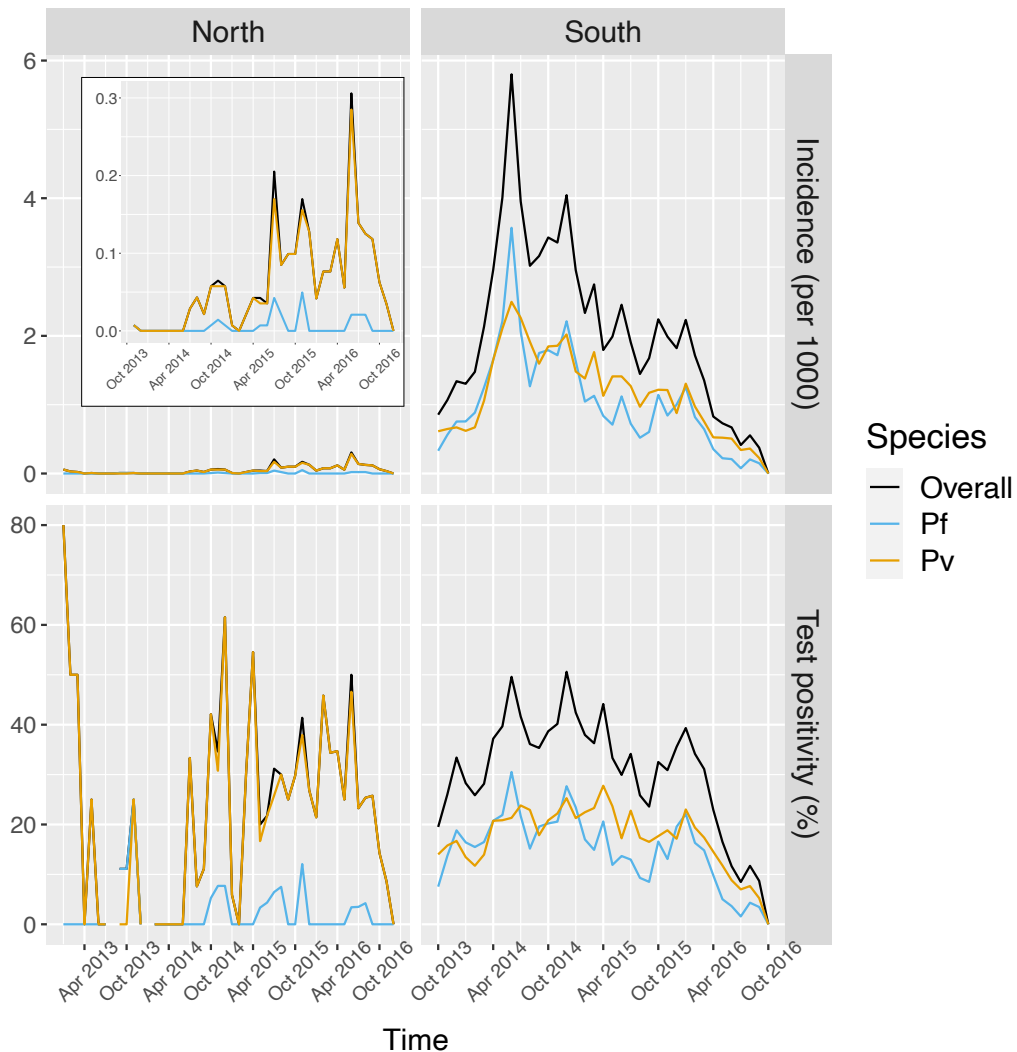
### ***1.3.3.2. Socio-demographics***

Age, gender and occupation of patients seeking treatment at health facilities were also recorded in the malaria registries. On average, patients in the south were older than patients in the north with mean age of 28 years and 23 years respectively. In the north, about half of the patients were male (53.1%), while most patients in the south were male (71.1%). Finally, the vast majority (68.2%) of patients in the south were farmers, whereas only 8% of patients in the north were farmers. Most patients in the north reported being unemployed (41.7%) or a student (31.2%) (Fig. 1.18 in the appendix).

### ***1.3.3.3. Geo-referencing***

Overall, 88.1% of malaria records were matched to one of the 491 villages in study districts. The remaining (11.7% in the south and 17.3% in the north) were removed from the analysis because of ambiguous village names, local nicknames for small villages and dissolving and grouping of

villages over time. Test positivity in the south was similar in matched (33.1%) and unmatched (34.2%) records but higher in matched (26.5%) than unmatched (10.5%) records in the north. No substantial difference was found in the distribution of socio-demographic variables available in malaria registries between matched and unmatched records (Fig. 1.19 in the appendix). Fewer than 0.3% of matched malaria records were missing dates and also removed from the analysis.



**Figure 1.2** - Malaria incidence (per 1000) and test positivity (%) over time. Upper left boxed indent zooms in malaria incidence in the North to better show the temporal variation (See y axis for scale).

## 1.3.4. Spatio-temporal analysis

### 1.3.4.1. Deforestation

Table 1.1 and Figure 1.3 show the adjusted incidence rate ratio (IRR) associated with deforestation, measured by a 0.1% increase in the percent area that experienced forest loss, in the previous 1 to 5 years within 1, 10 and 30 km of villages. Models controlled for various environmental factors and accounted for the probability of seeking treatment and the spatio-temporal structure of the data.

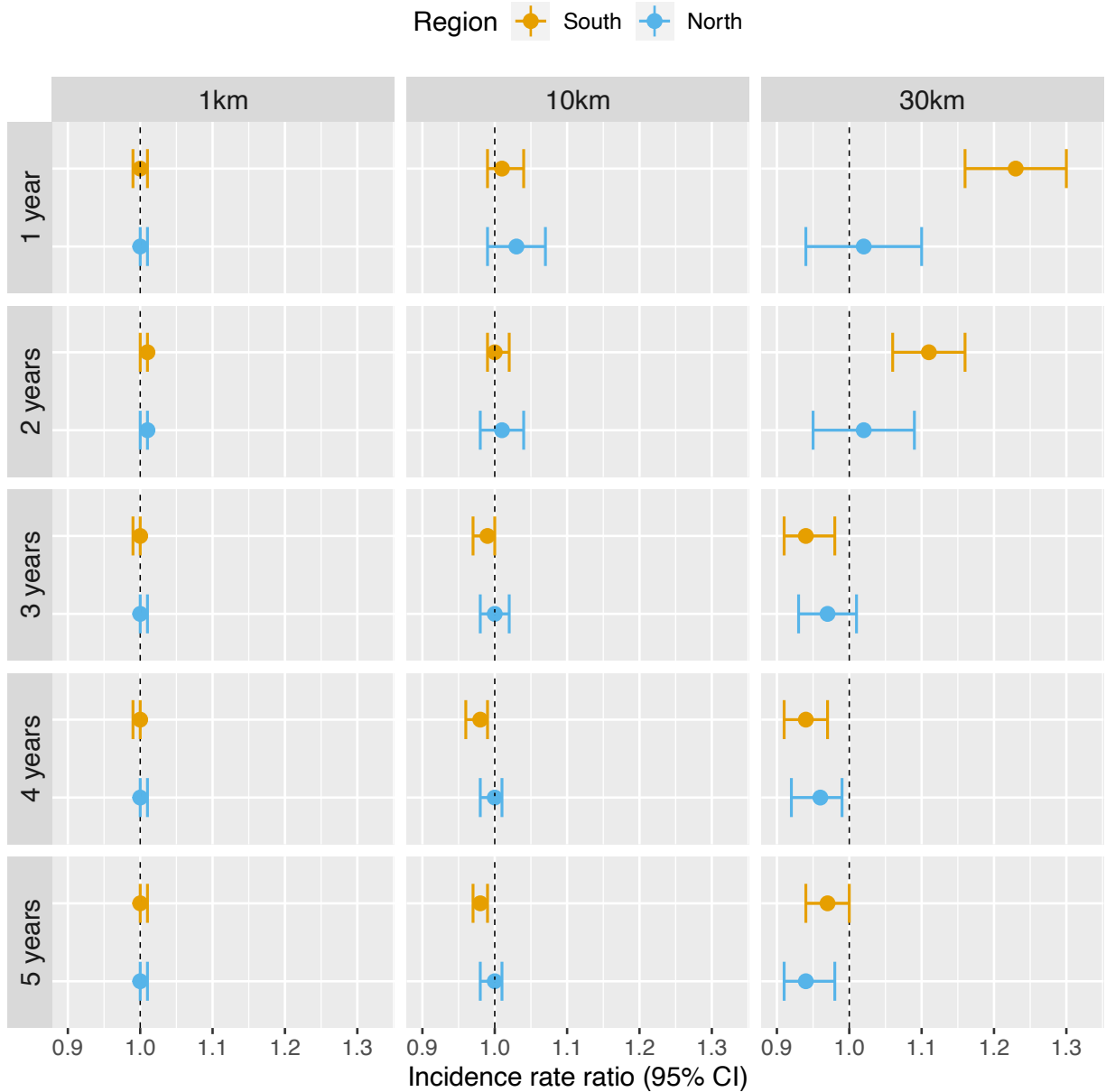
Deforestation within 1 or 10 km of a village was not associated with malaria incidence rate in either the south or the north, regardless of the temporal lag. However, in the south, deforestation within 30 km of a village in the previous 1 and 2 years was associated with higher malaria incidence rates (e.g. 1-year lag, IRR = 1.16, 95% CI: [1.10; 1.22]). In the north, where incidence was much lower, the results were not as clear, but a similar trend was observed with wide confidence intervals compatible with a short-term increased risk. On the other hand, deforestation within 30 km of a village in the previous 3, 4 or 5 years was associated with approximately a 5% lower malaria incidence rate both in the south (e.g. 5-year lag, IRR = 0.94, 95% CI: [0.91; 0.97]) and in the north (e.g. 5-year lag, IRR = 0.96, 95% CI: [0.93; 0.98]).

These results suggest deforestation around villages, but not in the near vicinity (1 or 10 km), is associated with higher risk of malaria in the first two years and lower risk of malaria beyond. There was stronger evidence of associations with deforestation in the south than in the north.

The IRR effect estimates in Table 1.1 assume a linear relationship between deforestation and malaria. Figure 1.13 in the appendix shows a few of these relationships when such linearity isn't assumed in the models. The functional forms reveal that they can be reasonably well summarized linearly, especially in the south. In the north, the functional forms highlight potential non-linearities for long term temporal lags but come with wide confidence intervals at extreme levels of deforestation.

**Table 1.1** - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km (left-right) of a village in the previous 1 to 5 years (top-down) in northern and southern Lao PDR. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 year. See Methods for details.

Time lag	South			North		
	1 km	10 km	30 km	1 km	10 km	30 km
Previous 1 year	1 [0.99; 1.01]	1.01 [0.99; 1.04]	1.16 [1.10; 1.22]	1 [1; 1.01]	1.03 [0.99; 1.06]	1.01 [0.94; 1.08]
Previous 2 years	1 [0.99; 1.01]	1 [0.98; 1.01]	1.08 [1.04; 1.13]	1 [1; 1.01]	1.01 [0.99; 1.04]	0.99 [0.95; 1.03]
Previous 3 years	0.99 [0.99; 1]	0.98 [0.97; 1]	0.93 [0.90; 0.97]	1 [1; 1.01]	1.01 [0.99; 1.02]	0.96 [0.94; 0.99]
Previous 4 years	0.99 [0.99; 1]	0.98 [0.97; 0.99]	0.94 [0.92; 0.97]	1 [1; 1.01]	1 [0.99; 1.02]	0.97 [0.94; 0.99]
Previous 5 years	0.99 [0.99; 1]	0.97 [0.96; 0.99]	0.94 [0.91; 0.97]	1 [1; 1.01]	1 [0.99; 1.02]	0.96 [0.93; 0.98]



**Figure 1.3** - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km (left-right) of a village in the previous 1 to 5 years (top-down) in Lao PDR. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 year. See Methods for details.

#### 1.3.4.2. *P. falciparum* and *P. vivax*

In addition to different overall levels of transmission in the north and south, the relative species composition also differs by region. In the north, *P. vivax* is more prevalent with only a few sporadic and seasonal *P. falciparum* infections, whereas *P. falciparum* and *P. vivax* are co-endemic in the south (Fig. 1.2). We used the co-endemicity and the larger amount of malaria case data collected in the south to assess the relationship between deforestation and malaria for both species separately.

Table 1.2 and Figure 1.4 show that the pattern of adjusted spatio-temporal associations identified in Table 1.1 is primarily driven by *P. falciparum*, with no associations for deforestation in the near vicinity of villages (1 or 10 km) but a short-term increase (e.g. 1-year lag, IRR = 1.27, 95% CI: [1.18; 1.36]) and long-term decrease (e.g. 5-year lag, IRR = 0.83, 95% CI: [0.80; 0.87]) in *P. falciparum* malaria incidence for deforestation within 30 km of villages.

On the other hand, all the associations were attenuated for *P. vivax* infections. In the previous 2 years and within 30 km of villages, deforestation is still associated with a higher incidence of *P. vivax* (e.g. 1-year lag, IRR = 1.07, 95% CI: [1.01; 1.13]) but less so than for *P. falciparum*.

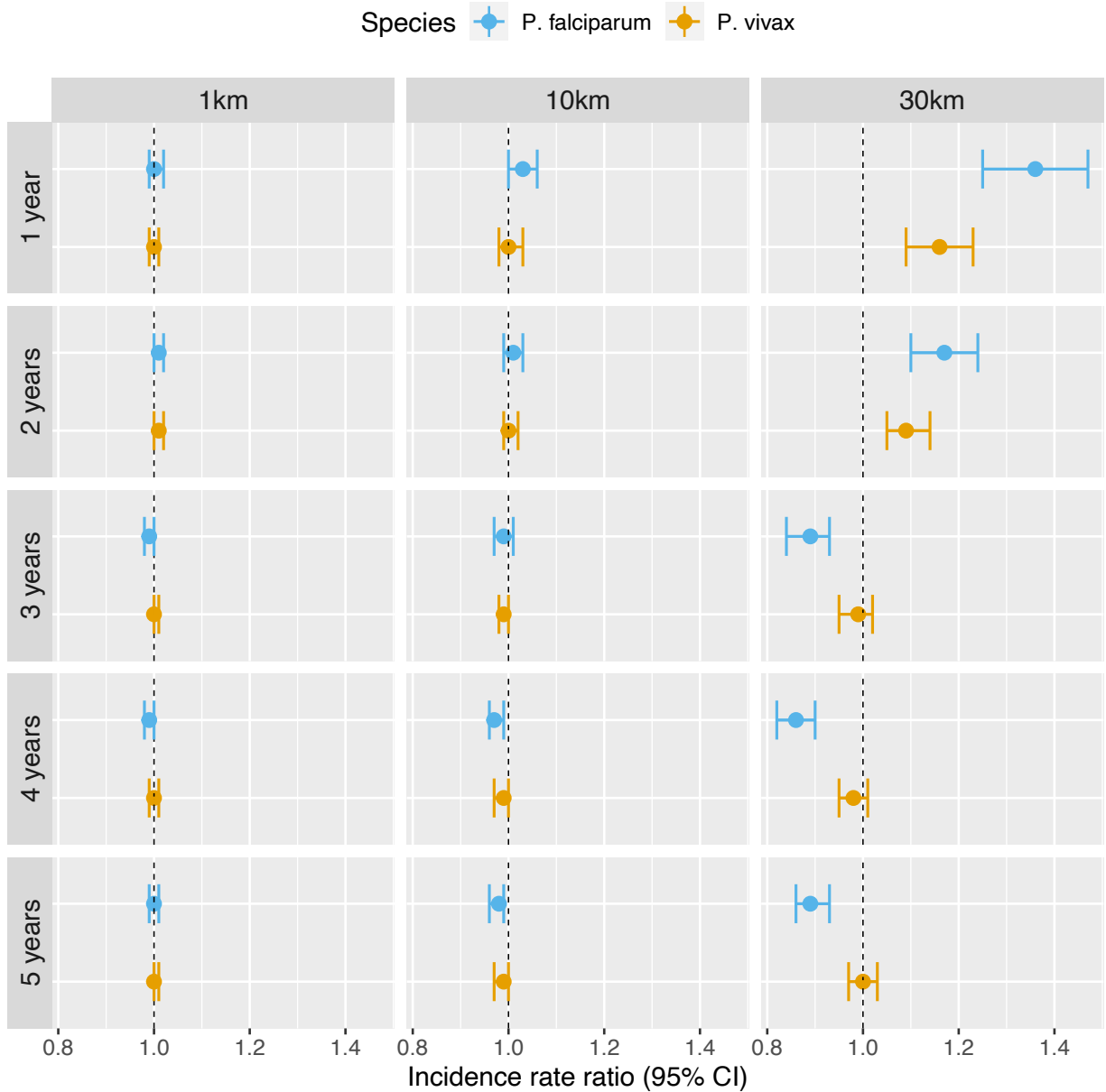
However, regardless of the temporal lag or spatial scale, deforestation was no longer associated with lower *P. vivax* malaria risks.

Figure 1.25 in the appendix plots the species-specific relationships when not assuming linearity in the models.



**Table 1.2** - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km (left-right) of a village in the previous 1 to 5 years (top-down) in southern Lao PDR, differentiated by malaria species. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 year. See Methods for details.

Time lag	<i>P. falciparum</i>			<i>P. vivax</i>		
	Buffer radius			Buffer radius		
	1 km	10 km	30 km	1 km	10 km	30 km
Previous 1 year	1 [0.99; 1.02]	1.04 [1.01; 1.07]	1.27 [1.18; 1.36]	1 [0.99; 1.01]	1 [0.97; 1.02]	1.07 [1.01; 1.13]
Previous 2 years	1 [0.99; 1.01]	1.01 [0.99; 1.03]	1.15 [1.08; 1.22]	1 [0.99; 1.01]	1 [0.98; 1.01]	1.06 [1.01; 1.11]
Previous 3 years	0.99 [0.98; 1]	0.99 [0.97; 1.01]	0.85 [0.80; 0.90]	1 [0.99; 1.01]	0.99 [0.98; 1.01]	1.02 [0.97; 1.06]
Previous 4 years	0.99 [0.98; 1]	0.98 [0.96; 0.99]	0.85 [0.81; 0.88]	1 [0.99; 1]	0.99 [0.98; 1.01]	1.01 [0.98; 1.04]
Previous 5 years	0.99 [0.98; 1]	0.97 [0.95; 0.98]	0.83 [0.80; 0.87]	1 [1; 1.01]	0.99 [0.98; 1]	1.01 [0.98; 1.04]



**Figure 1.4** - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 1, 10 or 30 km (left-right) of a village in the previous 1 to 5 years (top-down) in southern Lao PDR, differentiated by malaria species. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 year. See Methods for details.

### ***1.3.4.3. Alternative definitions of deforestation and interaction with forest cover***

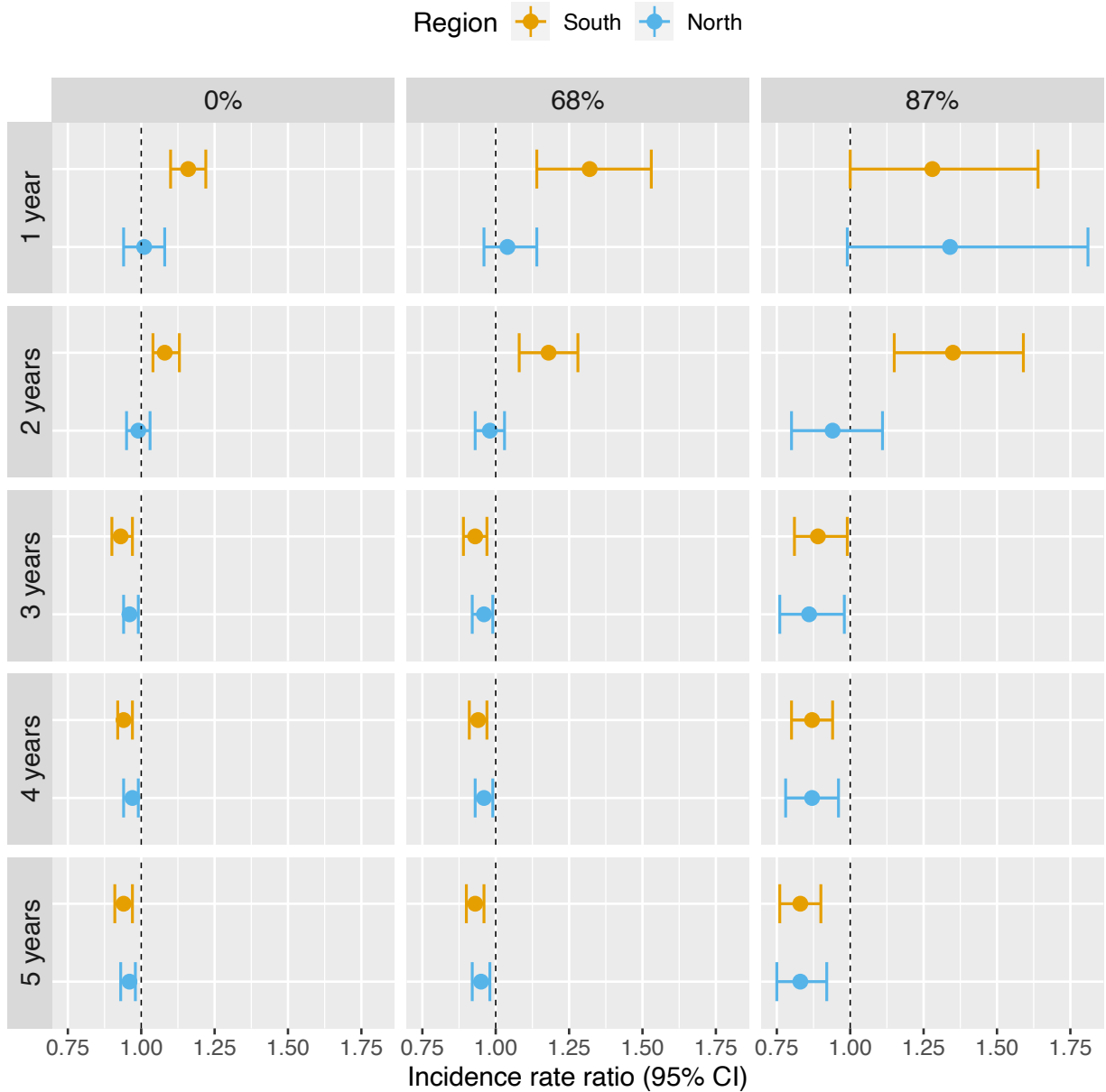
In previous models, our definition of deforestation did not distinguish between forest losses in densely forested areas and less forested areas. To explore potential interactions between deforestation and baseline forest cover, Table 1.3 and Figure 1.5 show how the adjusted IRR estimates vary as we consider deforestation in more densely forested pixels only (tree crown cover over 68% and 87% - see Methods for rationale on thresholds). We conducted this secondary analysis only for the non-null relationships previously identified, i.e. when considering a 30 km buffer radius around villages.

The associations with deforestation became more pronounced as we restricted forest losses to more forested areas: the adjusted IRR for deforestation in the previous 1 year, within 30 km of southern villages, increased from 1.16 (95% CI: [1.10; 1.22]) to 1.28 (95% CI: [1; 1.64]) when considering deforestation in areas with more than 0% and 87% tree crown cover respectively. On the other hand, the adjusted IRR for deforestation in the previous 5 years, within 30 km of southern villages, decreased from 0.94 (95% CI: [0.91; 0.97]) to 0.83 (95% CI: [0.76; 0.90]) when considering deforestation in areas with more than 0% and 87% tree crown cover respectively. A similar trend was observed in the north, although statistical significance wasn't reached as frequently as in the south.

These evidence strengthen our previous results and suggest that deforestation in deep and dense forests is more closely associated with malaria incidence in villages than deforestation in less forested areas.

**Table 1.3** - IRR between malaria incidence and a 0.1% increase in the area that experienced deforestation within 30 km of a village in the previous 1 to 5 years (top-down) and within areas with tree crown cover density above 0%, 68% and 87% (left-right) in Lao PDR. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 years. See Methods for details.

Time lag	South			North		
	Deforestation within areas with tree crown cover density above			Deforestation within areas with tree crown cover density above		
	0%	68%	87%	0%	68%	87%
Previous 1 year	1.16 [1.10; 1.22]	1.32 [1.14; 1.53]	1.28 [1; 1.64]	1.01 [0.94; 1.08]	1.04 [0.96; 1.14]	1.34 [0.99; 1.81]
Previous 2 years	1.08 [1.04; 1.13]	1.18 [1.08; 1.28]	1.35 [1.15; 1.59]	0.99 [0.95; 1.09]	0.98 [0.96; 1.03]	0.94 [0.80; 1.11]
Previous 3 years	0.93 [0.90; 0.97]	0.93 [0.89; 0.97]	0.89 [0.81; 0.99]	0.96 [0.94; 0.99]	0.96 [0.92; 0.99]	0.86 [0.76; 0.98]
Previous 4 years	0.94 [0.92; 0.97]	0.94 [0.91; 0.97]	0.87 [0.80; 0.94]	0.97 [0.94; 0.99]	0.96 [0.93; 0.99]	0.87 [0.78; 0.96]
Previous 5 years	0.94 [0.91; 0.97]	0.93 [0.90; 0.97]	0.83 [0.76; 0.90]	0.96 [0.93; 0.98]	0.95 [0.92; 0.98]	0.83 [0.75; 0.92]



**Figure 1.5** - Associations between malaria incidence and a 0.1% increase in the area that experienced deforestation within 30 km of a village in the previous 1 to 5 years (top-down) and within areas with tree crown cover density above 0%, 68% and 87% (left-right) in Lao PDR. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered as well as for malaria incidence in the previous 1 and 2 years. See Methods for details.

## 1.4. Discussion

Based on a large dataset of health facility surveillance records in two regions of Lao PDR, we found evidence that deforestation around villages is associated with higher malaria incidence over the short-term but lower incidence over the long-term (e.g. in the south, within 30 km of villages: IRR = 1.16 [1.10; 1.22] for deforestation in the previous year and IRR = 0.93 [0.90; 0.97] for deforestation in the previous 3 years). Our evaluation of alternative spatial scales identified strong associations for deforestation within a 30 km radius around villages but not for deforestation in the near (10 km) and immediate (1 km) vicinity. Our results incorporated correction for the probability of seeking treatment, modeled as a function of distance to the closest health facility, as well as adjustment for several environmental covariates. Results appear driven by deforestation in densely forested areas and the patterns exhibited are clearer for infections with *P. falciparum* than for *P. vivax*.

The wide availability and longitudinal nature of malaria surveillance records collected routinely by the national program enabled exploration of the relationship between deforestation and malaria incidence over multiple spatio-temporal scales and across different levels of forest density. The spatio-temporal variability highlighted here provides insights into the causal mechanisms driving local-scale malaria incidence in the GMS. This approach not only quantified the deforestation-malaria incidence association in the GMS, but also strengthened the evidence for the key influence of forest-going populations on malaria transmission in the GMS.

This study's results echo the frontier malaria hypothesis from the Amazon region, which posits an increase in malaria incidence in the first few years following deforestation and a decrease

over the long term. However, we found an earlier inflexion point, 1-3 years after deforestation compared to 6-8 years in the Amazon<sup>38</sup>, most likely because of very different underlying human processes. Indeed, the frontier malaria hypothesis considers non-indigenous human settlements sprouting deeper and deeper in the forest whereas forest-going populations in the GMS are primarily members of established forest-fringe communities who regularly tour the forest overnight to hunt and collect wood<sup>25</sup>. Industrial and agricultural projects or lucrative forest-based activities also attract mobile and migrant populations (MMPs)<sup>23</sup> in remote forested areas of the GMS but not on the same scale as the politically and economically driven unique colonization of the Amazon<sup>38</sup>.

Our results are also consistent with the three previous multivariable empirical studies<sup>54-56</sup> that assessed the effect of deforestation on malaria in Southeast Asia. Our study builds on these findings by using higher resolution forest data and exploring additional spatio-temporal scales. Using biennial village census data from Indonesia between 2003 and 2008 and district-aggregated remote sensing forest data, Garg<sup>55</sup> reported a 2 to 10.4% increase in the probability of a malaria outbreak in each village of districts that lost 1000 hectares of their forest cover in the same year. Using data from a 1996 cross-sectional household survey conducted in a quasi-experimental setting around a protected area in Indonesia, Pattanayak et al.<sup>54</sup> found a positive association between disturbed forest (vs undisturbed) and malaria in children under 5, again using no temporal lag. Our analysis plan was largely inspired by Fornace et al.<sup>56</sup>, which used similar high-resolution forest data<sup>60</sup> and 2008-2012 incidence data from Sabah, Malaysia. They reported a 2.22 (95% CI: [1.53; 2.93]) increase in the *P. knowlesi* incidence rate for villages where more than 14% (< 8%, being the reference) of the surrounding area (within 2 km)

experienced forest loss in the previous 5 years. On the other hand, our analysis explored wider spatial scales, bypassed any coarse categorization of forest and deforestation variables, corrected incidence for treatment-seeking probability, and most importantly focused on *P. falciparum* and *P. vivax*, the dominant malaria parasites in the GMS.

Engaging in forest activities, such as logging, hunting or spending the night in the forest, has been reported as a major risk factor by many studies in the region<sup>17–22</sup>. As countries of the GMS work towards malaria elimination, the literature stresses the key role of forest-going populations<sup>23,28,62–64</sup>, although research programs highlight the challenges of accessing them<sup>65,66</sup> as well as their diversity<sup>28,62</sup>. To our knowledge, no previous study has leveraged geo-spatial statistical analyses to characterize the importance of forest-going populations in the GMS. Our results suggest that deforestation in dense forests (Table 1.3) around villages, particularly areas further from the village (Table 1.1), is a driver of malaria in Lao PDR. We argue that this is indicative of the existence of a key high-risk group linking the deforestation patterns identified to malaria in the villages, namely a forest-going population. Deforestation captured by remote sensing in this setting likely reflects locations and times of heightened activity in the forest areas near villages, and therefore greater human-vector contact. We suspect longer and deeper trips into the forest result in increased exposure to mosquitoes, putting forest-goers at higher risk.

We conducted this study in northern and southern Lao PDR, where the malaria species composition differs, and assessed species-specific relationships in the south where *P. falciparum* and *P. vivax* are co-endemic. Our results highlight the challenges ahead of national programs with *P. vivax* elimination after successful *P. falciparum* elimination, as increasingly mentioned



in the literature<sup>67,68</sup>. This study identified a clear pattern of spatio-temporal associations between *P. falciparum* and deforestation, but these were not apparent for *P. vivax* (Table 1.2). The increase in *P. vivax* incidence in the first 2 years following deforestation was identified as well but the associations were smaller than for *P. falciparum*. Importantly, deforestation was never associated with lower risks of *P. vivax*. A recent study in the Amazon<sup>59</sup> reported a similar attenuation of the effects of deforestation on *P. vivax* compared to *P. falciparum*, most likely because of *P. vivax* parasites' ability to relapse months or even years after infection, which decouples the association between transmission and incidence data. These species-specific differences may also explain why the pattern of spatio-temporal associations between malaria and deforestation were markedly clearer in the south than in the north where *P. vivax* dominates.

Our results did have some inherent limitations based upon routine health facility surveillance data. First, reliability of such records varies across and within countries of the GMS and may depend on malaria incidence level. This could lead to unmeasured residual confounding, further exacerbated by the lack of available data on malaria control activities in the region. Another challenge with these data is obtaining an accurate denominator for incidence, as not everyone attends a public health center when febrile. We addressed this issue by modeling the probability of seeking treatment as a function of travel time to the closest health facility using data from two cross-sectional surveys. Last, the village-level geo-referencing of malaria registries ignores the possibility that patients may become infected elsewhere. Unfortunately, these surveillance records did not include information about patients' forest-going trips. Research to track and analyze micro-scale movements of forest-goers is needed to understand how they interact with the forest and where are the foci of infection.

The forest data we used has also been criticized, in particular for not distinguishing tropical forests from agroforestry<sup>69,70</sup> or man-made from natural causes of deforestation. The lack of temporal resolution for the forest gain variable (2000-2017 aggregate) as well as the assumption that forest loss happens all in 1 year are additional limitations of these data. Finally, our relative measure of deforestation, key to consistently compare the effects across different spatial scales, also implies that a 0.1% of the area that experienced forest loss within 30 km of a village is a much larger area (~ 280 hectares) than within 1 km (~ 0.3 hectare) and should be interpreted cautiously.

In conclusion, this study assessed the relationship between deforestation and malaria in Lao PDR. Our approach leveraged surveillance records collected by the national malaria program and high-resolution forest data and rigorously explored the spatio-temporal pattern of associations. As countries of the GMS work towards malaria elimination, our results highlight the challenges to transition from *P. falciparum* to *P. vivax* elimination, confirm and characterize the importance of high-risk populations engaging in forest activities and suggest malaria programs may benefit from monitoring areas of on-going deforestation using remotely sensed data.

## 1.5. Materials and methods

### 1.5.1. Study site and population

Lao PDR has seen a 92% reduction in cases between 2000 (280,000) and 2010 (23,000)<sup>61</sup>. Much of this progress has been attributed to heightened funding and better testing and treatments<sup>2</sup>.

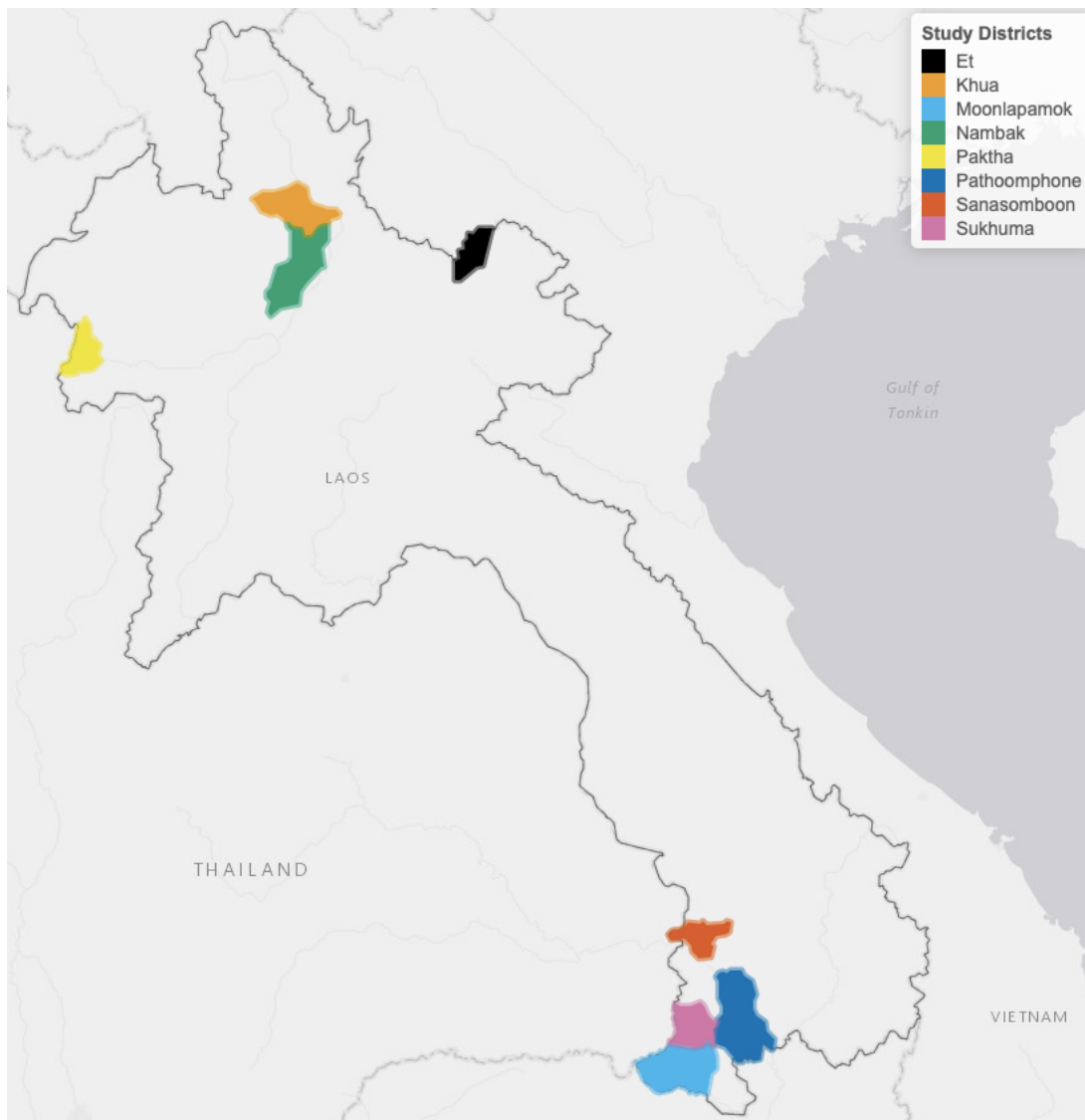
This study was conducted in eight districts (Fig. 1.6) to leverage the ecological and epidemiological diversity of Lao PDR. Four districts (Moonlapamok, Pathoomphone, Sanasomboon and Sukhuma) are situated in the southern province of Champasak where both *P. falciparum* and *P. vivax* are endemic. The four other districts (Et, Paktha, Nambak and Khua) each come from one of four northern provinces (Bokeo, Huaphanh, Phongsaly, Luang-Prabhang) where *P. vivax* is endemic but *P. falciparum* has reached historical lows<sup>71</sup>.

The four districts in the north were chosen in consultation with district and provincial level malaria staff to represent the epidemiology of malaria in the region. They were selected as part of a cross-sectional survey designed to assess the prevalence and risk factors for malaria in northern Lao PDR<sup>72</sup>. This region is very mountainous and characterized by a diverse climate, low-population density and limited road access<sup>73</sup>. Land clearing using fires for agriculture is customary.

The 4 districts in the south were selected within a larger cluster randomized controlled trial (RCT) study designed to assess the effectiveness of high-risk group targeted active case detection in southern Lao PDR<sup>66</sup>, where more than 95% of the country malaria burden is

concentrated<sup>71</sup>. This region is characterized by a moderately hilly and forested terrain and a workforce primarily engaged in forest-based and agricultural activities<sup>65</sup>.

When designing the study, in collaboration with the national control program, we purposefully excluded regions where we knew large programmatic activities were being implemented.



**Figure 1.6 - Map of study's districts.**

## **1.5.2. Malaria data**

### ***1.5.2.1. Malaria case data***

We conducted a retrospective review of malaria registries recorded at all health centers in the study districts between January 2013 and December 2016 in the north and between October 2013 and October 2016 in the south. The registries included information on every patient that was tested (RDT and/or microscopy) for malaria at the health center. Date, species-specific test results, demographic variables (age, gender and occupation) and the village of residence of the patient were recorded in the registries. With help from local Lao experts, village names were matched to a geo-registry of all villages in Lao PDR compiled from the 2005 and 2015 national census<sup>74</sup> and provided by the Center for Malariology, Parasitology and Entomology (CMPE). The geo-registry contains GPS coordinates and population of Lao PDR's villages. Unmatched records and records with missing date were removed from the analysis. Finally, these data were aggregated to extract the monthly village-level malaria incidence.

### ***1.5.2.2. Treatment-seeking data***

One issue with using passive surveillance data is that not everyone will seek treatment at a public health facility for a febrile illness, which can lead to an underestimate of the true incidence, if not accounted for. To correct for that, we modeled the probability that an individual in a given village of the study's district would seek treatment at a public health facility when febrile. We assumed that such probability is essentially driven by the travel time to the closest health facility. See Appendix 1 – S1.1 for methods used to calculate travel times to closest health facilities.

To model the probability of seeking treatment, we used data from two cross-sectional household surveys conducted in the eight districts where registries were collected. In the north, 1,480 households across 100 villages were surveyed in September-October 2016<sup>72</sup>. In the south, 1,230 households across 56 villages were surveyed in the baseline assessment of the RCT<sup>66</sup> in December 2017. In particular, survey respondents were asked whether or not they would seek treatment at the closest health facility for a febrile illness and GPS coordinates of their household were recorded.

We then used the cross sectional surveys to model the probability of seeking treatment (at a public health facility, implicit from now on),  $\theta$ , as a function of travel time to the closest health facility,  $\tau$  (Equation 1.1). To account for the correlation structure induced by the stratified sampling approach used in the surveys, we modeled the number of successes (febrile patients seeking treatment),  $S_{h,v}$ , at the household level and included a random intercept for village in the logistic regression.

$$S_{h,v} \sim Bin(\theta_{h,v}, N_{h,v})$$

$$\text{logit}(\theta_{h,v}) = \alpha_0 + \alpha_1 \times \tau_{h,v} + \alpha_v \quad (\text{Eq 1.1})$$

where  $N_{h,v}$  is the number of febrile individuals in household  $h$  of village  $v$  and  $\alpha_v \sim \mathcal{N}(0, \sigma_\alpha)$ .

We fit the models separately in the north and in the south and used the region-specific model to predict the probability of seeking treatment at all villages of the study districts based on their distance to the closest health facility. The population who seek treatment was then calculated by

multiplying the village population by the probability of seeking treatment. See Appendix 1 – S1.2 for travel times and treatment-seeking probabilities results.

### **1.5.3. Forest data**

For every 30m pixel in Lao PDR, tree crown cover density for the year 2000 and year of forest loss between 2000 and 2017, were obtained from Hansen et al.<sup>60</sup>. These layers were produced using decision tree classifiers on Landsat remote sensing imagery<sup>60</sup>. Trees are defined as “all vegetation taller than 5m in height”<sup>60</sup> and forest loss as “the removal or mortality of all tree cover in a Landsat pixel”<sup>70</sup>. For example, as depicted in Fig. 1.7, the Hansen data indicates that the tree crown cover in 2000 in pixel 1 is 54%, meaning that 54% of the 30m pixel is covered by vegetation taller than 5m. The Hansen data also indicates that forest loss occurred in pixel in 2013, meaning that all of the tree canopy disappeared in 2013.

#### ***1.5.3.1. Deforestation variable***

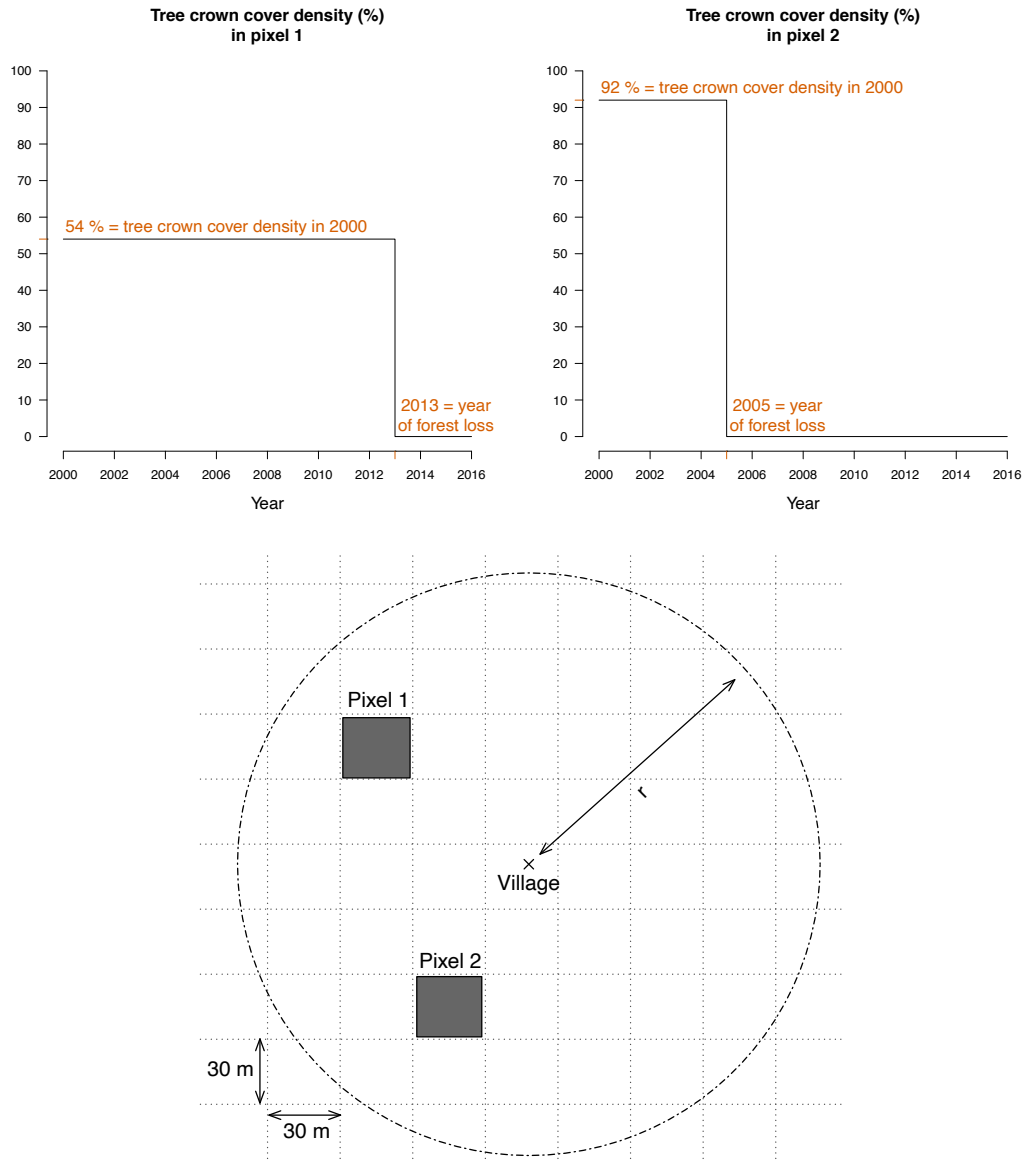
To define our primary exposure variable, for all villages in the study districts and year of the study period, we calculated the percent area within a buffer radius of 1, 10 and 30 km that experienced forest loss in the previous 1, 2, 3, 4 and 5 years (Fig. 1.7). These distances were chosen to explore a range of spatial scales at which the forest environment may be differentially relevant for village-based populations and forest-goers. To explore potential interactions between deforestation and forest cover, we computed an alternate exposure variable, restricting to areas that both experienced forest loss and had a tree crown cover density above 68% and 87%. Those thresholds are limits of the inter-quartile range (IQR) of the distribution of tree crown cover density in any 30m pixels within 10 km of study's villages that experienced forest loss between

2000 and 2017. This alternate definition captures deforestation activities occurring in areas with denser forest cover.

### ***1.5.3.2. Forest cover variable***

We also combined the two Hansen layers to produce annual tree crown cover maps of the study districts, assuming no changes prior to the year of forest loss but setting to 0 the pixel tree crown cover density afterwards (Fig. 1.7). For all villages in the study districts and year of the study period, we calculated the average tree crown cover density within a buffer radius of 1, 10 and 30 km and for 0, 1, 2 and 3-year lags. This is a secondary exposure, adjusted for in the primary analysis.





**Figure 1.7** – Forest data methods: for every 30m Landsat pixel within a buffer radius  $r$  (1, 10 and 30 km) of study's villages, the tree crown cover density in 2000 and the year of forest loss were combined to derive the deforestation and forest cover variables. The two upper plots highlight the raw data at two example pixels from the lower plot.

#### **1.5.4. Environmental covariates**

Village population sizes were needed to estimate monthly malaria incidence. 2005 and 2015 population estimates for the 491 villages of study districts were obtained from the national census<sup>74</sup>. The annual population growth rate (3.7%) was used to impute population values for two villages missing 2005 estimates and for two villages missing 2015 estimates. Then, village-level population growth rates were used to estimate villages' population per year between 2008-2016, assuming linear annual growth rate (median = 1.7%, IQR = [0%; 4.5%]).

Altitude, temperature, rainfall and access to health care were considered as potential village-level confounders of the relationship between malaria and forest cover factors. Travel time to closest health facility, computed for the treatment-seeking model, was used as a proxy for health care access and villages' remoteness. Altitude was extracted from SRTM<sup>75</sup> 1 km resolution layers. Monthly average day and night temperature were extracted from MODIS 1 km resolution product (MOD11C3<sup>76</sup>). Finally, monthly total rainfall was extracted from CHIRPS<sup>77</sup> 1 km resolution publicly available data. The average and standard deviation of the annual total precipitation and the average monthly temperature from the monthly time series was computed over the 2008-2012 period, which corresponds to the 5-year time period directly before our malaria data (2013-2016). This “long-term” aggregation of the climatic variables is included in the model to capture the spatial differences in overall climate between the villages of our study area. To account for the seasonal effect of these climatic variables, monthly temperatures and precipitation in the previous 1, 2 and 3 months were also extracted, as well as the average temperatures and total precipitation over the previous 1, 2 and 3 months (Seven “short-term”

variations: in current month, in previous 1, 2 or 3 months and aggregated over current and previous 1, 2 or 3 months). See “Details on covariates” below.

Altitude was missing for one village and we used an online elevation finder tool (FreeMapTools) for imputation. Temperature was missing for 2.4% of the village-months over the study period, most likely because of cloud coverage of the MODIS imagery. Monthly temperature was never missing more than two years in a row at villages of the study's districts and we imputed the temperature of the same month of the following year (or prior year when needed), adjusting for average district-level monthly temperature differences between the two consecutive years. Monthly rainfall was not missing at any of the villages.

## **1.5.5. Statistical analysis**

### ***1.5.5.1. Statistical model***

To model malaria incidence (Equation 1.2), the number of positive cases  $Y_{v,t}$  at village  $v$  over month  $t$  was modeled using a generalized additive model (GAM)<sup>78</sup>. To account for overdispersion, a negative binomial distribution was used, including an additional variance parameter  $\upsilon$ . The probability of seeking treatment  $\theta_v$ , estimated from the treatment-seeking model, was multiplied by the village population  $Pop_{v,t}$  to derive the population seeking treatment,  $Pop_{v,t}^{seek}$ . This was included as an offset term in the incidence model. Spatial autocorrelation was accounted for by the bivariate thin plate spline smoothing function on coordinates,  $f(\text{Lat}, \text{Long})$  and village random intercepts were included. A non-linear temporal trend was also included with the smoothing function on month,  $f(t)$ . Finally, the primary exposure, deforestation, and potential environmental confounders, including forest cover, were

modeled with splines in  $f(X_{v,t}^i)$ . Splines add up polynomial basis functions in between knots and allow to control for very flexible relationships with covariates and spatio-temporal trends. Regularization was used to integrate model selection into the model fitting step by adding an extra penalty to each term so that the coefficients for covariates can be penalized to zero, also meaning that splines can be kept minimal if the data does not support more flexibility. See Figure 1.8 for a graphical visualization of our conceptual model for this analysis.

$$\begin{aligned}
 Y_{v,t} &\sim \text{NegBin}(E[Y_{v,t}], \nu) \\
 \log(E[Y_{v,t}]) &= \log(\mu_{v,t} \times \text{Pop}_{v,t}^{\text{seek}}) = \log(\mu_{v,t}) + \log(\text{Pop}_{v,t} \times \theta_{v,t}) \quad (\text{Eq 1.2}) \\
 \log(\mu_{v,t}) &= \sum_i \beta^i \times f(X_{v,t}^i) + f(\text{Lat}, \text{Long}) + f(t) + \beta_v
 \end{aligned}$$

with  $\beta_v \sim \mathcal{N}(0, \sigma_\beta)$ .

We ran 15 models separately in the north and the south, each varying the buffer radius (1, 10 and 30 km) and temporal scale for deforestation (previous 1, 2, 3, 4 and 5 years). The coefficients of the linear effect for deforestation were extracted and exponentiated to get the incidence rate ratio (IRR) associated with a 0.1% increase in the percent area that experienced forest loss around villages.

### 1.5.5.2. *Secondary analyses*

As secondary analyses, we ran the same models, separately by malaria species (*P. falciparum* and *P. vivax*), leveraging the large amount of data and co-endemicity in the south. We also used our alternative definitions for deforestation, restricted to areas that both experienced forest loss

and had a tree crown cover density above 68% and 87%, to explore the interaction between deforestation and the amount of forest cover.

To further strengthen the robustness of our analysis, we conducted a sensitivity analysis where villages' populations in the surveillance system registries were not adjusted for the probability of seeking treatment. See Appendix 1 – S1.3.

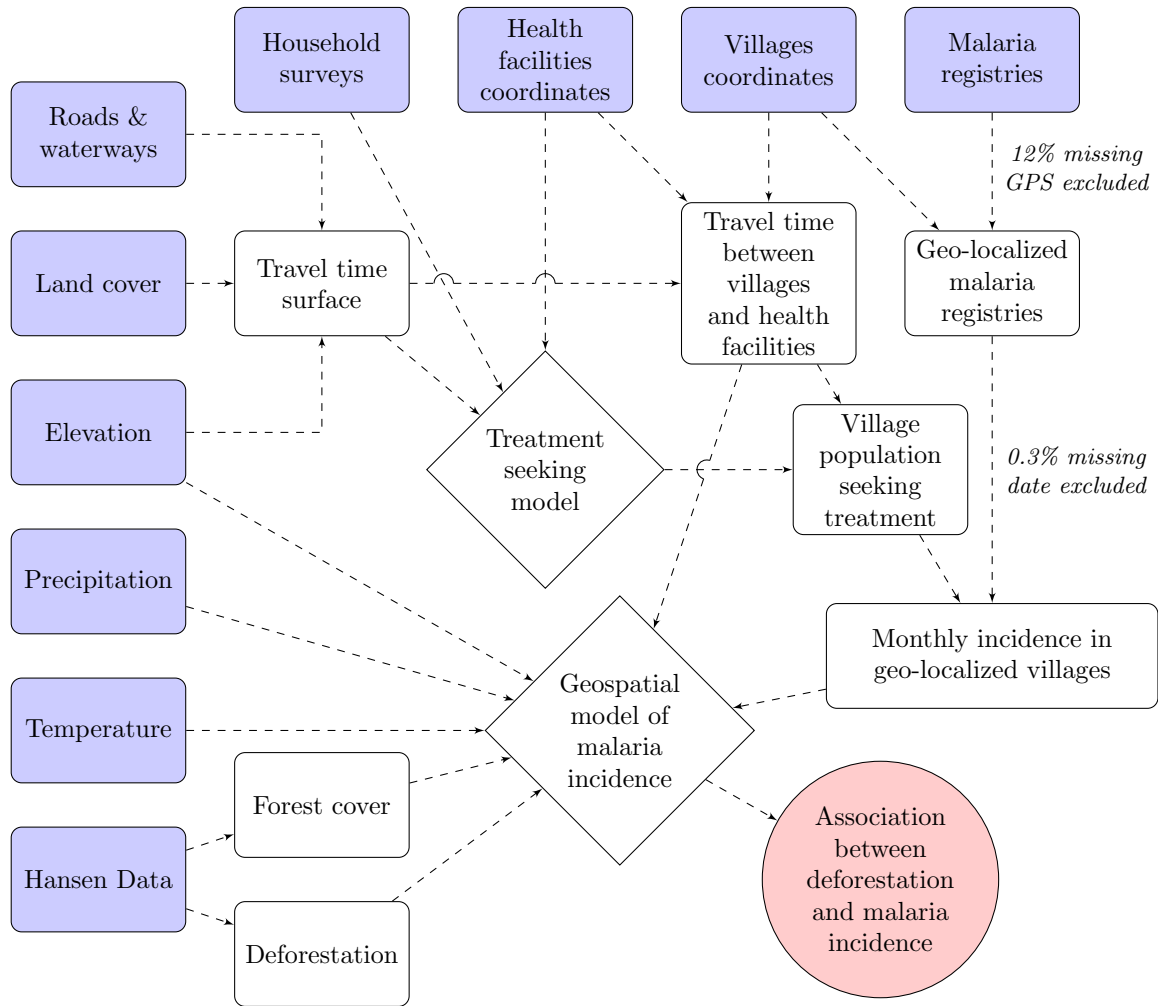
### ***1.5.5.3. Details on covariates***

To prevent collinearity in the final model, for each of the three monthly climatic variables (precipitation, day and night temperature), we first selected the one of its seven "short-term" variations (in current month, in previous 1, 2 or 3 months and aggregated over current and previous 1, 2 or 3 months) that provided the best AIC fit in an univariate model, solely adjusted for the spatio-temporal structure of the data ( $f(t)$ ,  $f(\text{Lat}, \text{Long})$  and village random intercepts). See Appendix 1 – S1.6.

Malaria incidence in the previous 1 and 2 months were included in the model. Results in Appendix 1 – S1.4 show this was necessary to fully address temporal autocorrelation and led to a better AIC fit. Different shape of the temporal trend  $f(t)$  were also explored (up to 25 spline knots, auto-regressive, cyclic cubic spline) but none accounted for temporal autocorrelation better.

In a preliminary analysis, before including the deforestation variable, we ran our model in equation 1.2 with forest cover as the primary exposure. We ran 12 models separately in the north

and the south, each varying the buffer radius (1, 10 and 30 km) and temporal scale (0, 1, 2 and 3 year lag) for the forest cover variable. The coefficients of the linear effect for forest cover were extracted and exponentiated to get the incidence rate ratio (IRR) associated with a 1% increase in the average tree crown cover density around villages (Table 1.6 in the appendix). The model including average tree crown cover density within 30 km of villages with no temporal lag provided the smallest AIC value. In the final models with the deforestation variables we therefore included the average tree crown cover density within 30 km of villages in the starting year of the temporal scale for the deforestation variable considered (e.g. 3 year lag in the model with percent area that experienced forest loss in previous 3 years as the deforestation variable) to adjust for baseline forest cover.



**Figure 1.8** - Conceptual model for our analysis showing how the raw input data (blue boxes) were combined via intermediate data (white boxes) and models (white diamonds) to produce our estimated outputs (red circle).

## **1.6. Appendix 1**

### **1.6.1. S1.1: Travel times methods**

To calculate the travel time along a path linking any two points of the map, we defined a transition matrix that gives the speed at which one may travel between two adjacent pixels. We followed the parameterization suggested by Alegana et al.<sup>79</sup> and demonstrated by Sturrock et al.<sup>80</sup> (see Table 1.4), which first uses Toblers' hiking function to specify the travel speed between two points of different altitudes. Intuitively, it is faster to travel downhill than uphill. Second, the speed is adjusted based on the type of landcover travelled through: a forested or a flooded area for instance slows you down. Last, the network of roads and major rivers may be used to catch a bus or a boat and therefore increases the travel speed. Altitude (SRTM 90m<sup>75</sup>) was aggregated and resampled at the land cover (ESA GlobCover 2009 Project<sup>81</sup>) 300m-resolution and roads and waterways from Open Street Map<sup>82</sup> were rasterized to calculate the transition matrix all across Lao PDR. The raster package in R<sup>83</sup> was used.

We then used the Djisktra's algorithm from the R-package igraph<sup>84</sup> and the gdistance package<sup>85</sup> to find the fastest route between every village (or every household in the cross-sectional surveys) and its closest health facility. Coordinates of health facilities across Lao PDR came from the 2017 stratification exercise and were provided by CMPE. We authorized travel through non-study districts but not across international borders.



**Table 1.4** - Data used to parameterize the transition matrix with the travel speed between any 2 adjacent pixels of the map.

Data layer	Category	Speed (km/h)
Digital elevation	0° (flat)	5
	5° (uphill)	3.71
	-5° (downhill)	5.27
Land cover	Cropland	No adjustment
	Artificial and bare areas	No adjustment
	Open deciduous forest	0.8 * hiking speed
	Sparse herbaceous	0.8 * hiking speed
	Closed deciduous forest	0.6 * hiking speed
	Herbaceous	0.6 * hiking speed
	Flooded	0.5 * hiking speed
	Other forest cover	0.4 * hiking speed
	Water	0.2 * hiking speed
Roads and rivers	Motorway/trunk	80
	Primary/secondary	60
	Tertiary/unclassified	10
	Major rivers	5

## **1.6.2. S1.2: Travel times and treatment-seeking results**

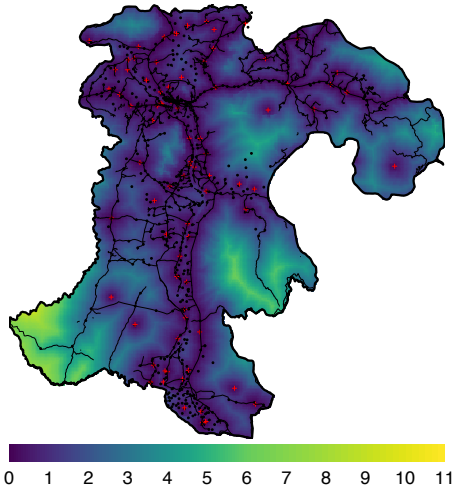
Figure 1.9a shows how the travel time to closest health facility varies across Champasak province in southern Lao PDR, influenced by both distance and road connectivity. Figure 1.9c presents a right-skewed distribution of travel time from study villages to the closest health facility. Most villages are within 2 hours of the closest health facility but some are as far as 6 hours away. The distribution is similar for villages in the northern and southern study districts.

In the southern household survey, 243 individuals reported fever in the past 2 weeks. 225 (92.6%) of them, from 156 households, answered whether or not they sought treatment and were included in the treatment-seeking model. 219 (97.3%) reported seeking treatment and they all reported where they did so: 154 (70.3%) of them sought treatment at a public health facility (Village malaria worker (VMW), health center, district hospital or provincial hospital) and would therefore appear in the malaria registries collected.

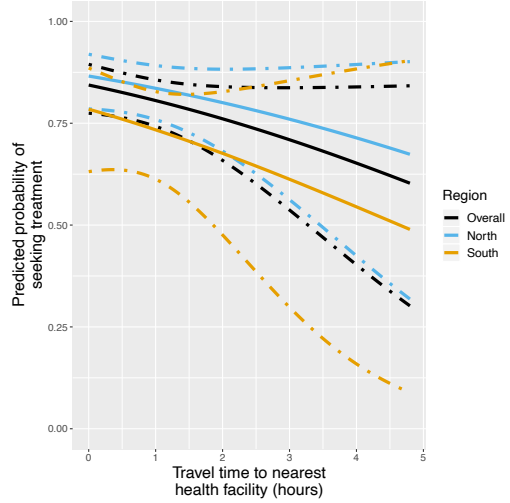
In the northern household survey, 378 individuals reported fever in the past 2 weeks. 360 (95.2%) of them, from 297 households, answered whether or not they sought treatment and were included in the treatment-seeking model. 283 (78.6%) reported seeking treatment. Only 40 (14.1%) of them reported where they did so but all of them sought treatment at a public health facility and we therefore upweighted the population that sought treatment at a public health facility accordingly.

Most surveyed households included in the treatment-seeking model were within 2 hours of travel time to the closest health facility but some were almost 5 hours away (Fig. 1.20 in the appendix).

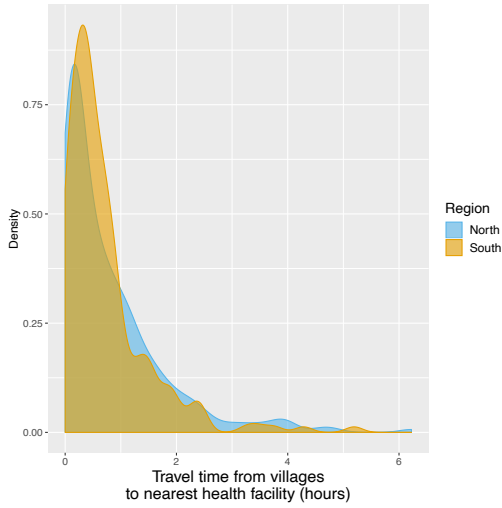
Figure 1.9b shows the modeled relationship between the probability of seeking treatment (at a public health facility, implied from now on) and distance to the closest health facility. For villages within the same 300m<sup>2</sup> pixel as a health facility (estimated travel time of 0 hour), the predicted probability of seeking treatment was 0.87 (95% CI: [0.79; 0.92]) in the north and 0.78 (95% CI: [0.63; 0.89]) in the south. A 1 hour increase in travel time to the closest health facility was associated with a similar 0.79 (95% CI: [0.55; 1.13]) reduction in the odds of seeking treatment in the north and 0.76 (95% CI: [0.43; 1.34]) in the south, almost reaching statistical significance when pooling data from both regions: 0.77 (95% CI: [0.56; 1.04]). Figure 1.9d shows the resulting distribution for the probability of seeking treatment for all villages in study's districts. Monthly village-level malaria incidence was adjusted accordingly.



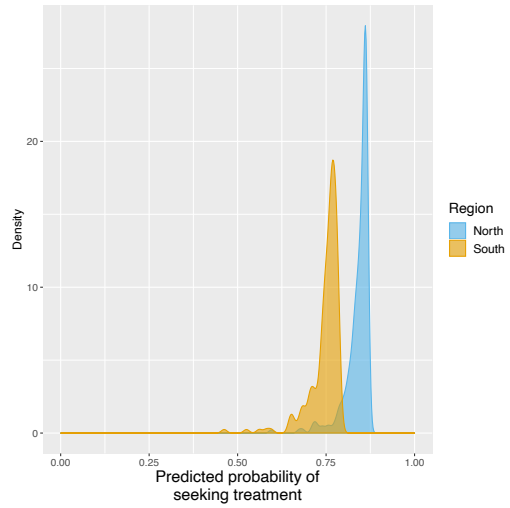
(a) Travel time (hours) to closest health facility (red crosses) in Champasak province, southern Lao PDR. Black dots represent villages and lines show main roads that may be used to travel.



(b) Modeled relationship between treatment-seeking probability and travel time to closest health facility. Dashed lines represent the 95% confidence boundaries.



(c) Distribution of travel time (in hours) from villages to closest health facilities.



(d) Distribution of the predicted probability of seeking treatment.

**Figure 1.9 - Treatment-seeking modeling plots.** Note that treatment-seeking at public health facilities is implied all along the manuscript.

### 1.6.3. S1.3: Sensitivity Analysis

We conducted a sensitivity analysis where village population at risk of appearing in the surveillance system registries were not adjusted for the probability of seeking treatment. The effect estimates and confidence intervals were virtually unchanged, strengthening the robustness of our primary analysis (Table 1.5 below).

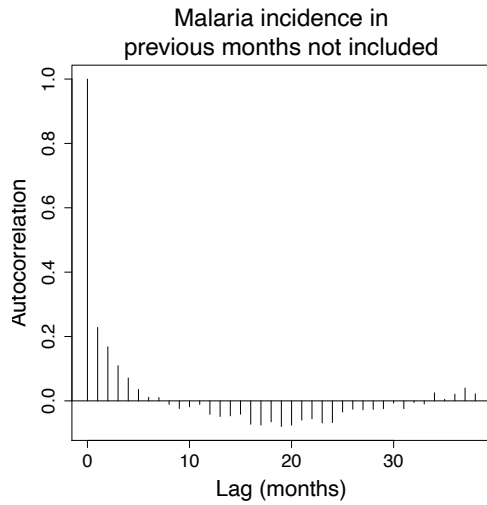
**Table 1.5** - IRR associated with a 0.1% increase in forest loss. Adjusted for the spatio-temporal structure of the data, the environmental covariates selected in the model and forest cover within 30 km in the year before the deforestation temporal scale considered and malaria incidence in the previous 1 and 2 months. See Methods for details. Sensitivity analysis: village population unadjusted for probability of seeking treatment.

Time lag	South			North		
	1 km	Buffer radius 10 km	30 km	1 km	Buffer radius 10 km	30 km
Previous 1 year	1 [0.99; 1.01]	1.01 [0.99; 1.04]	1.16 [1.10; 1.22]	1 [1; 1.01]	1.03 [1; 1.07]	1.01 [0.94; 1.08]
Previous 2 years	1 [0.99; 1.01]	1 [0.98; 1.01]	1.09 [1.04; 1.13]	1 [1; 1.01]	1.01 [0.99; 1.04]	0.98 [0.94; 1.01]
Previous 3 years	0.99 [0.99; 1]	0.98 [0.97; 1]	0.93 [0.90; 0.97]	1 [1; 1.01]	1.01 [0.99; 1.02]	0.96 [0.93; 0.99]
Previous 4 years	0.99 [0.99; 1]	0.98 [0.97; 0.99]	0.94 [0.92; 0.97]	1 [1; 1.01]	1 [0.99; 1.02]	0.97 [0.94; 0.99]
Previous 5 years	0.99 [0.99; 1]	0.97 [0.96; 0.99]	0.94 [0.91; 0.97]	1 [1; 1.01]	1.01 [0.99; 1.02]	0.95 [0.93; 0.98]

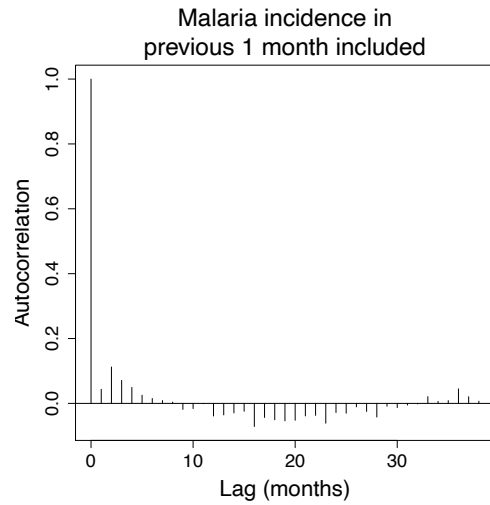
#### **1.6.4. S1.4: Inclusion of malaria cases in previous months**

Figure 1.10 shows residual temporal auto-correlation plots in models from equation 1.2, when malaria incidence in previous 1 and 2 months are included or not. These plots show that including covariates for malaria incidence in the previous 1 and 2 months is necessary to address residual temporal autocorrelation and keep each lag-wise individual autocorrelation estimate below 5 %. The plots presented here are for the model in the south with a 30 km buffer radius and a 1 year temporal lag but similar results were observed across all 15 models both in the north and in the south.

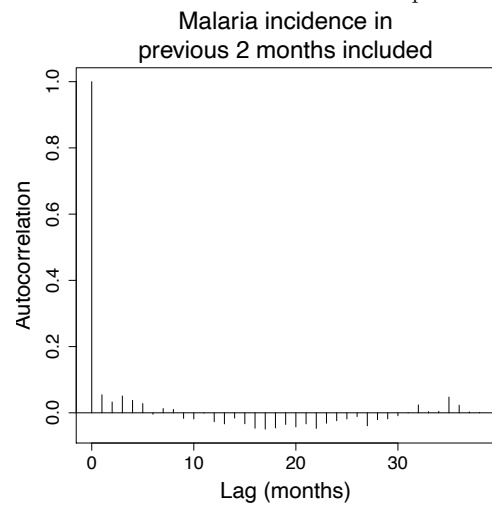
The AIC fit also substantially improved from 18667 when no malaria incidence is included to 18152 when malaria incidence in previous month is included and to 17687 when both malaria in the previous 1 and 2 months are included.



(a) Residual temporal autocorrelation when malaria incidence in previous months is not included.



(b) Residual temporal autocorrelation when malaria incidence in previous 1 month is included but not in previous 2 months.



(c) Residual temporal autocorrelation when malaria incidence in previous 1 and 2 months are included.

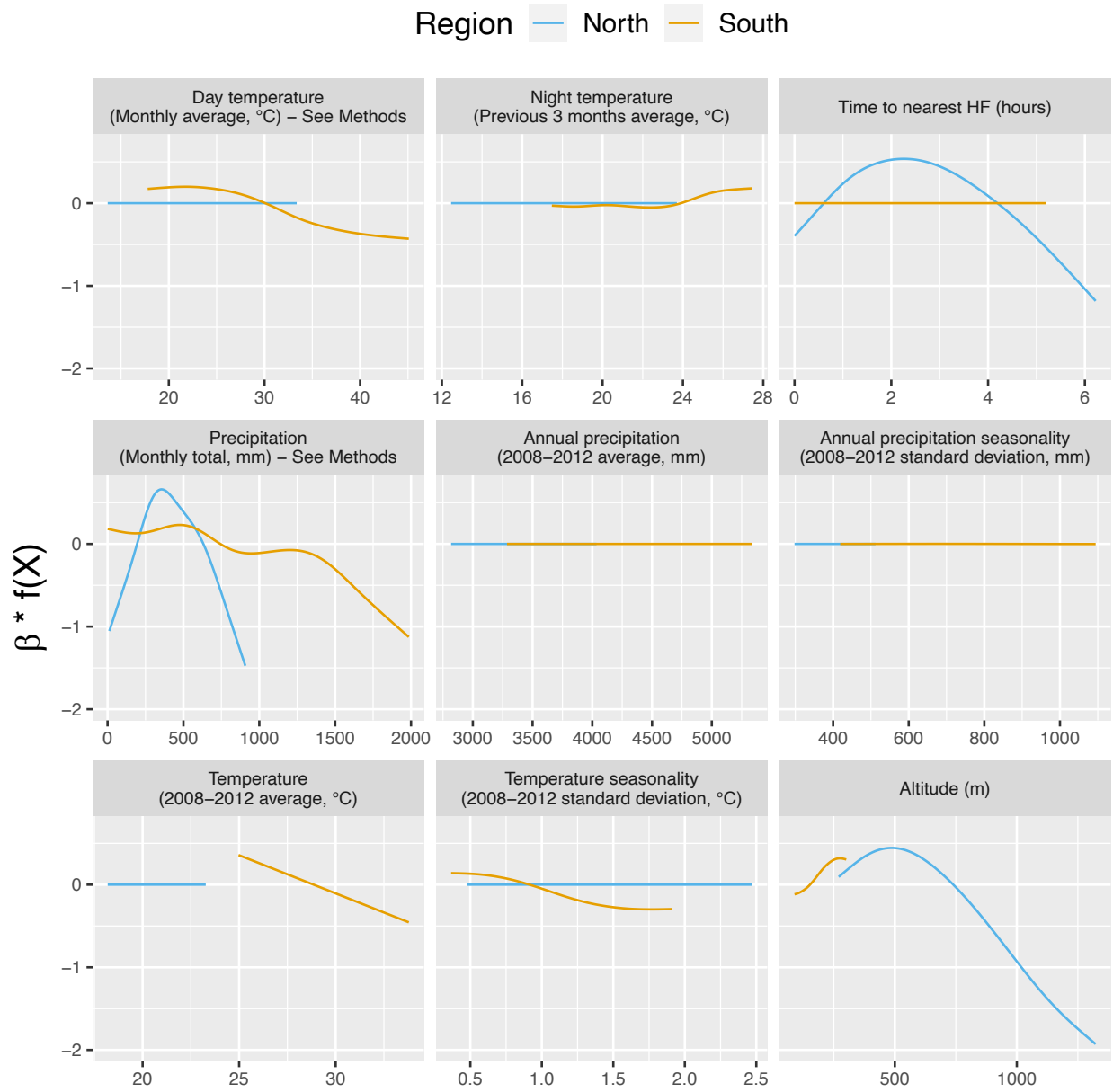
**Figure 1.10** - Residual temporal autocorrelation when malaria incidence in previous 1 and 2 months are included or not.

## 1.6.5. S1.5: Additional Results

### 1.6.5.1. *Environmental covariates*

Figure 1.11 shows the relationship - via their individual contribution  $\beta \times f(X)$  in equation 1.2 - between malaria incidence and the environmental covariates included in the model (30 km radius and 1 year temporal lag). These plots show that relationships differ slightly by region although the range covered by the environmental variables also differs by region. We also see the effect of regularization, that penalized some covariates to zero, like our long-term precipitation covariates. This penalization happened more frequently in the north, where we had much less data. Note that 95% confidence intervals (see Fig. 1.21 in the appendix) have been hidden for better visualization. The larger amount of data in the south also allowed the identification of more precise relationships than in the north.

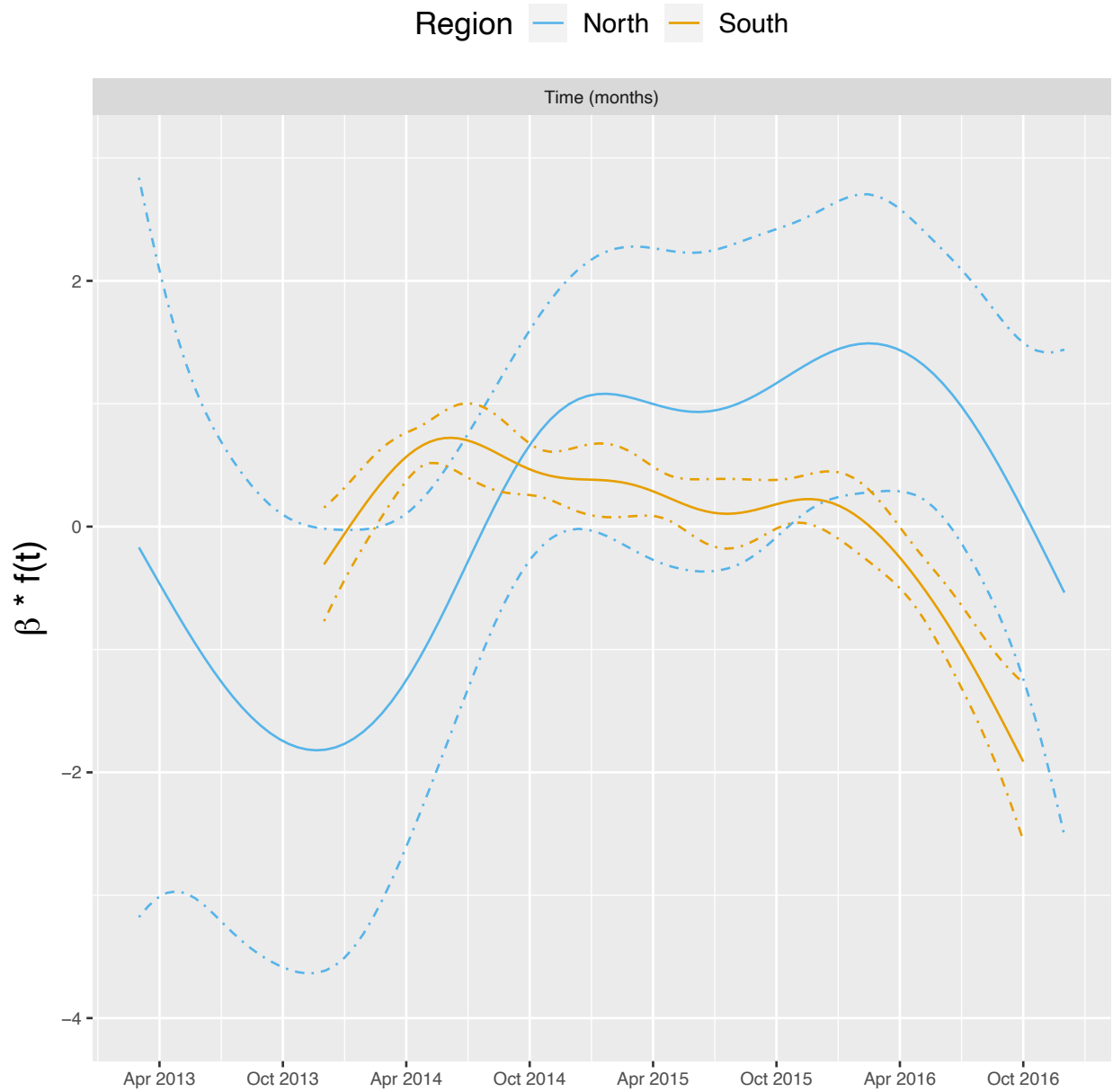




**Figure 1.11** - Relationships between malaria incidence and the environmental covariates in the multivariable model described in equation 1.2 (30 km radius and 1 year temporal lag), additionally adjusted for the probability of seeking treatment, the spatio-temporal structure of the data ( $f(t)$ ,  $f(\text{Lat}, \text{Long})$  and village random intercepts) and malaria incidence in the previous 1 and 2 months. See Methods for details. Note that 95% confidence intervals (see Fig. 1.21 in the appendix) have been hidden for better visualization.

### ***1.6.5.2. Temporal trend***

Figure 1.12 shows the relationship - via its individual contribution  $\beta \times f(X)$  in equation 1.2 - between malaria incidence and the temporal trend included in the model (30 km radius and 1 year temporal lag). These plots show that relationships are quite similar in both regions with an increase in 2014, followed by a plateau in 2015 and a decrease in 2016. The larger amount of data in the south also allowed the identification of a more precise relationship than in the north.



**Figure 1.12** - Relationships between malaria incidence and the temporal trend in the multivariable model described in equation 1.2 (30 km radius and 1 year temporal lag), additionally adjusted for the probability of seeking treatment, the spatial structure of the data ( $f(\text{Lat}, \text{Long})$  and village random intercepts) and malaria incidence in the previous 1 and 2 months. See Methods for details.

### **1.6.5.3. Forest cover**

Table 1.6 shows the incidence rate ratio (IRR) associated with forest cover, measured by a 1% increase in the average tree crown density, in current and previous 3 years within 1, 10 and 30 km of villages.

Forest cover within 1 km of a village was not associated with malaria incidence rate in either the south or the north, regardless of the temporal lag. However within 10 and 30 km of a village, increased forest cover tended to be associated with higher malaria incidence rates both in the north and the south (e.g. 30 km buffer, 1-year lag, IRR = 1.09, 95% CI: [1.03; 1.15] in the south; IRR = 1.12, 95% CI: [0.99; 1.26] in the north). The associations were higher when considering a larger spatial scale (30 km) but were already statistically significant for a 10 km buffer radius in the south. None of the associations reached statistical significance in the north, where the sample size is small. The temporal scale considered did not affect the associations much.

Statistical significance wasn't necessarily reached for all the associations highlighted, but the trends observed suggest forest cover around villages but not in the immediate vicinity (1 km) leads to higher risk of malaria both in the north and in the south.

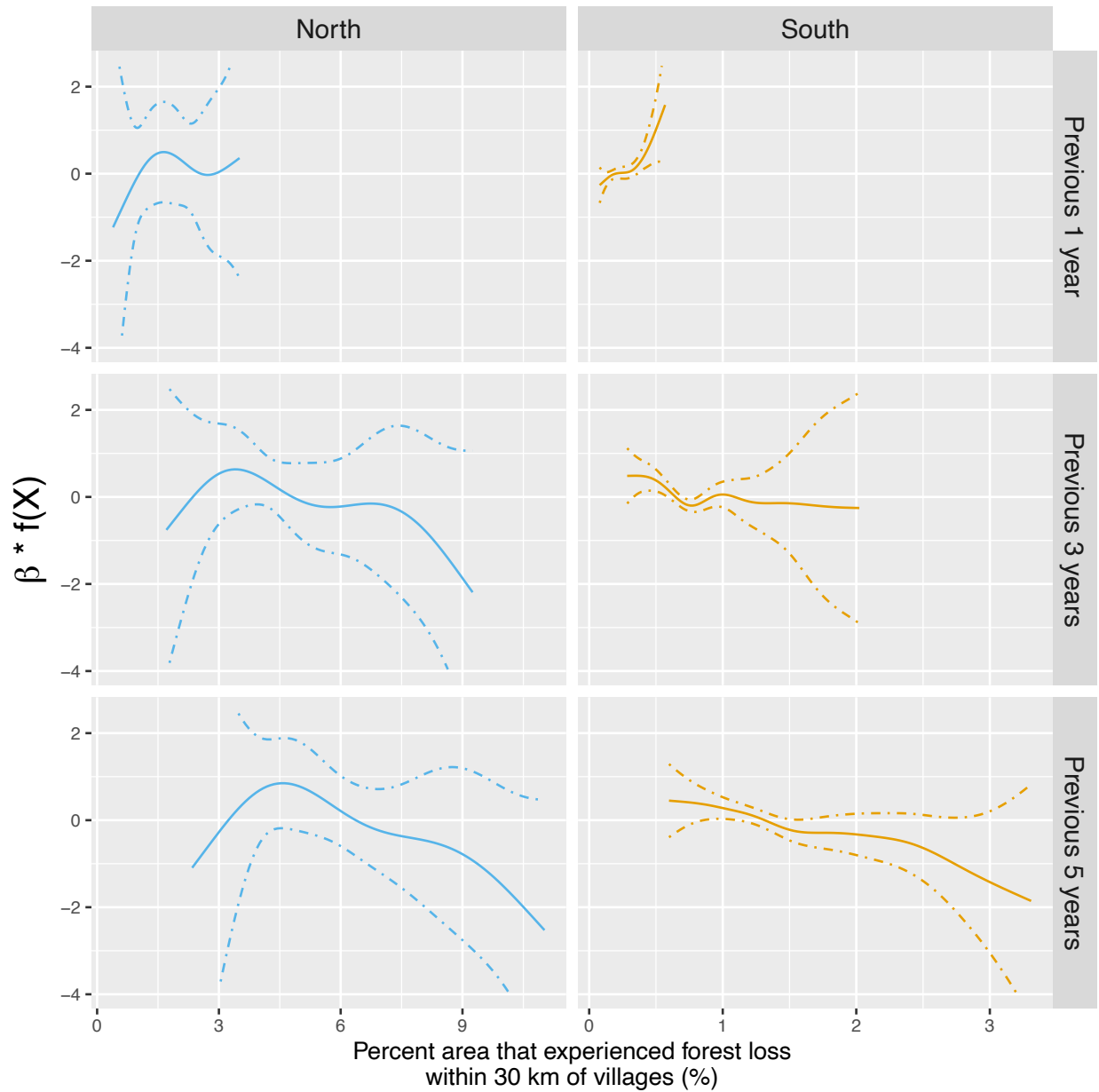
**Table 1.6** - IRR [95% CI] associated with a 1% increase in average tree crown density. Adjusted for the probability of seeking treatment, the spatio-temporal structure of the data, the environmental covariates selected in the model and malaria incidence in the previous 1 and 2 months. See Methods for details.

Time lag	South			North		
	1 km	10 km	30 km	1 km	10 km	30 km
Current year	1 [0.99; 1.01]	1.07 [1.04; 1.10]	1.06 [1; 1.12]	0.99 [0.97; 1.02]	1.01 [0.96; 1.05]	1.10 [0.99; 1.23]
Previous 1 year	1 [0.99; 1.02]	1.07 [1.05; 1.10]	1.09 [1.03; 1.15]	1 [0.97; 1.02]	1.01 [0.97; 1.06]	1.12 [0.99; 1.26]
Previous 2 years	1 [0.99; 1.02]	1.07 [1.05; 1.10]	1.09 [1.03; 1.16]	1 [0.98; 1.03]	1.02 [0.97; 1.06]	1.10 [0.98; 1.25]
Previous 3 years	1 [0.99; 1.02]	1.07 [1.04; 1.10]	1.10 [1.04; 1.16]	1.01 [0.98; 1.03]	1.02 [0.97; 1.07]	1.10 [0.98; 1.24]

The model including average tree crown cover density within 30 km of villages with no temporal lag provided the best AIC. In the final models with the deforestation variables we therefore included the average tree crown cover density within 30 km of villages in the starting year of the temporal scale for the deforestation variable considered (e.g. 3 year lag in the model with percent area that experienced forest loss in previous 3 years as the deforestation variable).

#### ***1.6.5.4. Deforestation - Non linearities***

The IRR effect estimates in Table 1.1 and Figure 1.3 assume a linear relationship between deforestation and malaria. Figure 1.13 shows a few of these relationships - via their individual contribution  $\beta \times f(X)$  in equation 1.2 - when such linearity isn't imposed in the GAM models. Although the AIC fit is slightly better when modeling non-linearities, these plots show that the linearity assumption is mostly warranted.



**Figure 1.13** - Adjusted relationship between deforestation and malaria incidence. All models were adjusted for environmental covariates and forest cover on top of the probability of seeking treatment, the spatio-temporal structure of the data ( $f(t)$ ,  $f(\text{Lat}, \text{Long})$  and village random intercepts) and malaria incidence in the previous 1 and 2 months. See Methods for details. Note that scales are different between buffer radius for better visualization. Figure 1.22 in the appendix shows the raw scatterplot between monthly village malaria incidence rate and deforestation. Figures 1.23 and 1.24 in the appendix show the raw time series of malaria incidence, forest cover and percent area that experienced forest loss.

### **1.6.6. S1.6: AIC fit of the seven monthly climatic variables variations**

To avoid collinearity, we have selected (based on the best AIC fit) the 1 of the 7 variations (In current month, in previous 1, 2 or 3 months and aggregated over previous 1, 2 or 3 months) of the three monthly climatic variables (Precipitation, Day temperature and Night temperature) to be included in the final model. This was done independently for each of the four outcome models (South, North, South *Pf* and South *Pv*).

For South *P. falciparum*, the second best fitting AIC day temperature (in current month) was selected (rather than in previous month) because of very similar AIC fits and to ensure better comparability with the overall South and South *P. vivax* models.



**Table 1.7** - AIC fit of univariate models when including each of the seven monthly climatic variation one at a time as unique covariate in equation 1.2 solely adjusted for the probability of seeking treatment, the spatio-temporal structure of the data ( $f(t)$ ,  $f(\text{Lat}, \text{Long})$  and village random intercepts). AIC selected are in bold.

	Outcome model			
	South	North	South <i>P. falciparum</i>	South <i>P. vivax</i>
<b>Day temperature</b>				
Current month	<b>18546</b>	1671	<b>13226</b>	<b>14575</b>
Previous month	18556	1702	13224	14590
2 months ago	18578	<b>1669</b>	13249	14594
3 months ago	18559	1672	13232	14593
Over current and previous month	18556	1670	13231	14583
Over current and previous 2 months	18570	1670	13248	14588
Over current and previous 3 months	18573	1680	13249	14592
<b>Night temperature</b>				
Current month	18413	1669	13120	14474
Previous month	18453	1670	13155	14520
2 months ago	18547	1673	13231	14576
3 months ago	18581	1672	13251	14596
Over current and previous month	18296	1664	13044	14397
Over current and previous 2 months	18263	1669	13014	14385
Over current and previous 3 months	<b>18262</b>	<b>1663</b>	<b>13007</b>	<b>14385</b>
<b>Precipitation</b>				
Current month	18532	1693	13198	14593
Previous month	<b>18520</b>	1669	<b>13181</b>	<b>14575</b>
2 months ago	18538	<b>1658</b>	13207	14594
3 months ago	18579	1664	13243	14596
Over current and previous month	18570	1672	13239	14595
Over current and previous 2 months	18543	1670	13187	14580
Over current and previous 3 months	18555	1674	13212	14591

## 1.6.7. S1.7: Additional figures

This section presents additional figures mentioned in the text and in the additional results section of the appendix.

### 1.6.7.1. Forest and environmental variables

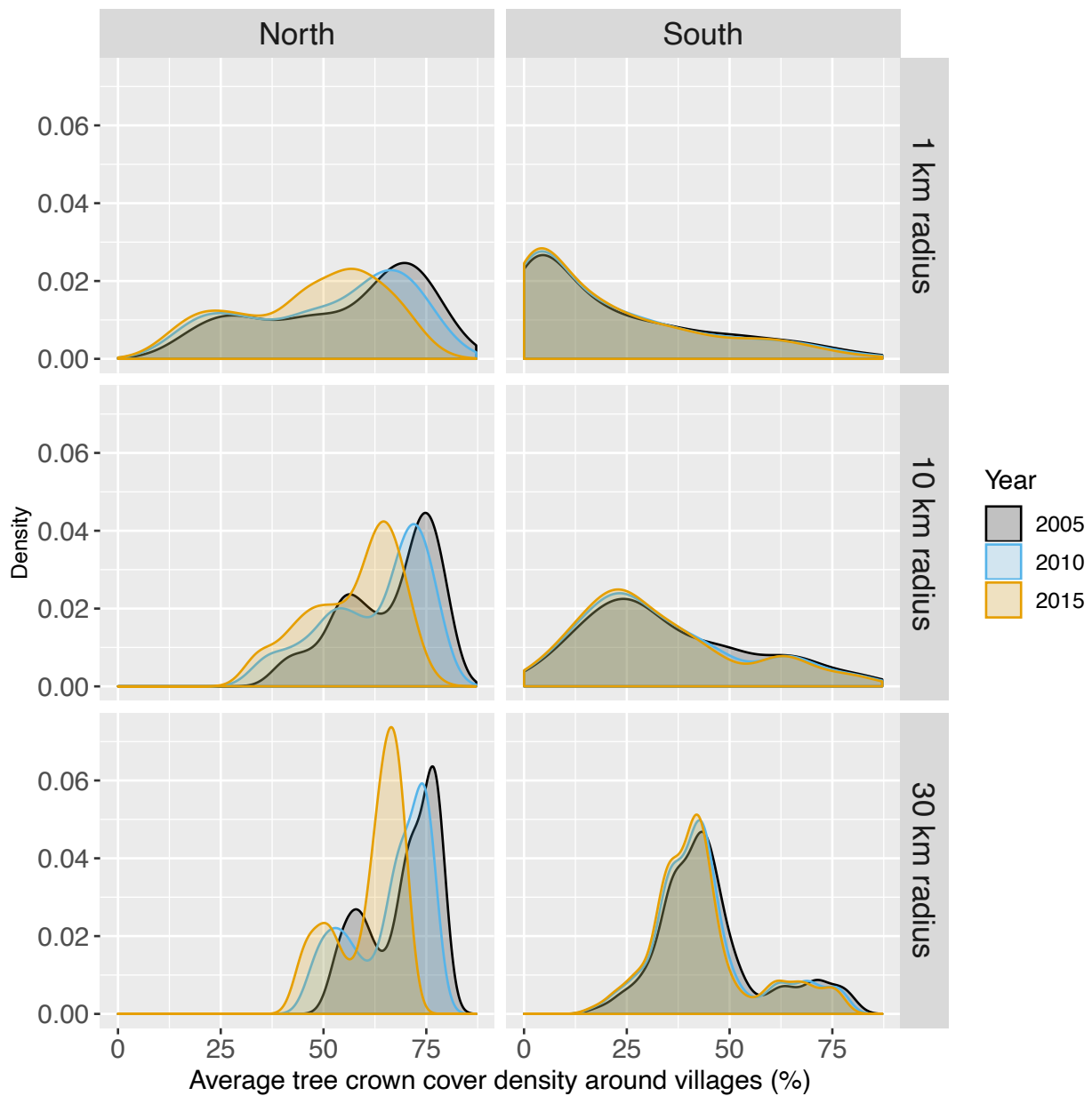
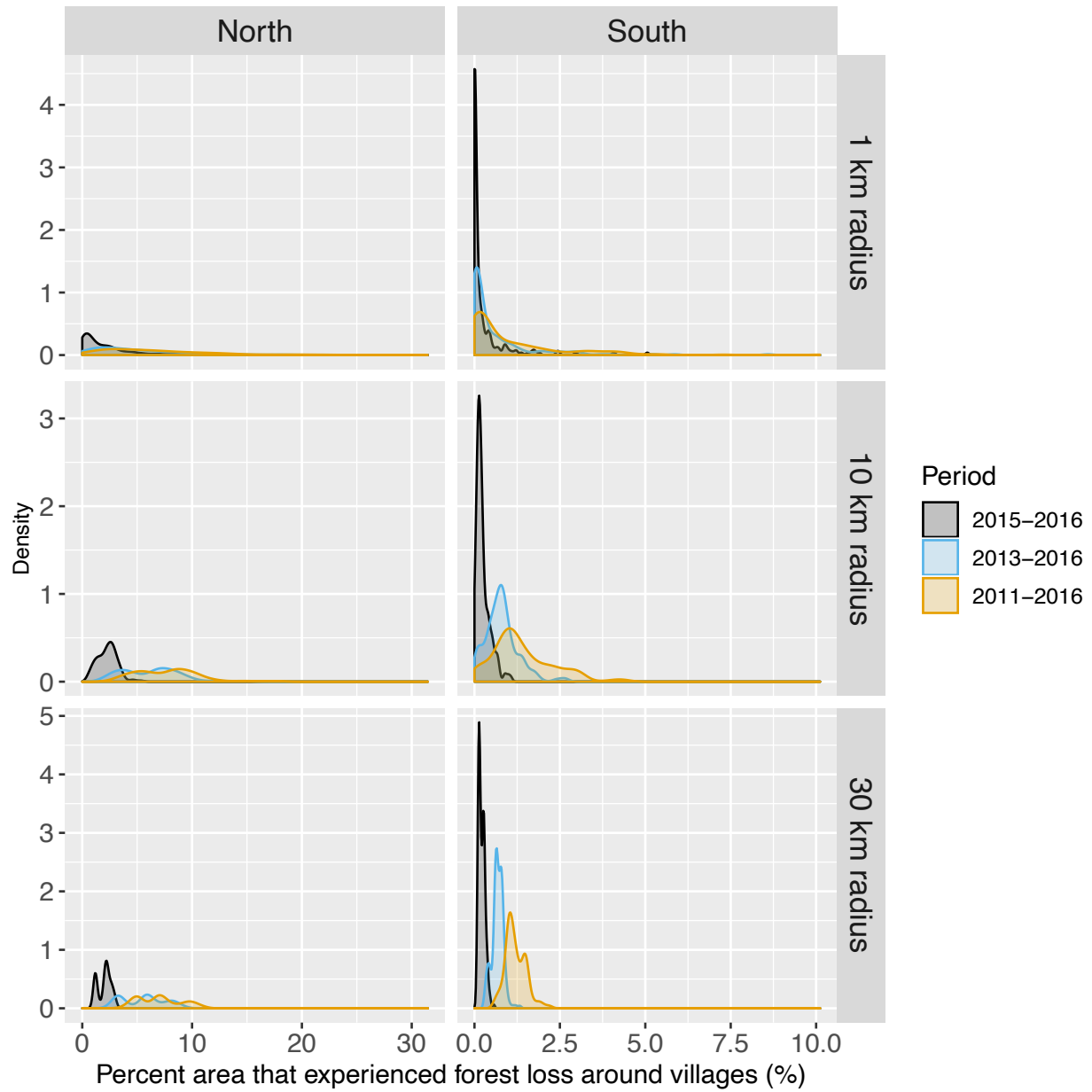
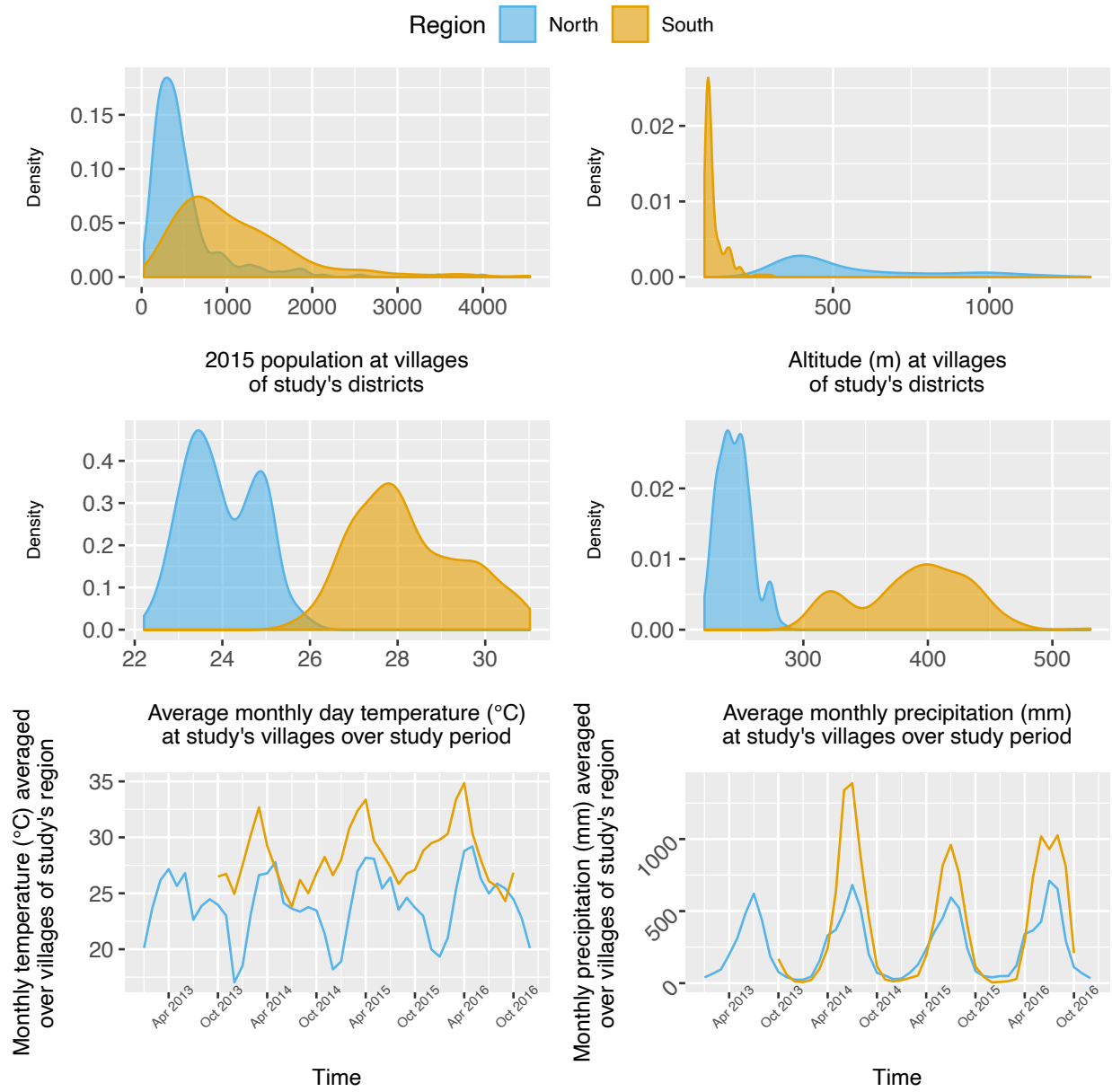


Figure 1.14 - Distribution of average tree crown cover density within 1, 10 and 30 km of villages.

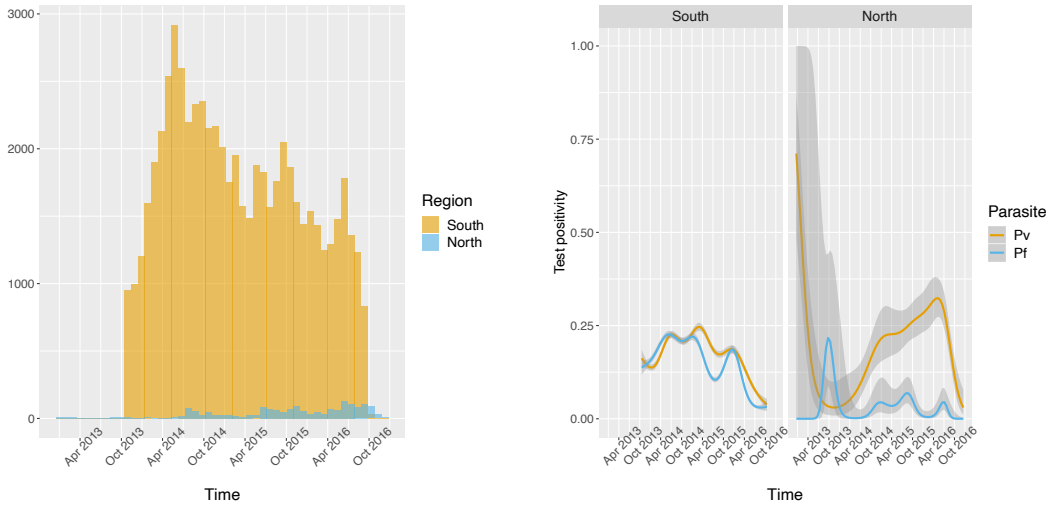


**Figure 1.15** - Distribution of percent area within 1, 10 and 30 km of villages that experienced forest loss between 2011 and 2016. Note that the scales are different for every panel for better visualization of the distributions.



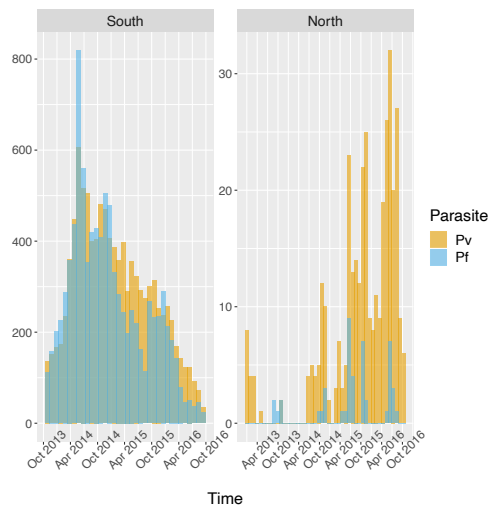
**Figure 1.16** - Distribution and time series of environmental covariates (population, altitude, monthly day temperature and monthly total precipitation) at study's villages.

### 1.6.7.2. Malaria registries - Malaria infections



(a) Number of patients recorded per month in health facilities malaria registries over time.

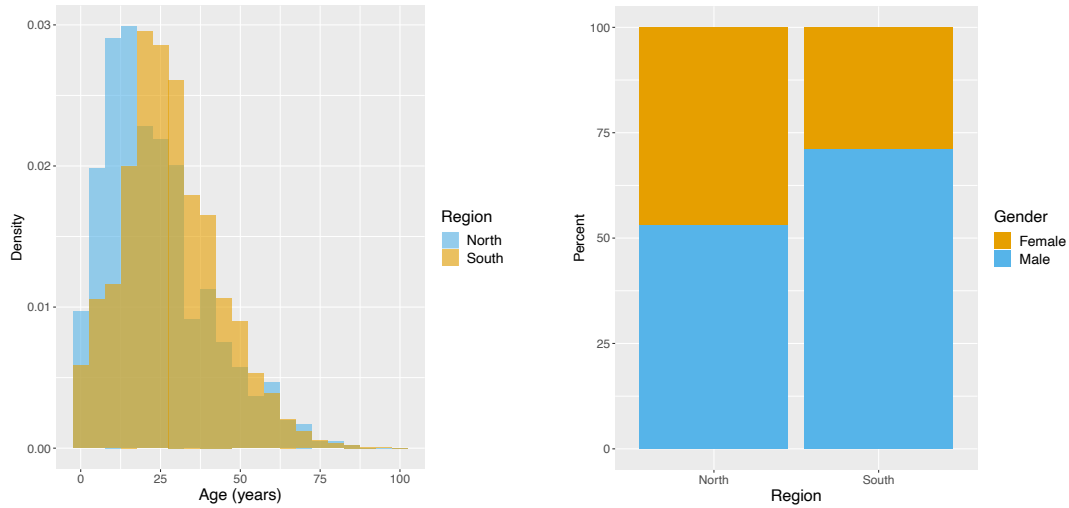
(b) Smoothed (binomial GAM) *P. falciparum* and *P. vivax* test positivity rate over time.



(c) Monthly case count recorded in health facilities malaria registries. Note that scales are different between regions for better visualization.

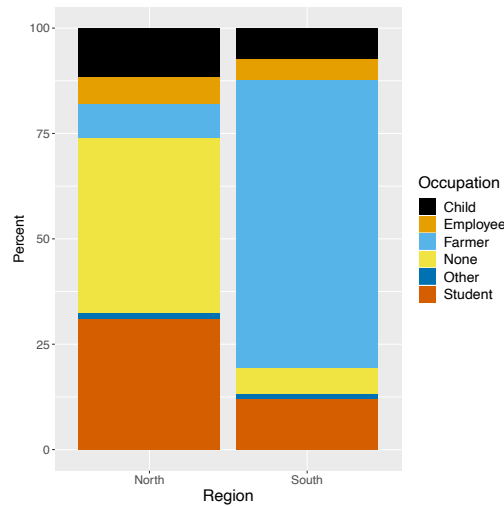
**Figure 1.17 - Additional figures from malaria registries: malaria infections.**

### 1.6.7.3. Malaria registries - SES



(a) Distribution of age (in years) of patients recorded in the malaria registries.

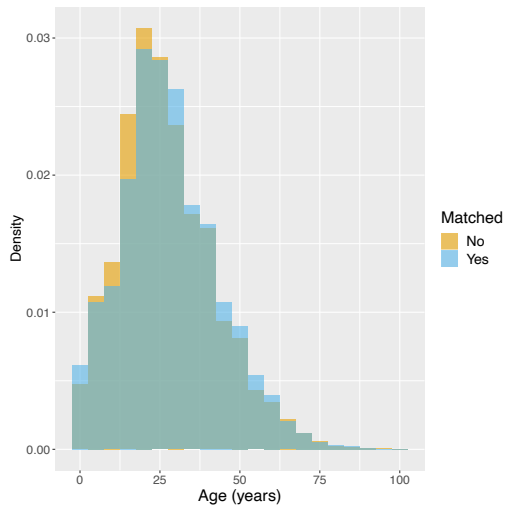
(b) Distribution of gender of patients recorded in the malaria registries.



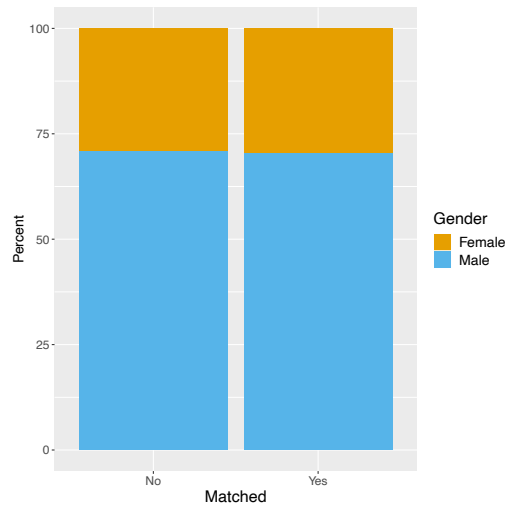
(c) Distribution of occupation of patients recorded in the malaria registries.

**Figure 1.18** - Distributions of socio-economical variables of all patients recorded in the malaria registries.

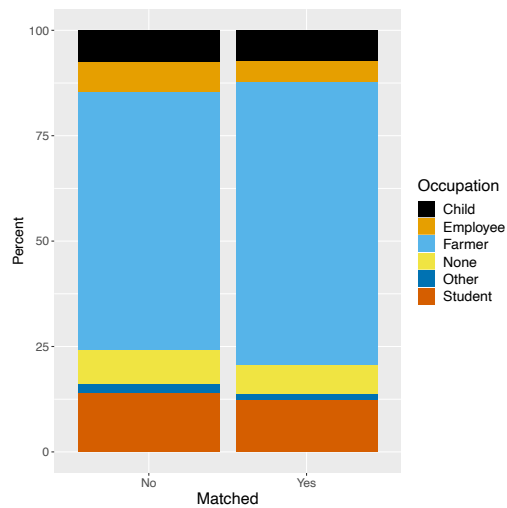
### 1.6.7.4. Malaria registries - Matched vs unmatched



(a) Distribution of age (in years) of patients recorded in the malaria registries.



(b) Distribution of gender of patients recorded in the malaria registries.



(c) Distribution of occupation of patients recorded in the malaria registries.

**Figure 1.19** - Additional figures from malaria registries: matched vs unmatched SES variables.

### 1.6.7.5. Treatment-seeking

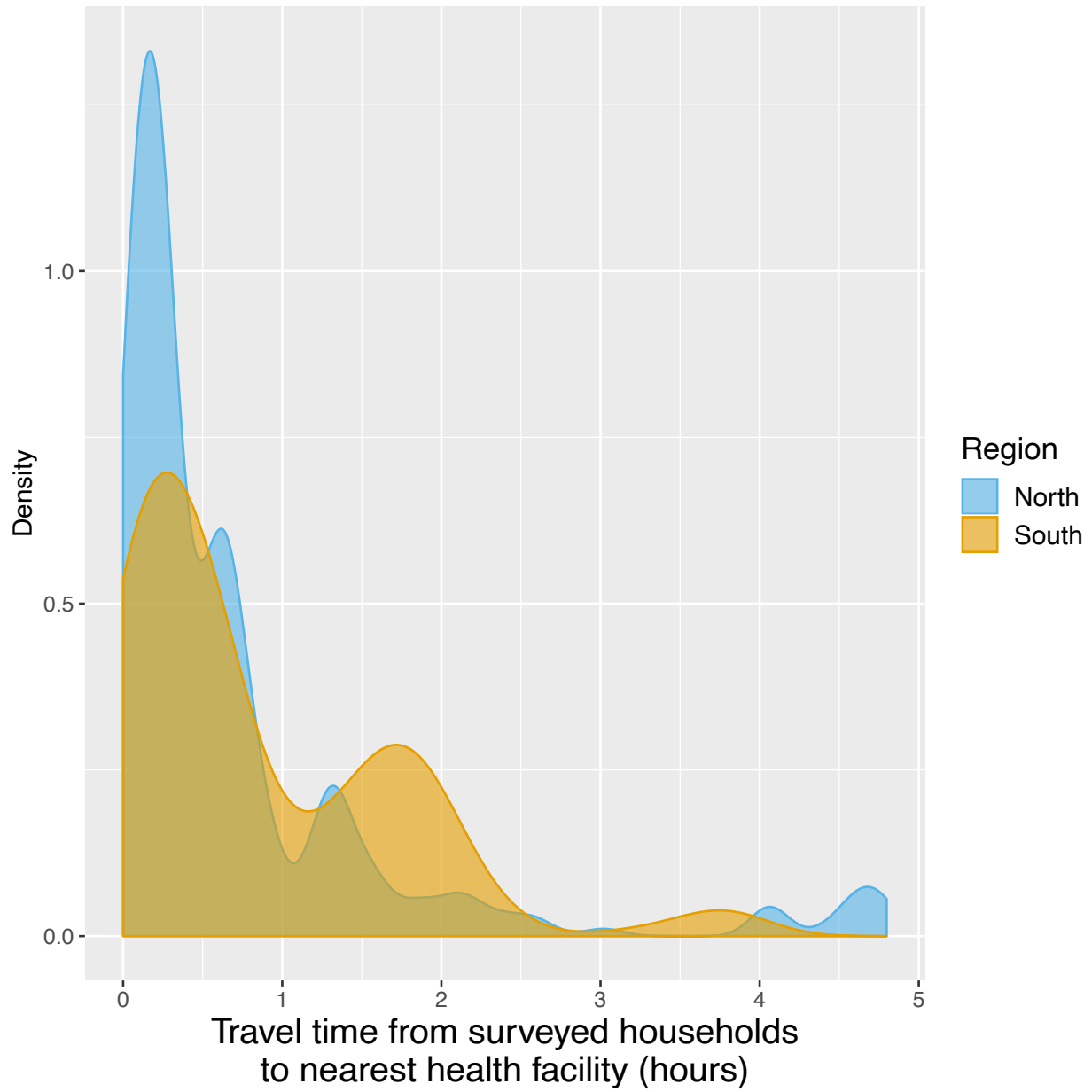
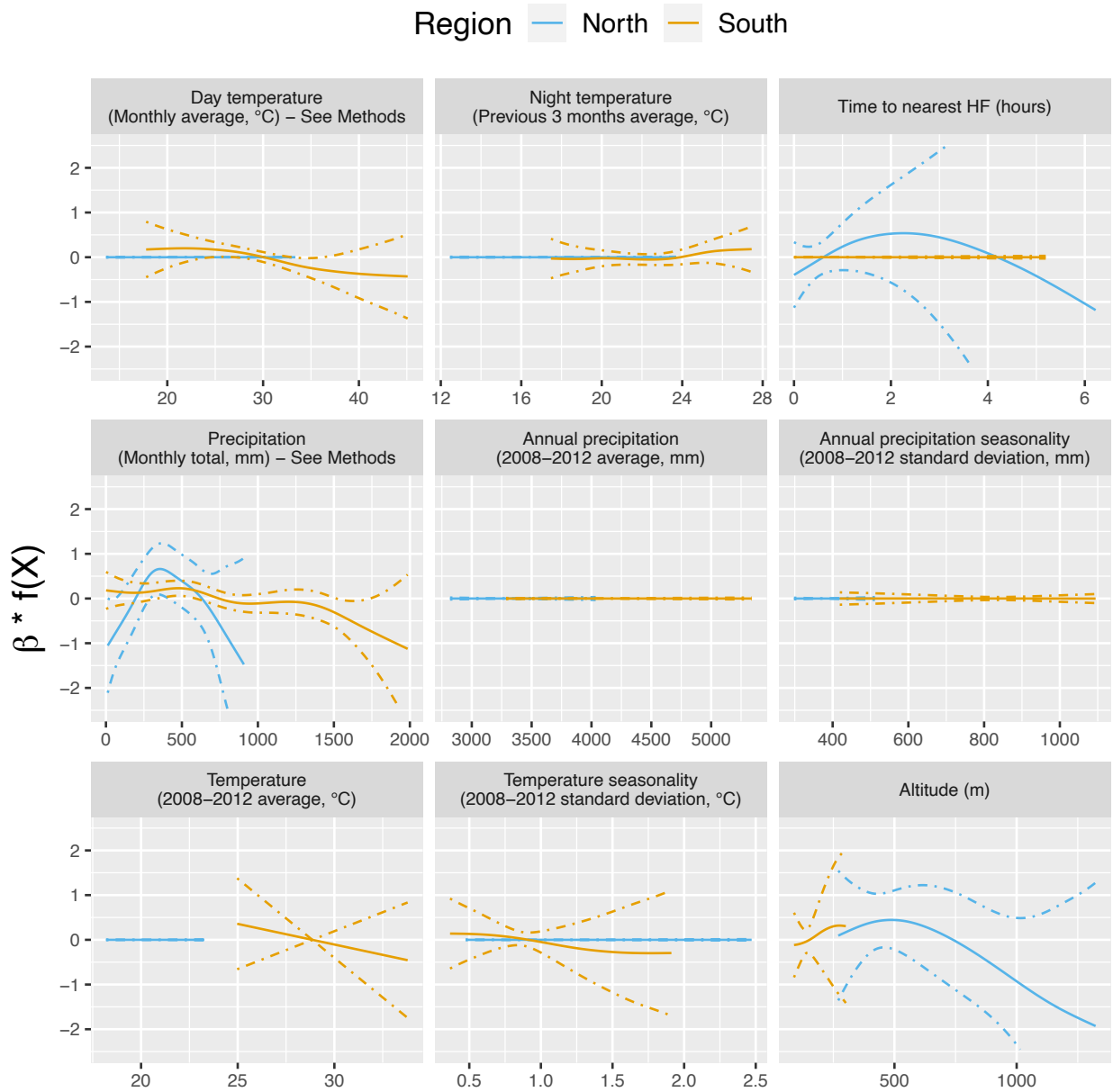


Figure 1.20 - Distribution of travel time (in hours) from surveyed households to closest health facilities.

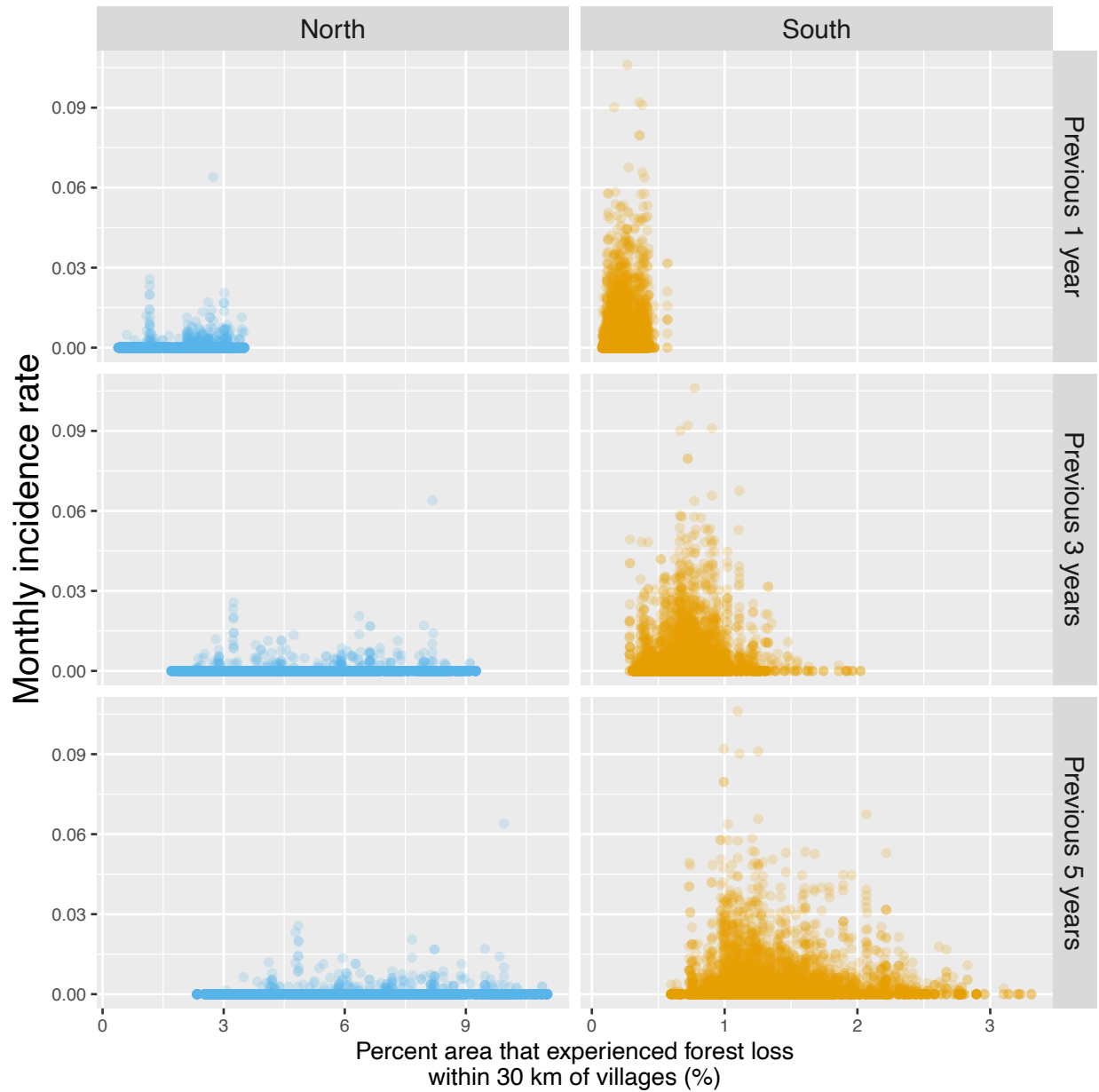


### 1.6.7.6. Statistical analysis - Environmental covariates



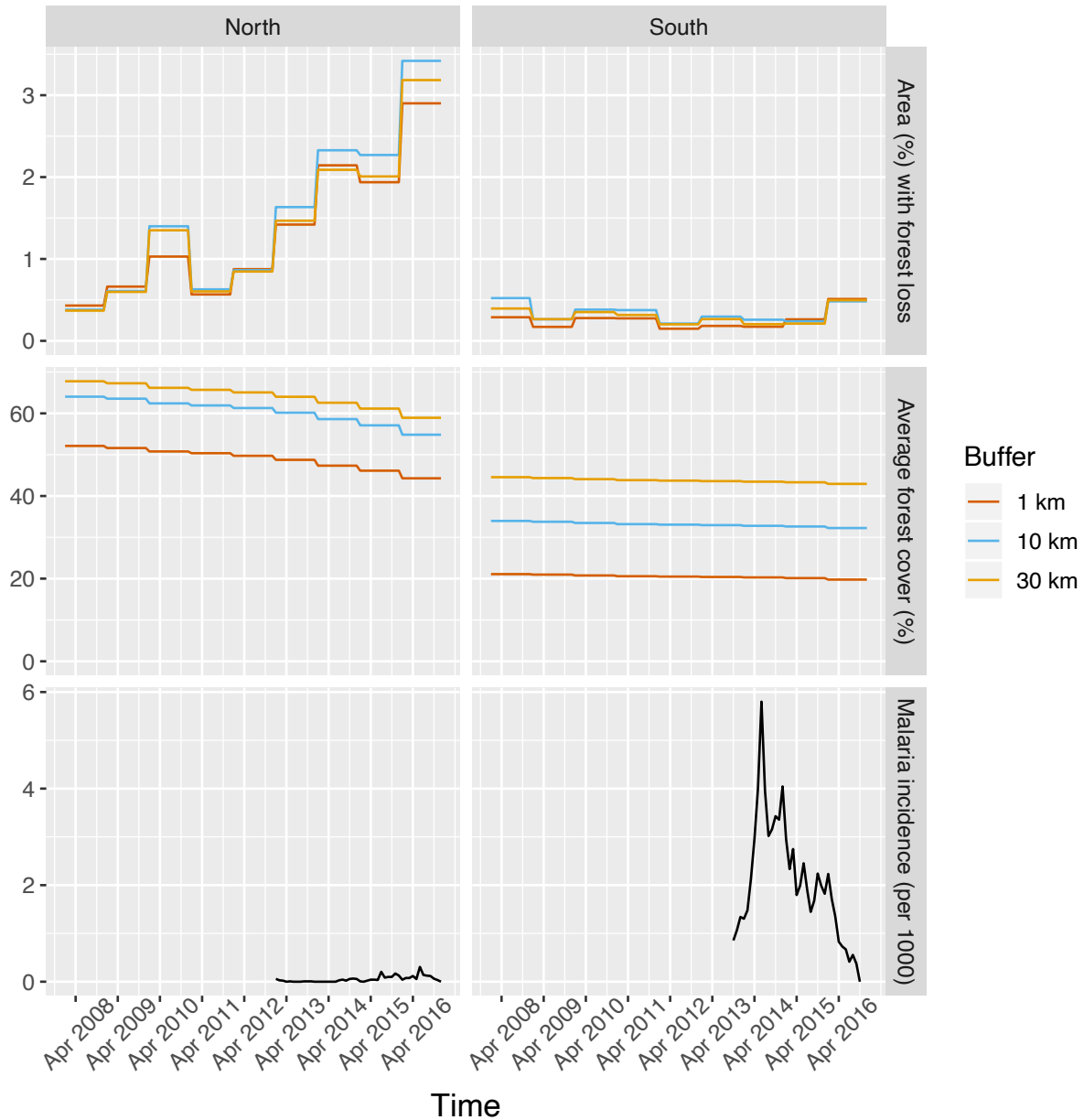
**Figure 1.21** - Relationships between malaria incidence and the environmental covariates in the multivariable model described in equation 1.2 (30 km radius and 1 year temporal lag), additionally adjusted for the probability of seeking treatment, the spatio-temporal structure of the data ( $f(t)$ ,  $f(\text{Lat}, \text{Long})$ ) and village random intercepts) and malaria incidence in the previous 1 and 2 months. Dashed lines are for 95% confidence intervals. Note that the y scale has been trimmed a bit for better visualization.

**1.6.7.7. Raw association between malaria incidence and deforestation**

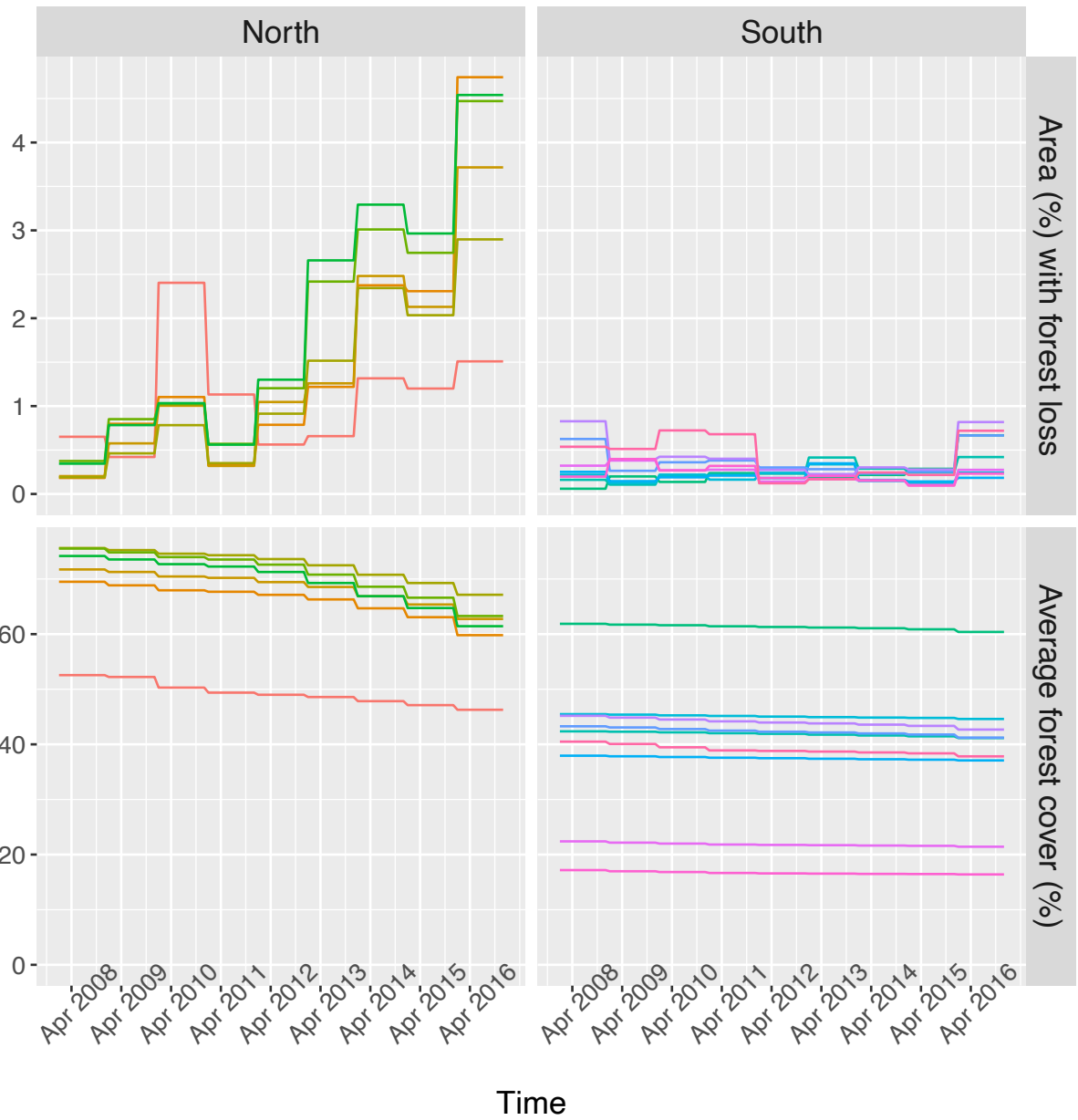


**Figure 1.22** - Raw scatterplot between monthly village malaria incidence rate and the percent area within 30 km of villages that experienced forest loss in the previous 1, 3 and 5 years. Note that scales are different between regions for better visualization.

**1.6.7.8. Raw time series of malaria incidence, forest cover and deforestation**

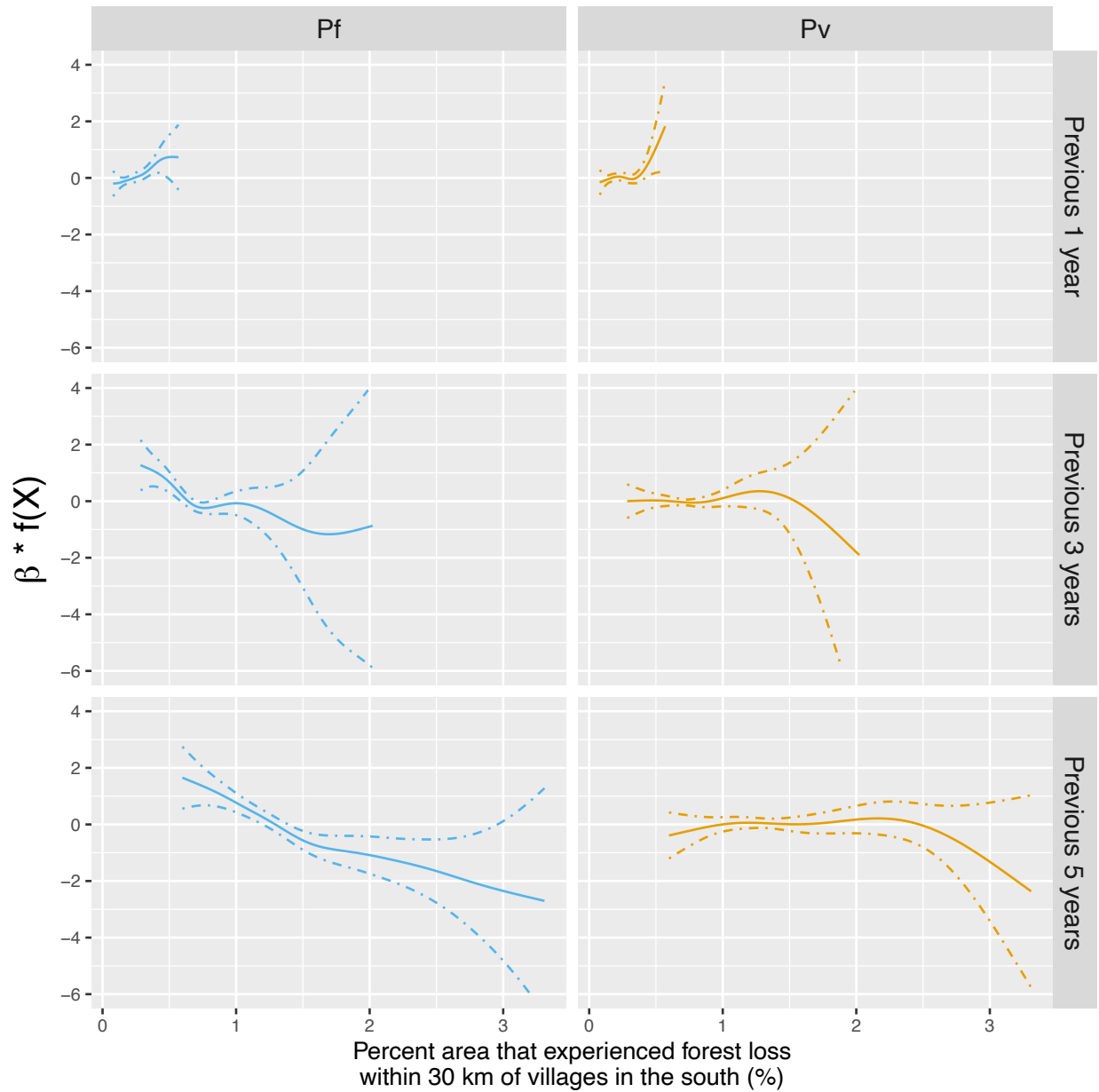


**Figure 1.23** - Time series of deforestation (percent area that experienced forest loss around villages), forest cover (average tree crown cover around villages) and malaria incidence, averaged over study's villages and for varying buffer radius around villages (1, 10 and 30 km).



**Figure 1.24** - Time series of deforestation (percent area that experienced forest loss within 30 km of villages) and forest cover (average tree crown cover within 30 km of villages), for a few randomly sampled study's villages. Each color represents 1 village.

1.6.7.9. Statistical analysis - *P. falciparum* and *P. vivax*



**Figure 1.25** - Adjusted relationship between deforestation and species-specific malaria incidence in southern Lao PDR. All models were adjusted for environmental covariates and forest cover on top of the probability of seeking treatment, the spatio-temporal structure of the data ( $f(t)$ ,  $f(Lat, Long)$  and village random intercepts) and malaria incidence in the previous 1 and 2 months.

## **Chapter 2: Population size estimation of seasonal forest-going populations in southern Lao PDR**

François Rerolle, Jerry O. Jacobson, Paul Wesson, Emily Dantzer,  
Andrew A. Lover, Bouasy Hongvanthong, Jennifer Smith,  
John M. Marshall, Hugh Sturrock, Adam Bennett

## 2.1. Abstract

Forest-going populations are key to malaria transmission in the Greater Mekong Sub-region (GMS) and are therefore targeted for elimination efforts. Estimating the size of this population is essential for programs to assess, track and achieve their 2030 elimination goals.

Leveraging data from three cross-sectional household surveys and one survey among forest-goers, the size of this high-risk population in a southern province of Lao PDR between December 2017 and November 2018 was estimated by two methods: population-based household surveys and capture-recapture.

During the first month of the dry season, the first month of the rainy season, and the last month of the rainy season, respectively, 16.2% [14.7; 17.7], 9.3% [7.2; 11.3], and 5.3% [4.4; 6.1] of the adult population were estimated to have engaged in forest-going activities. The capture-recapture method estimated a total population size of 18,426 [16,529; 20,669] forest-goers, meaning 61.0% [54.2; 67.9] of the adult population had engaged in forest-going activities over the 12-month study period.

This study demonstrates two methods for population size estimation to inform malaria research and programming. The seasonality and turnover within this forest-going population provide unique opportunities and challenges for control programs across the GMS as they work towards malaria elimination.

## 2.2. Introduction

Malaria transmission in the Greater Mekong Sub-region (GMS) is commonly described as “forest malaria”<sup>15</sup>, and is attributed to the dominance of forest-dwelling malaria vectors such as *Anopheles dirus* and *Anopheles minimus*<sup>13,14</sup>. Activities that result in contact with these vectors in the forest, such as logging, hunting or sleeping, and common forest-fringe activities such as farming or “slash and burn” agriculture near forest areas<sup>20,62,86</sup> are major risk factors for malaria in the GMS<sup>16,17,22,25,26,29,87–89</sup>. As malaria declines in the region, transmission often clusters in forest-going populations that are increasingly targeted for prevention and treatment efforts by national control programs across the GMS<sup>23,24</sup>. Yet, the size of this high-risk population (HRP) remains unknown and difficult to quantify.

Estimating the size of HRPs is important for several reasons<sup>90</sup>. Population size estimates (PSE) can be used to inform policies and mobilize support for control and elimination programs. They are essential to determine the required scale of preventive interventions, to assess intervention coverage, parametrize transmission models, and monitor programs.

There are numerous studies as well as international guidelines<sup>90</sup> focusing on size estimation of HRPs for HIV<sup>91–96</sup> but, to our knowledge, none for malaria. In regions where HIV transmission clusters in HRPs, the Second Generation Surveillance (SGS) guidelines for HIV<sup>97</sup> recommend routine PSE<sup>98</sup>. HRPs for HIV such as sex workers or injecting drug users are considered hard to reach populations because of stigma or discrimination and require sophisticated PSE methods. In malaria, HRPs may not be as hidden, although there may be concerns about the illegal nature of large-scale logging in the GMS<sup>99</sup>. PSE methods, originating in animal ecology, can also be used



for non-stigmatized populations and researchers have noted the need for PSE in malaria surveillance where HRPs are key to transmission<sup>100</sup>.

A major difference between forest-going HRPs for malaria in the GMS and HRPs for HIV is the marked seasonality of their high-risk activities. In this region, the monsoon transforms the environment and affects HRPs' forest-going activities. For instance, while the rainy season draws populations to rice fields for agriculture, heavy precipitation may also deteriorate roads so that traveling to the forest often becomes challenging. Therefore, evaluating the population size of forest-going HRPs at different time points is essential to identify the appropriate timing of interventions in the GMS.

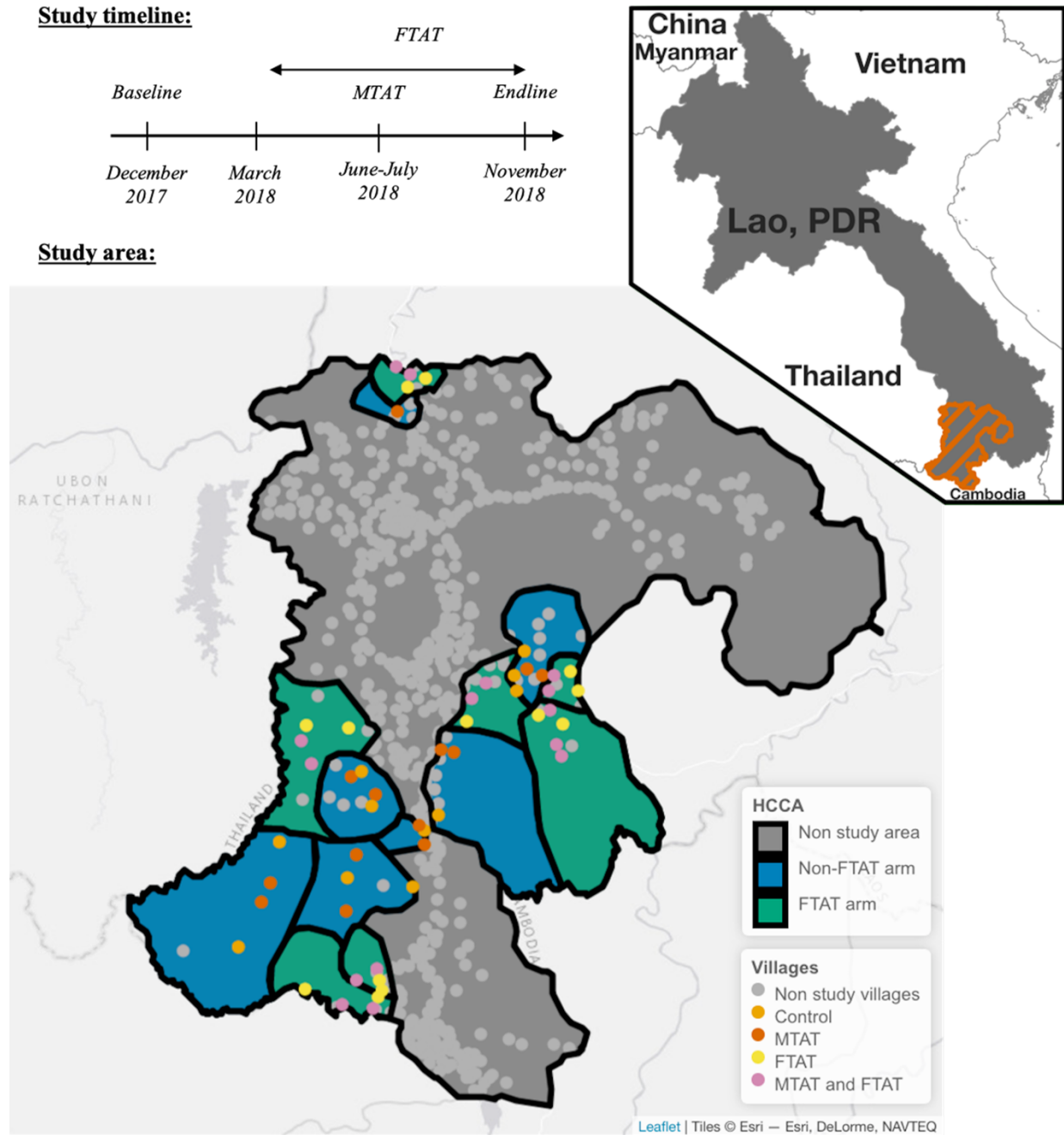
In this analysis, we estimated the population size of forest-goers in southern Lao People's Democratic Republic (PDR). Population-based surveys from a randomized controlled trial were used to produce PSEs at three different time points and capture-recapture methodology - drawing on those surveys in addition to a rolling survey of forest-goers - estimated the total number of forest-goers in the study area over the study period.

## **2.3. Methods**

### **2.3.1. Study area**

This study was conducted in Champasak Province, one of the five southernmost provinces in Lao PDR, together accounting for 95% of the country's malaria burden<sup>71</sup>. As part of a randomized controlled trial, surveys were conducted between December 2017 and November 2018 to assess the effectiveness of active case detection in village-based and forested-based settings<sup>66</sup>. Across four districts, 56 villages in 14 health center catchment areas (HCCA) were randomized to one of four arms: no intervention, Focal Test-And-Treat (FTAT), an intervention specifically targeting forest-goers, Mass Test-And-Treat (MTAT), where everyone was tested for malaria using rapid diagnostics tests (RDTs) and treated if positive or both interventions. The study area was selected in consultation with the national malaria program based on malaria burden (highest API in 2016). See Figure 2.1 for the study timeline and a map of the study area.

The rainy and dry seasons were defined, respectively, as the June to October and November to May periods in consultation with local health ministries and corroborated by actual precipitation data<sup>77</sup> (see Supplementary Fig. S4.1).



**Figure 2.1** - Study timeline and study area. Top left: Study timeline with 3 cross-sectional surveys conducted in December 2017 (Baseline), June-July 2018 (MTAT) and November 2018 (Endline) and a rolling FTAT survey between March and November 2018. Bottom: Study area with 7 of 14 health center catchment areas (HCCA) randomly assigned to FTAT and 28 of 56 villages randomly assigned to MTAT. The study was conducted in Champasak province in southern Lao PDR neighboring Thailand and Cambodia (see upper right indent).

### **2.3.2. Population Size Estimation**

We defined the HRP target population as individuals at increased exposure to malaria vectors due to spending the night outdoors for forest or agriculture activities.

In this paper, we report results from two population size estimation methods: population-based household surveys and capture-recapture. The first approach estimated the population proportion of HRP in the study area from three cross-sectional household surveys conducted at different time points during the year. Each proportion was combined with a census count of the total population in the area to produce three distinct PSEs. The capture-recapture methodology drew on individual information from the household surveys and data collected from an intervention among forest-goers conducted over the course of a year to produce another PSE.

These PSEs are complementary but do not estimate the same quantity. The population-based household surveys estimates are “snapshots” of the population size, corresponding to the time frame when the household surveys were conducted. The capture-recapture estimate represents the total population size of HRPs in the study area over the study period, from December 2017 to November 2018. These four estimates would be equal only if, every month, the same HRP individuals spent at least one night outdoors for forest or agriculture activities. If there is seasonality in forest-going, these PSEs should be different.

### ***2.3.2.1. Baseline and endline surveys***

For the baseline (December 2017) and endline (November 2018) cross-sectional surveys, simple random sampling was used to select 22 and 35 households respectively in each of the 56 study villages. Following written consent, all residents and visitors present in the household at the time of the visit were invited to participate in the survey. Heads of household were asked to answer questions on behalf of absent household members. Primary caretakers answered any questions pertaining to their children when they could not answer themselves. If no householder was at home at time of visit, the study team tried to revisit three times before randomly selecting a replacement household in the village from the household census. The survey was conducted in Lao language by local members of the ministry of health and the national research institution (Lao Tropical Public Health Institute) after receiving comprehensive training<sup>66</sup>. The surveys questioned participants on demographics, forest-going behaviors, treatment-seeking attitudes and malaria knowledge.

### ***2.3.2.2. MTAT survey***

Between June 12<sup>th</sup> and July 23<sup>rd</sup> 2018, the MTAT intervention was conducted, targeting every household in 28 villages randomly selected from among the 56 villages in the study area. Although questions differed slightly, data collection methods for the survey embedded in this intervention were the same as in the baseline and endline surveys. The study team attempted to visit an absent household three times before marking that household as ‘absent’. The households included in baseline, endline and MTAT were sampled independently from one another<sup>66</sup>.

### **2.3.2.3. FTAT survey**

In the FTAT intervention, conducted continuously between March and November 2018, peer navigators (PNs) were employed in intervention HCCAs to conduct test-and-treat activities amongst members of their communities presumed to be “forest-goers” because of their activities in or near the forest. PNs were themselves forest-goers recruited from the local communities via health authorities and trained to conduct continuous surveillance by testing for malaria using Rapid Diagnostic Tests (RDTs)<sup>66</sup>. PNs were instructed to actively target HRP individuals, and to enroll, once outside the villages, anyone meeting the FTAT HRP eligibility criteria: aged 15 years or older and having spent at least one night outside a formal village in the past 30 days. For 16 HRP individuals interviewed twice in FTAT, we included only data from the first interview.

### **2.3.3. HRP eligibility criteria**

Participants in the baseline, endline, and MTAT surveys were classified as members of the HRP target population if they were aged 15 years or older, were usual residents of the household, and met any of the criteria listed in Table 2.1. These criteria were based on responses to survey questions and varied slightly by survey due to differences in questionnaires. All participants in FTAT were classified as HRP due to the intervention’s eligibility criteria; however, we limited the FTAT sample to individuals who reported residing in the study area (56 villages) to ensure geographic alignment with the other surveys.

**Table 2.1 - HRP eligibility criteria.**

<b>Baseline and Endline criteria</b>	<b>MTAT criteria</b>
<b>A</b> - During the past month, stayed overnight away from home AND reason for the absence was working in the rice field, plantation or forest in this province or another province	<b>D</b> - During the past month, stayed overnight away from home village AND reason for travel was working in a rice field, agricultural or other plantation work, forest foraging, collecting small wood or timber, or logging
<b>B</b> - Did not sleep in the household the previous night due to working in the rice field, plantation or forest in this province or another province	<b>E</b> - During the past month, stayed overnight within 10km of home village AND travel destination was forest, forest fringes, rice field, other field or plantation
<b>C</b> - Spent at least 1 night in the forest, forest fringe, farms, or rice fields in the past month	<b>F</b> - Spent at least 1 night in the forest in the past month

### **2.3.4. PSE method 1: population-based household surveys**

First, we estimated the population proportion of HRP in the study area,  $p_{hrp}$ , as the percentage of participants aged 15 years and older in each household survey—baseline, endline and MTAT—that fulfilled the HRP eligibility criteria. Sampling weights and the clustering structure of the respective surveys were specified using the Survey<sup>101</sup> R<sup>83</sup> package to correctly estimate population proportions and standard errors.

Second, we developed a pooled estimate of the population proportion of individuals aged 15 and older in households in the study area,  $p_{15}$ , by combining, in a meta-analysis using inverse variance, the individual estimates from the 3 surveys.

Third, the total household population in the study area,  $Pop$ , was obtained by summing the population count listed in the household census across the 56 villages.

Finally, the population-based survey PSE was calculated for each survey as follows:

$$PSE = p_{hrp} \times p_{15} \times Pop \quad (\text{Eq 2.1})$$

The delta method<sup>102</sup> was used to calculate 95% confidence intervals for each PSE.

The three PSEs obtained from this method pertain to different time periods starting 1 month prior to the first day of the household survey until the last day of the survey (see Table 2.2).

Two sensitivity analyses were conducted to strengthen the robustness of our results. First, we considered how the differences among criteria may lead to an underestimate of the PSE for the MTAT survey. Second, we attempted to adjust for potential selection bias because of absent households. See Appendix 2 – S2.5 for details.

### **2.3.5. PSE method 2: Capture-recapture**

Survey participation represented “capture” in the respective survey. To identify participation of the same individual across surveys (i.e., “recapture”), survey records were matched based on age, sex, level of education, first initial and home village. Together, these identifying variables were unique for 99.5% of participants. The matching algorithm allowed plus or minus 2 years for age and 1 level apart for education because rounding age and self-reported education may have introduced errors. See Appendix 2 - S2.7 for details.



The overlap among the 4 lists of HRP individuals participating in surveys was analyzed using log-linear models<sup>103–107</sup> by the Rcapture<sup>108</sup> R<sup>83</sup> package. The models allowed for temporal dependence due to the potential seasonality of forest-going activities in two ways. First, we estimated a closed population model, where HRP individuals remain in the population all year long but where the probability of being captured differs across surveys because of varying probability of spending a night outside in a given month ( $M_t$  models). Second, we estimated an open population model, in which HRP individuals may migrate in and out of the population depending on whether or not they spent a night outside in a given month. Both models were designed to estimate the same PSE: the total number of HRP individuals in the study area any time during the 1-year study period from December 2017 to November 2018. See Appendix 2 – S2.8 for details.

Two sensitivity analyses estimated a lower bound of the PSE by either relaxing the matching criteria or augmenting the eligibility criteria in FTAT. In a third sensitivity analysis, we leveraged the participation of non-HRP individuals in the three household surveys to assess and correct for potential matching errors in the record linkage algorithm. See Appendix 2 – S2.6 for details.

## 2.4. Results

### 2.4.1. Data description

#### 2.4.1.1. Household-based surveys

In the baseline, MTAT and endline surveys respectively, 5,723, 18,143 and 7,870 individuals across 1,310, 4,489 and 2,081 households, were interviewed. Responses required to construct HRP criteria were provided by 99.6%, 97.4%, and 99.9% of baseline, MTAT and endline participants, respectively (see Supplementary Tables 2.4, 2.5 and 2.6 in appendix 2).

Of those 47,575 inhabitants living in the study area - *Pop* in equation (1) -, 63.5% (95% CI: [62.9%; 64.2%]) were estimated to be older than 15 years -  $p_{15}$  in equation (1). See Appendix 2 – S2.3 for details.

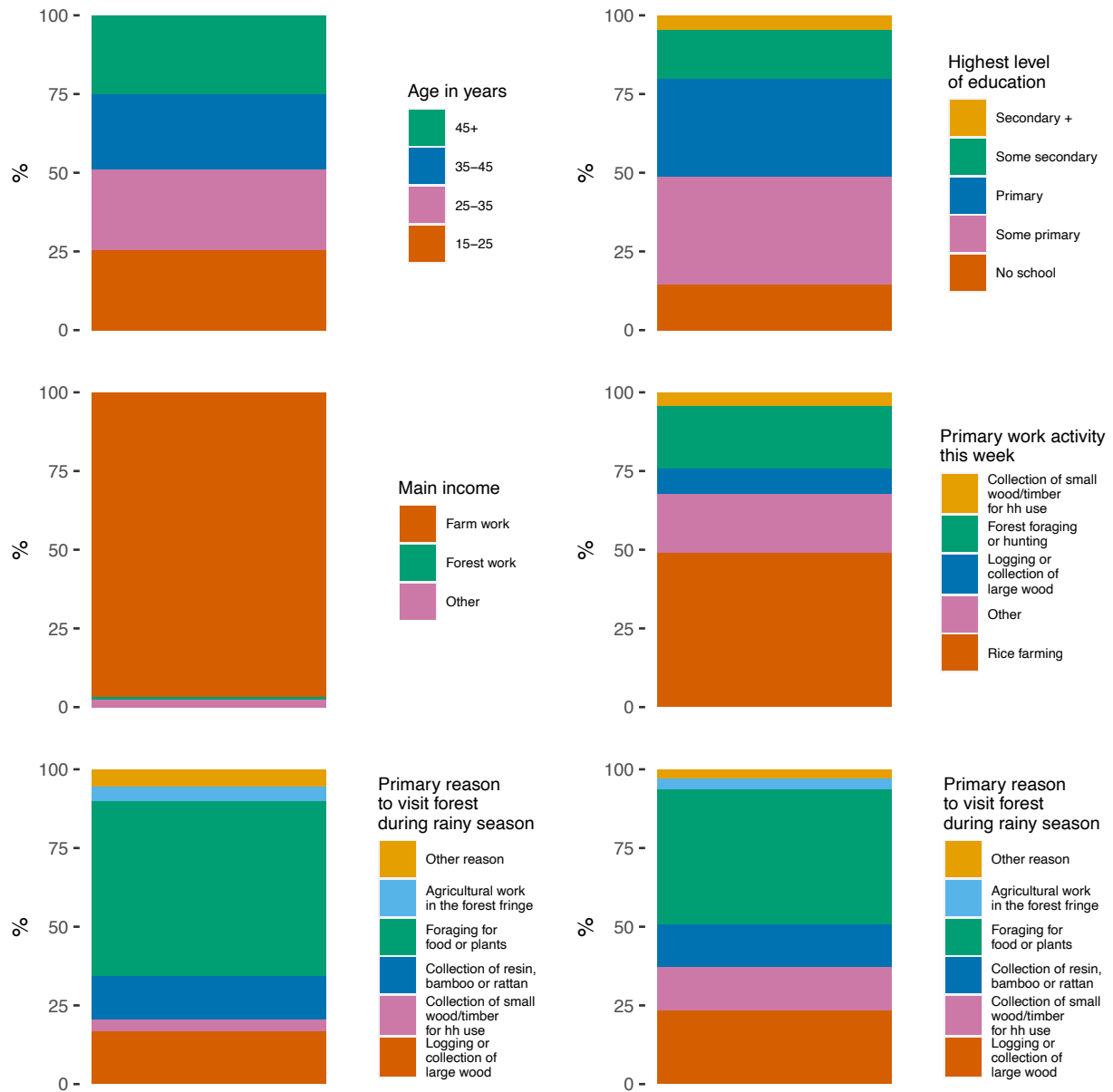
#### 2.4.1.2. FTAT survey

Among the 2,888 HRP individuals recruited into the FTAT survey, 2,305 (79.8%) came from one of the 56 villages in our study area and were included in this study. Supplementary Fig. S4.2 shows the weekly enrollment.

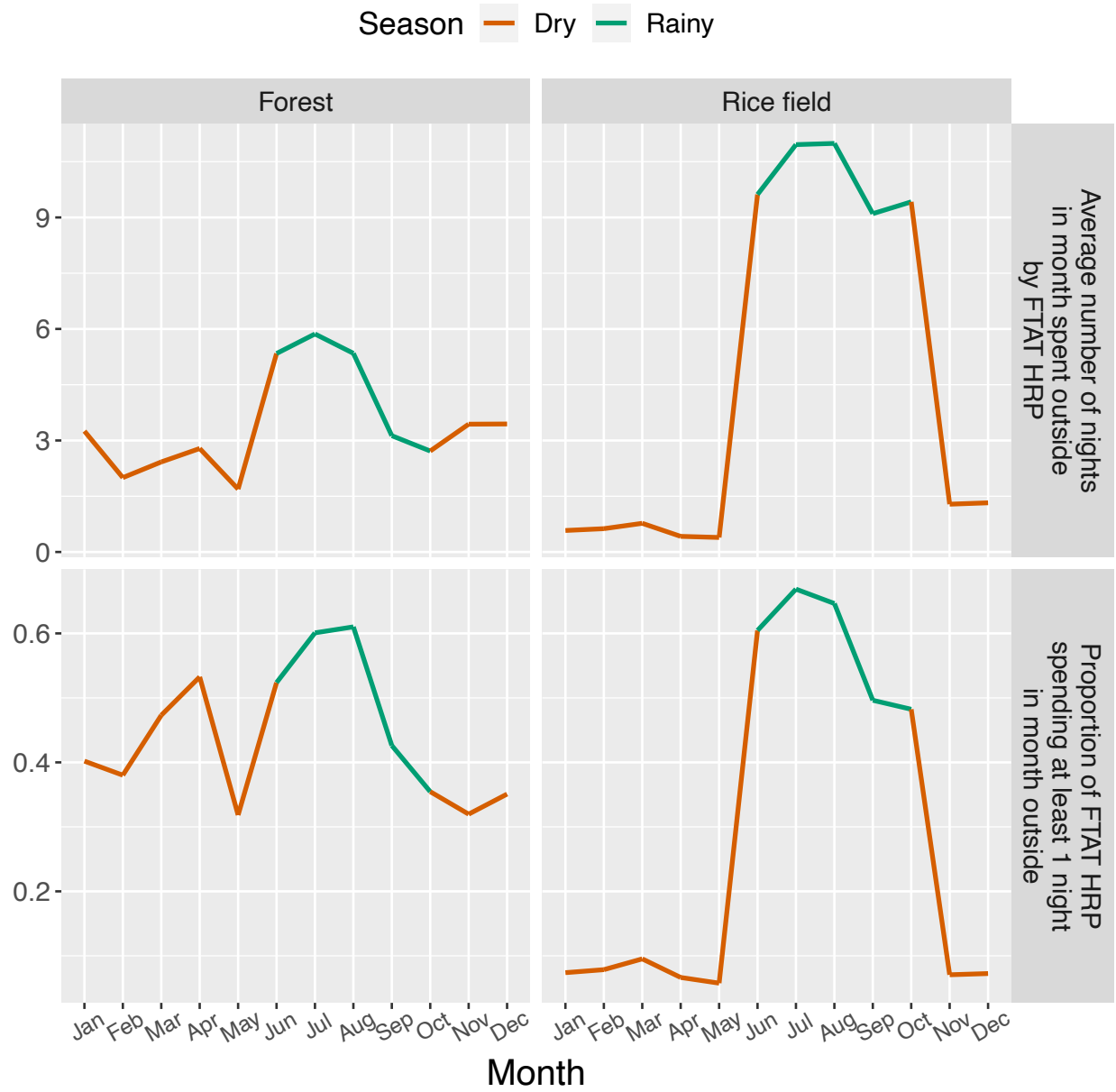
Figure 2.2 shows the distribution of selected variables from the FTAT survey. Males were more represented (67.2%) than females (32.8%) and the average age was 36.4 years. A majority (96.6%) of HRP individuals earned their primary income from agricultural work and about 50% reported rice farming as their primary activity. The proportion of HRP individuals reporting the

collection of wood as the primary reason to visit the forest almost doubled between the rainy (20.4%) and dry (37.3%) seasons.

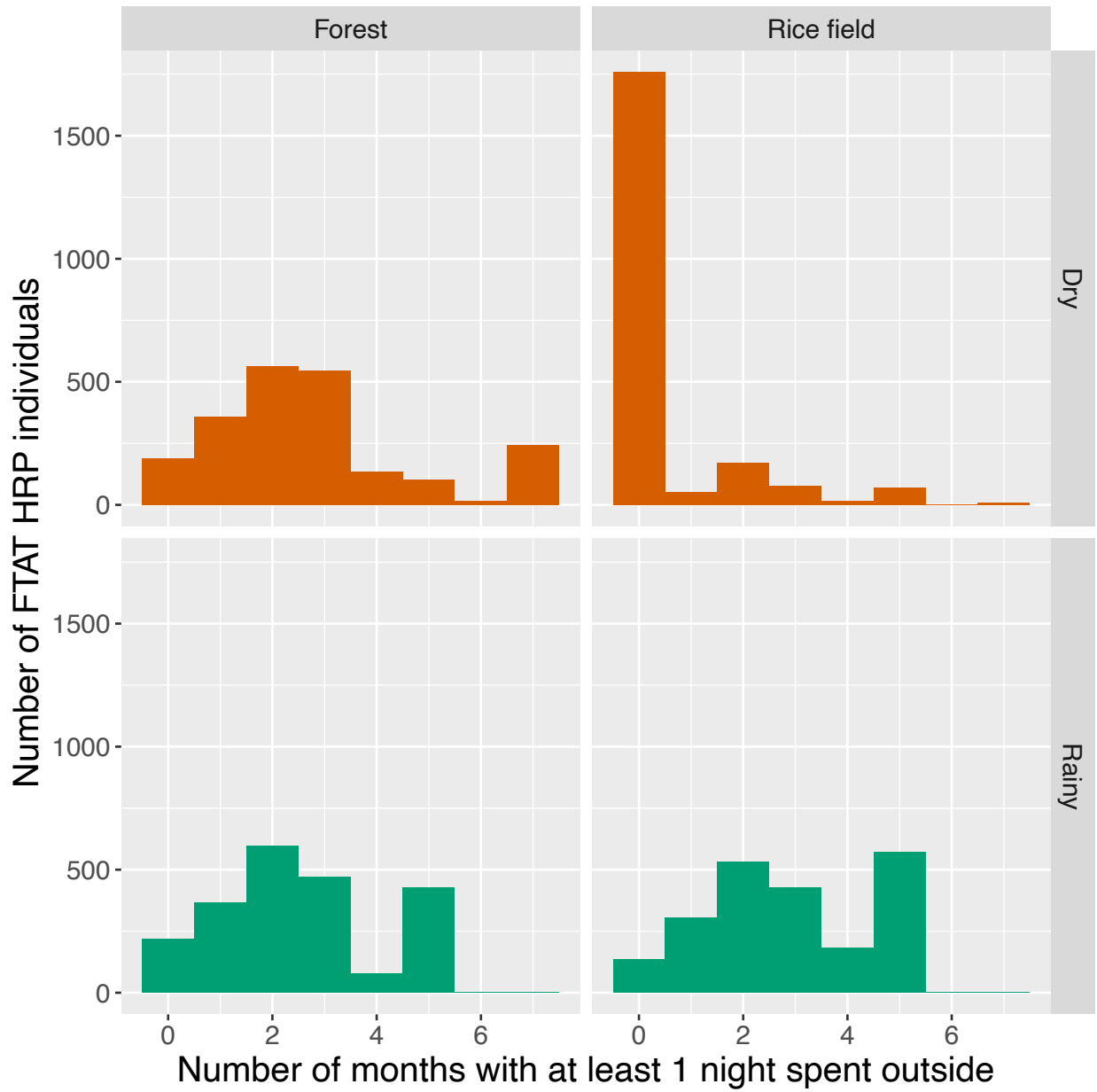
The number of nights typically spent outside each month of the year was reported by FTAT HRP individuals and summarized in Figures 2.3 and 2.4. During the rainy season, activities in the rice field intensified, with about 60% of HRP individuals spending at least one night outside in any given month and about 10 nights per month spent outside on average. During the dry season, few HRP individuals reported spending a night in the rice field. In contrast, forest-going was characterized by a greater average number of nights and a greater proportion of HRP individuals spending a night outside during the rainy season and occurred more regularly throughout the year. Across all months, at least 30% of HRP individuals reported spending at least 1 night in the forest. These plots also suggest a high level of turnover with many HRP individuals reporting spending nights outside in only 1 to 3 months of either the dry or rainy season.



**Figure 2.2 - Demographics of FTAT HRP.** Age, education, income, work activity, reasons to visit the forest of HRP individuals enrolled in FTAT survey.



**Figure 2.3** - Seasonality of FTAT HRP. Top row - Average number of nights spent outside in the forest or rice field by FTAT HRP individuals over time. Bottom row - Proportion of FTAT HRP individuals spending at least 1 night in month outside in the forest or rice field over time.



**Figure 2.4** - Turnover of FTAT HRP. Distribution of the number of months in which FTAT HRP individuals reported spending at least 1 night outside in the forest or rice field during the rainy (5 months between June and October) and dry (7 months between November and May) season.

## 2.4.2. PSE method 1: Population-based household surveys

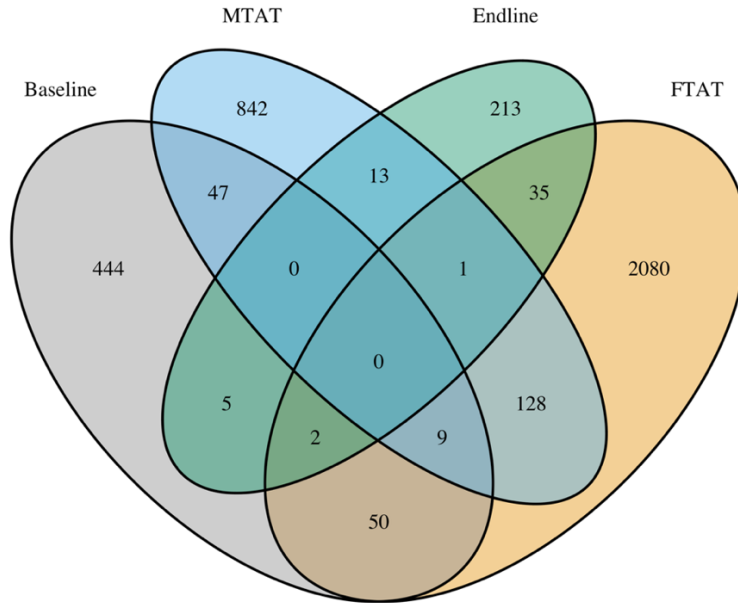
Table 2.2 presents the estimated population proportion of HRP and the resulting PSE from each of the three cross-sectional household surveys.

**Table 2.2** - Results for the population-based household survey method for population size estimation of HRP individuals.

	PSE time period	Rainy/Dry season	% HRP [95% CI]	PSE [95% CI]
Baseline	Oct 28 <sup>th</sup> – Dec 9 <sup>th</sup> 2017	First month of dry season	16.2 [14.7; 17.7]	4,898 [4,445; 5,361]
MTAT	May 12 <sup>th</sup> – July 23 <sup>rd</sup> 2018	First month of rainy season	9.3 [7.2; 11.3]	2,801 [2,180; 3,395]
Endline	Sep 31 <sup>st</sup> – Nov 19 <sup>th</sup> 2018	Last month of rainy season	5.3 [4.4; 6.1]	1,586 [1,328; 1,844]

## 2.4.3. PSE method 2: Capture-Recapture

A total of 557, 1,040, 269 and 2,305 HRP individuals from the study area were captured in the baseline, MTAT, endline and FTAT surveys, respectively. After matching participants, 3,869 unique HRP individuals were identified across the four surveys. Figure 2.5 presents a Venn Diagram of these capture history data.



**Figure 2.5** - Capture history. Venn Diagram of the capture history data. For instance, 128 HRP individuals were captured both in the MTAT and FTAT surveys but not in baseline or endline surveys.

Table 2.3 shows the capture-recapture results from log-linear models fit to our data, assuming a closed population. Models that allowed for correlation of the probability of selection across surveys (denoted  $M_b$ ), did not perform well. Their fit to the data, as indicated by the AIC and BIC, was poor and their PSEs were barely above 3,869, the total number of unique HRP individuals captured. The model allowing for temporal dependence ( $M_t$ ) yielded the best fit in terms of both AIC and BIC. Allowing for heterogeneity ( $M_h$ ) between individuals in terms of their selection probability did not result in a better fit in any parametrization (i.e., Chao, Poisson2, Darroch or Gamma 3.5). Additionally, Appendix 2 – S2.9 details a diagnostic test that determined it was not necessary to incorporate heterogeneity in the models.



**Table 2.3** - Capture-recapture PSE results using log-linear models and assuming closed population.

	PSE	Standard Error	Deviance	Df	AIC	BIC
M0	20,892.7	1,103.9	2,384.2	13	2461.4	2474
Mt	17,106.6	878.1	23	10	106.3	137.6
Mh Chao (LB)	21,136.3	1,192.3	2,383.8	12	2,463.1	2,481.9
Mh Poisson2	22,317.5	3,240.9	2,383.9	12	2,463.2	2,482
Mh Darroch	24,222.1	6,664.4	2,383.9	12	2,463.1	2,481.9
Mh Gamma3.5	26,272.7	10,795.7	2,383.8	12	2,463.1	2,481.9
Mth Chao (LB)	17,476	953.1	21.4	9	106.7	144.3
Mth Poisson2	19,900	2,782.9	21.7	9	107	144.6
Mth Darroch	23,693.6	6,493.3	21.6	9	106.9	144.4
Mth Gamma3.5	28,267.4	11,782.7	21.5	9	106.8	144.4
Mb	6,052.6	210.2	2,182.4	12	2,261.7	2,280.5
Mbh	3,998.6	30.7	200.2	11	281.5	306.6

Profile likelihood<sup>109</sup> was used to calculate a 95% CI of [15,502; 18,959] for the  $M_t$  capture-recapture PSE of 17,107. Changing our conceptual framework for temporal dependence and modeling an open population did not improve the fit to our data (AIC = 108.2) and yielded a similar PSE of 17,008 [15,136; 18,880]. In our final  $M_t$  model, an additional interaction term between baseline and MTAT improved the fit (AIC = 93.0) and led to a final PSE of 18,426 [16,529; 20,669], representing 61.0% [54.2; 67.9] of the household population aged 15 years or older in the study area (using equation 1). See Appendix 2 – S2.2 for additional results.

## 2.5. Discussion

Based on data from a randomized controlled trial conducted between December 2017 and November 2018, we applied two methods to estimate the number of forest-goers in Champasak province in southern Lao PDR. Leveraging the different timing of three cross-sectional household surveys, forest-going HRPs were found to represent 16.2% [14.7; 17.7], 9.3% [7.2; 11.3], and 5.3% [4.4; 6.1] of the household population older than 15 years during the first month of the dry season, the first month of the rainy season, and the last month of the rainy season, respectively. The capture-recapture method estimated a total population size of 18,426 [16,529; 20,669] forest-goers present at any time over the period, representing 61.0% [54.2; 67.9] of the population 15 years or older.

A key finding from this study is that a large majority of adult residents in the study area spent at least one night outdoors for forest or agricultural activities over the course of a year. This has important implications for malaria control programs, suggesting they may have underestimated the size of forest-going populations that are increasingly targeted by prevention and treatment efforts<sup>23,24</sup>. Alternatively, these results call for a more stringent definition of the forest-going label to identify higher-risk forest-goers. For instance, in crude analyses, some HRP eligibility criteria in Table 2.1 such as criteria A, B, C or F were positively associated with PCR malaria whereas no associations were found for criteria D or E. As highlighted in a recent systematic review of the qualitative literature on forest-goers in the GMS<sup>28</sup>, a better characterization of the activities that put forest-goers at increased risk for malaria is needed. This is critical to clearly identify and count HRPs.

Data from the FTAT survey showed that forest-going HRP individuals were much more active during the rainy season, especially in the rice fields. In contrast, the household surveys identified a greater number of forest-goers during the dry season than the rainy season. Yet, this difference may be an artifact of selection bias since twice as many households approached during the rainy season (i.e. in the MTAT survey) could not be enrolled due to householders being away, compared to the dry season (i.e. in the Baseline survey). Anecdotal evidence from field teams suggests that households were often vacant because household members were working in the forest or at agriculture sites. That said, sensitivity analyses found that no more than 25% of the population had spent a night outside for forest or agricultural activities in a given month during either the rainy or dry season. This implies a high turnover among the forest-going HRPs with individuals spending a night outside for forest or agriculture activities only in certain months of the year. Seasonality and turnover thus appear to be important considerations when designing interventions to access and treat these forest-going HRPs. For instance, our results show a drop in the number of forest-goers active toward the end of the rainy season which could be leveraged by interventions to more effectively target forest-goers both in the forest and in the villages, where many may have already returned.

In our statistical models, the closed population assumption was most consistent with the data, suggesting there was no change in the HRP population over the one-year study period. The seasonality and turnover among forest-going HRPs highlighted in the FTAT data and, additionally, by variation across the household-based PSEs, was accommodated in closed population models  $M_t$  with a capture probability allowed to vary among surveys. Another way to account for this temporal dependence was to change our conceptual framework and restrict our

HRP definition to individuals spending a night outside for forest or agriculture activities in a *given month* of the one-year study period. As the number of HRP individuals can now vary between two months, open population models were used. Importantly though, we considered this alternative approach to estimate the same PSE, i.e. the total population size of HRPs in the study area during the study period. Results from both conceptual frameworks were consistent.

Routinely used in HIV surveillance<sup>97</sup>, PSEs could strengthen the malaria surveillance arsenal<sup>100</sup>. Population-based surveys, unsuitable to identify hidden and hard-to-reach HRPs in HIV, are simple methods frequently used in malaria research that could be leveraged for PSEs of forest-goers in the GMS and other high-risk subgroups such as cattle-herders in southern Africa<sup>110</sup>. As illustrated here, in the presence of seasonal risk behaviors, PSEs reflecting different periods over the malaria season can be obtained by conducting multiple surveys at different time points. A cumulative PSE can be obtained by applying capture-recapture to three or more data sources, ideally including a longitudinal survey, such as our FTAT survey. Additional sources might include surveillance data routinely collected by malaria programs or more targeted data such as surveys at known venues where HRP congregate. Equipped with routine PSEs, malaria control program could better serve their respective HRP. In the GMS for instance, they could determine how many hammock nets to distribute and at which time of the year.

Our results do have limitations. First, the study did not use a standardized definition for forest-going HRPs across surveys. To address this issue, we combined multiple survey questions to identify HRP individuals. Our sensitivity analysis estimated a possible 17% undercount in the MTAT survey. Second, asking heads of households to answer question items to assess HRP

criteria on behalf of absent household members may have contributed to more complete data, but also to misclassification. Our formative work indicated forest and agricultural activities were not significantly stigmatized in the study area so this should not have led to a meaningful bias.

Lastly, individuals could not be matched across surveys on full names so that initials were used, potentially leading to matching errors. In a sensitivity analysis, captures of non-HRP individuals in household surveys were leveraged to evaluate that such matching errors could have led to no more than an 11 % undercount in our capture-recapture PSE.

In conclusion, this study estimated the overall proportion that forest-going HRPs represent in southern Lao PDR, highlighted an important seasonality in malaria risk behaviors, and illustrates population size estimation methods that can be replicated to support national control programs in the GMS to assess and meet the 2030 malaria elimination goals<sup>7,12</sup>.

## 2.6. Appendix 2

### 2.6.1. S2.1: additional tables for population-based household survey method

**Table 2.4** - Identification of HRP individuals in baseline survey.

Measure	N	n	% [95% CI]	Missing observations n (%)
Usually resident in HH	5,723	5,593	97.3 [96.5; 98.0]	1 (0)
Age older than 15 among usual residents	5,593	3,378	60.2 [59.1; 61.3]	2 (0)
HRP criteria:				
A	3,378	383	11 [9.7; 12.4]	1 (0)
B	3,378	138	3.9 [3.0; 4.7]	5 (0.1)
C	3,378	443	13.1 [11.7; 14.4]	8 (0.2)
Any of A, B, C	3,378	557	16.2 [14.7; 17.7]	13 (0.4)

**Table 2.5** - Identification of HRP individuals in MTAT survey.

Measure	N	n	% [95% CI]	Missing observations n (%)
Age older than 15	18,143	11,526	63.5 [62.0; 65.1]	5 (0)
HRP criteria:				
D	11,526	879	7.8 [6.0; 9.7]	280 (2.4)
E	11,526	831	7.4 [5.6; 9.2]	277 (2.4)
F	11,526	284	2.5 [1.5; 3.6]	288 (2.5)
Any of D, E, F	11,526	1,040	9.3 [7.2; 11.3]	302 (2.6)

**Table 2.6** - Identification of HRP individuals in endline survey.

Measure	N	n	% [95% CI]	Missing observations n (%)
Usually resident in HH	7,870	7,678	97.5 [97.0; 98.1]	1 (0)
Age older than 15 among usual residents	7,678	5,023	65.9 [64.9; 66.8]	0 (0)
HRP criteria:				
A	5,023	189	3.7 [3.0; 4.5]	0 (0)
B	5,023	85	1.6 [1.1; 2.0]	0 (0)
C	5,023	215	4.3 [3.5; 5.0]	1 (0)
Any of A, B, C	5,023	269	5.3 [4.4; 6.1]	1 (0)

## 2.6.2. S2.2: additional tables for capture-recapture method

**Table 2.7** - Capture-recapture  $M_t$  PSE using 2, 3 or 4 of the survey lists available.

Number of lists	Surveys	PSE [95% CI]
4	FTAT, MTAT, baseline, endline	17,107 [15,502; 18,959]
3	FTAT, MTAT, baseline	16,999 [15,241; 19,061]
	FTAT, MTAT, endline	17,247 [15,221; 19,679]
	FTAT, baseline, endline	19,578 [16,459; 23,532]
	MTAT, baseline, endline	12,946 [10,594; 16,089]
2	FTAT, MTAT	17,371 [15,012; 20,295]
	FTAT, baseline	21,047 [16,836; 26,886]
	FTAT, endline	16,317 [12,387; 22,256]
	MTAT, baseline	10,344 [8,216; 13,333]
	MTAT, endline	19,983 [12,567; 34,880]
	baseline, endline	21,405 [11,238; 49,131]

**Table 2.8** - Capture-recapture PSE for various models considered.

Model	PSE [95% CI]	AIC
Closed population $M_0$	20,886 [18,877; 23,223]	2461.44
Closed population $M_t$	17,107 [15,502; 18,959]	106.30
Open population	17,008 [15,136; 18,880]	108.23
<b>Final model: Closed population <math>M_t</math> with baseline-MTAT interaction</b>	<b>18,426 [16,529; 20,669]</b>	<b>92.97</b>

**Table 2.9** - Capture-recapture PSE for Mt models with additional interaction terms between surveys.

	PSE	Standard Error	Deviance	Df	AIC	BIC
Baseline & MTAT	18,425.9	1,050.7	7.7	9	93.0	130.5
Baseline & MTAT; Baseline & FTAT	17,640.5	1,139.2	5.9	8	93.2	137.0
Baseline & MTAT; FTAT & MTAT	19,703.4	1,681.1	6.5	8	93.8	137.6
No interaction	17,106.6	878.1	23	10	106.3	137.6
Baseline & MTAT; FTAT & Endline	18,788.3	1,160.9	7.0	8	94.2	138.1
Baseline & MTAT; Baseline & Endline	18,361.6	1,058.3	7.5	8	94.8	138.6
Baseline & MTAT; MTAT & Endline	18,354.3	1,071.8	7.6	8	94.9	138.7
Baseline & FTAT	16,173.5	914.4	18.7	9	104.0	141.6

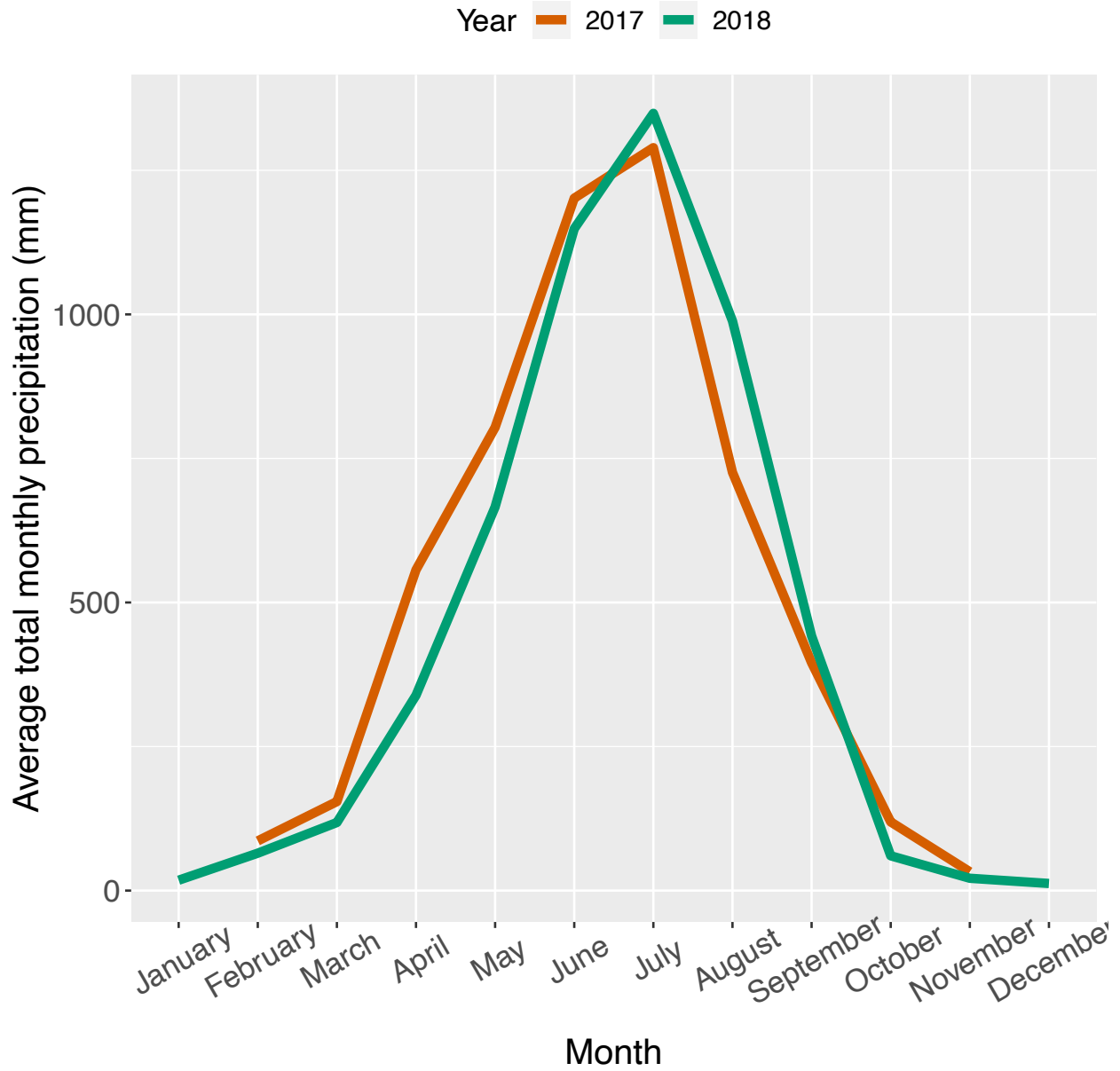


### 2.6.3. S2.3: additional table for meta-analysis estimating p15

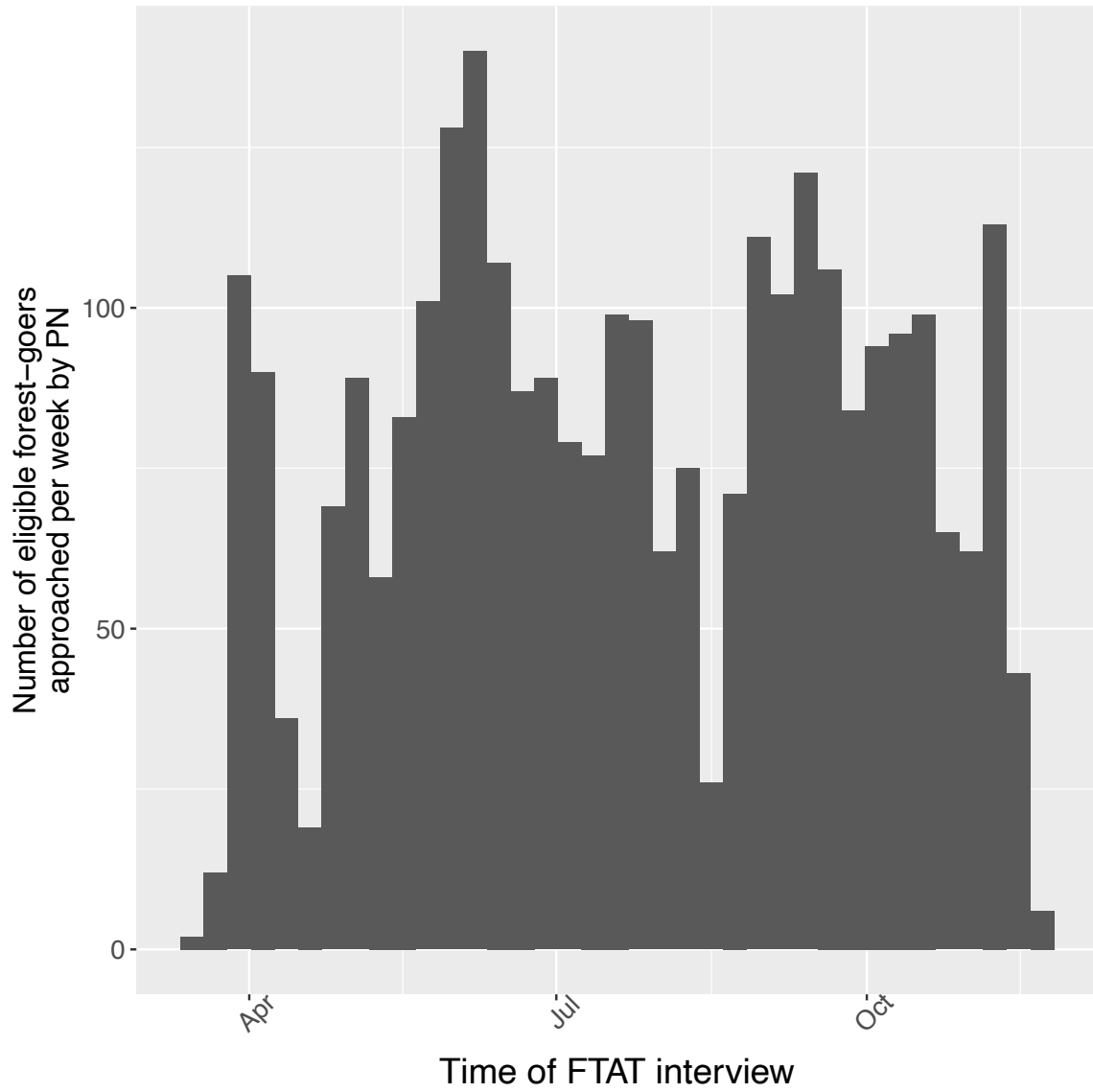
**Table 2.10** - *Meta-analysis to estimate proportion of population older than 15.*

	Survey period	Age older than 15 (%) [95% CI]	Age older than 15 (SE)
Baseline	Nov 28 <sup>th</sup> – Dec 9 <sup>th</sup> 2017	60.2 [59.1; 61.3]	0.00577
MTAT	June 12 <sup>th</sup> – July 23 <sup>rd</sup> 2018	63.5 [62.0; 65.1]	0.00794
Endline	Oct 31 <sup>st</sup> – Nov 19 <sup>th</sup> 2018	65.9 [64.9; 66.8]	0.00485
Overall		63.5 [62.9; 64.2]	0.00336

## 2.6.4. S2.4: Additional figures



**Figure 2.6** - Precipitation time series. Average total monthly precipitation (mm) in Champasak, southern Lao PDR. Precipitation data from CHIRPS in all 300m<sup>2</sup> pixels of Champasak province were averaged for all months of 2017 and 2018.



**Figure 2.7 - FTAT HRP enrollment.** Enrollment of HRP individuals in FTAT survey over time.

## **2.6.5. S2.5: Sensitivity analyses for population-based household survey PSE**

In a first sensitivity analysis, we considered how the differences among criteria may lead to an underestimate of the PSE for the MTAT survey. Indeed, criterion A in the baseline and endline surveys maps onto criterion D in the MTAT survey, but the differences between criteria C and F and the absence of an equivalent for criteria B in MTAT may result in an undercount of HRP individuals in MTAT. Therefore, we calculated the proportion of HRP individuals missed in the baseline and endline surveys if only criteria A was to be used, and upweighted the count of HRP individuals in MTAT accordingly.

In a second sensitivity analysis, we attempted to adjust for potential selection bias because of absent households during surveys. Indeed, if absence was related to forest or agriculture activities, as was informally reported by survey teams, our PSE estimates would be biased, presumably downwards. In this sensitivity analysis, we replaced absent households by households with 25%, 50% or 75% household members qualifying as HRP. Absent and surveyed households were taken to be of the same size. This rather conservative sensitivity analysis estimates upper bounds for the PSE in each of the three population-based surveys and quantifies the possible impact of such selection bias. Adjustment for differences in HRP eligibility criteria across surveys, as discussed in the first sensitivity analysis, was also included here.

Table 2.11 presents the results of our sensitivity analyses. In the first sensitivity analysis, we estimated that 30% of HRP individuals would be missed if only criterion A was to be used in the baseline and endline surveys. As a result, we upweighted the MTAT HRP count (if only criteria

D was to be used) to estimate 11.1% [8.6; 13.9] for the population proportion of HRP in MTAT. This is slightly higher than the 9.3% [7.2; 11.3] reported in the primary analysis (Table 2.3) but not massively different. In the second sensitivity analysis, we estimated an upper bound for each of the three population-based PSEs when attempting to adjust for selection bias and replaced absent households by households with 25%, 50% or 75% household members qualifying as HRP. At baseline, MTAT and endline respectively, 6.2%, 16.3% and 7.4% of surveyed households were absent.

**Table 2.11** - Results for the population-based survey method for population size estimation of HRP individuals. Sensitivity Analyses.

		Sensitivity analysis 1	Sensitivity analysis 2		
		Address differences in HRP's eligibility criteria among surveys	Absent households during surveys replaced by households where X% of households' members are HRP individuals		
	Primary analysis		25%	50%	75%
	% HRP [95% CI]	% HRP [95% CI]	% HRP [95% CI]	% HRP [95% CI]	% HRP [95% CI]
Baseline	16.2 [14.7; 17.7]	16.2 [14.7; 17.7]	16.7 [15.3; 18.2]	18.3 [16.9; 19.7]	19.8 [18.4; 21.3]
MTAT	9.3 [7.2; 11.3]	11.2 [9.3; 13.0]	13.4 [11.6; 15.0]	17.5 [15.9; 19.0]	21.6 [20.0; 23.1]
Endline	5.3 [4.4; 6.1]	5.3 [4.4; 6.1]	6.8 [5.9; 7.5]	8.6 [7.8; 9.3]	10.5 [9.6; 11.2]

### **2.6.6. S2.6: Sensitivity analyses for capture-recapture PSE**

Three sensitivity analyses were conducted to strengthen the robustness of our results. The first two sensitivity analyses estimate a lower bound for our PSE by either relaxing the matching criteria or augmenting the eligibility criteria in FTAT. In the first one, matches with plus or minus 3 years for age and plus or minus 2 ordered education categories apart were additionally accepted in our algorithm to relax our matching constraints. In the second one, the eligibility criteria for FTAT were augmented to only include individuals who, when asked specifically which months of the year they tended to spend a night outside in the forest or in the rice field, listed one of the two months prior their FTAT interview (May and June for someone interviewed in June for instance).

The third sensitivity analysis attempts to assess and correct for potential matching errors among HRP individuals, either because of which identifying variables were selected or how they were used in the matching algorithm. Unlike conventional capture-recapture studies, some of our traps, namely the three population-based surveys, not only capture HRP but also non-HRP individuals. These can be considered captures of individuals living in the study area, regardless of their HRP status. Using these captures, the total population in the study area was estimated by running the same capture-recapture methodology with the same matching algorithm and the same identifying variables on the complete four survey lists. This estimate was compared to the actual total population, known from the census count, to infer how biased the capture-recapture PSE may be and correct it accordingly. Such correction relies on the reasonable assumption that matching errors between two individuals are as likely to happen in the general population as among HRP individuals.

Results from our first two sensitivity analyses, which can be viewed as tentative to estimate a lower bound, yielded  $PSE = 15,305$  [13,886; 16,959] when matches with plus or minus 3 years of age and plus or minus 2 ordered education categories apart were allowed and  $PSE = 16,380$  [14,555; 18,566] when FTAT eligibility criteria was augmented. In the third sensitivity analysis, trying to adjust for potential mismatch among HRP individuals because of which identifying variables were selected and how they were used in the matching algorithm, the  $M_t$  log-linear model estimated a total population in the study area of 52,739 [51,655; 53,878] which is close to the true 47,575 in the household census count. This discrepancy means that our matching algorithm slightly underestimates the true overlap among the four lists of HRP individuals. Assuming the degree of mismatching error is the same among the general population as among HRP individuals, correcting the estimate by a 1.11 factor would yield a  $PSE = 16,622$  [14,911; 18,646]. All of these three sensitivity analyses yielded fairly similar PSEs and strengthened the robustness of our estimate.

## **2.6.7. S2.7: Details on record matching for population-based PSEs**

First, each survey sample was restricted to participants who met the same HRP criteria applied in the population-based method. Four different lists of HRP individuals were therefore extracted: baseline, MTAT, endline and FTAT.

HRP individuals needed to be uniquely identified in a consistent manner across surveys in order to ascertain overlap. Age, sex, highest level of education, first name initial and home village were extracted as variables collected in all surveys with the potential to uniquely identify HRP individuals across surveys. Age was either self-reported or computed from date of birth and was rounded to years. Because surveys were conducted at different times, age was standardized across surveys using July 1<sup>st</sup>, 2018 (middle of MTAT) as reference. Education was self-reported using 6 categories (none, some primary school, completed primary school, some secondary school, completed secondary school, more than secondary school). Ethnicity was not used because it was unavailable for participants other than heads of household in the 3 cross-sectional surveys. Only the initials of the first and last names were reported in the FTAT survey while the complete first name was reported in the 3 cross-sectional surveys. Thus, names could only be matched based on the first initial: the initial was extracted from the complete first name after removing titles (e.g., “Miss”, “Mister”), which were identified in collaboration with a Lao consultant. Overall, the combination of these five identifying variables was unique for 99.5% of HRP individuals in all data sources (100% in baseline and endline, 99.5% in MTAT and 98.9% in FTAT).



HRP individuals from the 4 different lists were matched based on age, sex, level of education, first initial and home village. Matches with plus or minus 2 years for age and plus or minus 1 ordered education categories apart (or missing) were accepted. This flexibility was allowed because rounding age may have introduced errors and self-reported education was considered to be less reliable and prone to social desirability and recall biases with individuals potentially mixing up secondary and primary school or reporting school completion instead of some school in two different surveys. For names, records from cross-sectional surveys were matched with FTAT participants when the first initial extracted from the cross-sectional survey matched one of the initials reported in FTAT. Indeed, the first initial could not be isolated from the last initial in the FTAT survey. When a record appeared to match multiple records from another data source, perfect matches were favored. In 19% of all matches, ties persisted, and we randomly selected one of them.

## 2.6.8. S2.8: Methodological details on capture-recapture

### 2.6.8.1. Theory

The capture-recapture methodology, originating in wildlife ecology, relies on the overlap of animals captured on different trap occasions to estimate the unknown total population size. Animals captured in the first sample are marked and then released back in the population. In the second sample, there will be animals tagged, meaning they were captured in the first sample, and untagged animals. This results in a two-source capture-recapture dataset with two samples of respective size  $n_1$  and  $n_2$  comprising  $m_2$  recaptures. The Petersen estimator<sup>111</sup> equates the proportion of tagged animals in the population and in the second sample to estimate the total population size of animals  $\hat{N} = n_1 \times n_2 / m_2$ . If more than two samples are collected, the process is repeated, uniquely tagging captured individuals on each occasion before releasing them. The capture histories for each individual - e.g., 01011 for an animal captured on occasions 2, 4 and 5 but not on occasions 1 and 3 - are then analyzed to estimate the total population size.

Typically, the simple estimators rely on the key assumptions that 1) the population is closed to additions and deletions, 2) all animals are equally likely to be caught on each capture occasion and 3) marks are reliable to assess the capture histories. That said, advanced methods such as log-linear models<sup>103–107</sup>, can relax some of those assumptions.

Here, the different surveys serve as traps and HRP individuals are “captured” when they participate in a survey. Hence, each of the four surveys represent a different capture occasion. Unique identifying variables can be used to track in which surveys individuals were captured.

### 2.6.8.2. *Statistical analysis*

The overlap among the 4 lists of HRP individuals was analyzed using log-linear models<sup>103–107</sup> available in the Rcapture<sup>108</sup> R package. Ubiquitous in the capture-recapture methodology, log-linear models leverage the overlap among lists to estimate the capture probability  $p$ , i.e the probability of being captured in a survey. In the simplest model  $M_0$ , the population is assumed to be closed with no in or out immigration and  $p$  is the same for all individuals and all traps, i.e surveys here. These stringent assumptions can be relaxed to model an open population or to allow heterogeneity in  $p$ , with additional parameters to be estimated, depending on the data available and assumptions researchers are willing to make. Three main sources of heterogeneity are discussed in the literature and available for modeling in the Rcapture<sup>108</sup> package. First, models  $M_t$  allow temporal dependence where the capture probability  $p_t$  can change between capture occasions, i.e surveys here. Second, models  $M_h$  allow the capture probability  $p_h$  to vary between individuals of the population. Third, trap dependence can be modeled in  $M_b$  where individuals' capture probability  $p_b$  can depend on their previous capture history. Any combination of the 3 sources of heterogeneity can be modeled such as in  $M_{th}$  where temporal dependence and individual heterogeneity are allowed but not trap dependence. Last, interaction terms can be included in log-linear models<sup>112</sup> when capture probabilities are correlated between two lists at a population level - referred as list dependency in the epidemiological terminology<sup>105</sup>. The fit to the data, as indicated by the AIC or BIC, should guide model selection.

In our study, trap dependence does not make sense as the probability of being in one survey should not depend on whether or not an individual was in a previous survey. In particular, surveys' samples were randomly selected independently from one another. Heterogeneity in

gender, age or ethnicity may definitely result in capture probability heterogeneity among HRP individuals. Our formative work indicated that spending a night outside in the forest or in the rice field was not stigmatized in the region, but some HRP criteria were answered by the head of the household, who may not be fully aware of the forest going habits of their household members. In addition, our capture instruments, i.e surveys, may not cover the whole spectrum of HRP forest-going activities equally and may therefore introduce some heterogeneity in the capture probabilities. For instance, HRP individuals traveling frequently in and out of the forest or through major forest entry points may be more likely to be ascertained in the FTAT survey. Surveys' questions also rely on a local understanding of what "sleeping", "forest", "home" or "sleeping outside in the forest away from home" mean and may result in different capture probabilities among HRP individuals.

The number of HRP individuals captured is expected to fluctuate across different months because of the influence of the dry and rainy seasons on livelihoods and forest or agricultural activities. In our modeling framework, we can conceptualize this temporal dependence in two ways. First, we considered a closed population where HRP individuals remain so all year long but where the probability of being captured and identified as a HRP in a survey varies across surveys because of varying probability of spending a night outside in a given month ( $M_t$  models). Second, we considered an open population with HRP individuals migrating in and out of the population depending on whether or not they spent a night outside in a given month. Both models estimated the same PSE, i.e the total population size of the forest-going HRP in the study area during the study period, which covers 1 full year between December 2017 and November 2018. As noted by Pollock<sup>113</sup>, one of the founders of the capture-recapture methodology, the

distinction between open and closed population may be artificial and mainly resides in their aptitude to estimate different parameters. Closed population models focus on estimating capture probabilities whereas open population models focus on estimating migration rates.

Baillargeon and Rivest<sup>108</sup> provide a nice illustration of the log-linear models fitting process with the simplest model for a closed population  $M_0$ . For a dataset with  $t$  capture occasions,  $2^t - 1$  capture histories,  $\omega$ , are observables. Again, with  $t = 5$ , 01011 is the capture history for an individual captured on occasions 2, 4 and 5 but not on occasions 1 and 3. In  $M_0$ , which has a single capture probability  $p$ , the probability for an individual to experience a capture history  $\omega$  is

$$P(\omega) = p^{\sum \omega_j} (1 - p)^{t - \sum \omega_j} \quad (\text{Eq 2.2})$$

where  $\sum \omega_j$  is the number of times the individual is captured. Therefore, the expected number of units in the population with capture history  $\omega$  is given by:

$$\begin{aligned} \mu_\omega &= N \times P(\omega) \\ &= N p^{\sum \omega_j} (1 - p)^{t - \sum \omega_j} \\ &= \exp(\log(N(1 - p)^t) + \sum \omega_j \log\left(\frac{p}{1 - p}\right)) \end{aligned} \quad (\text{Eq 2.3})$$

The log-linear model therefore fits  $E[Y] = \exp(\alpha + X\beta)$ , where  $Y$  is the vector of observed capture history frequencies and  $X$  is a vector defined by  $\sum \omega_j$ . Then, the total population size is estimated as  $\hat{N} = n + \exp(\hat{\alpha})$  where  $n$  is the total number of units captured in the data. This is because:

$$\exp(\alpha) = \exp(\log(N(1 - p)^t)) = N(1 - p)^t = N \times P(\omega_0) = \mu_0 \quad (\text{Eq 2.4})$$

where  $\omega_0$  is the unobservable capture history of zero captures and  $\mu_0$  is the expected number of units never captured. Estimation is done using maximum likelihood for a Poisson count random variable for the number of individuals with a certain capture history. See Rivest and Baillargeon<sup>112</sup> for more details on other log-linear models.

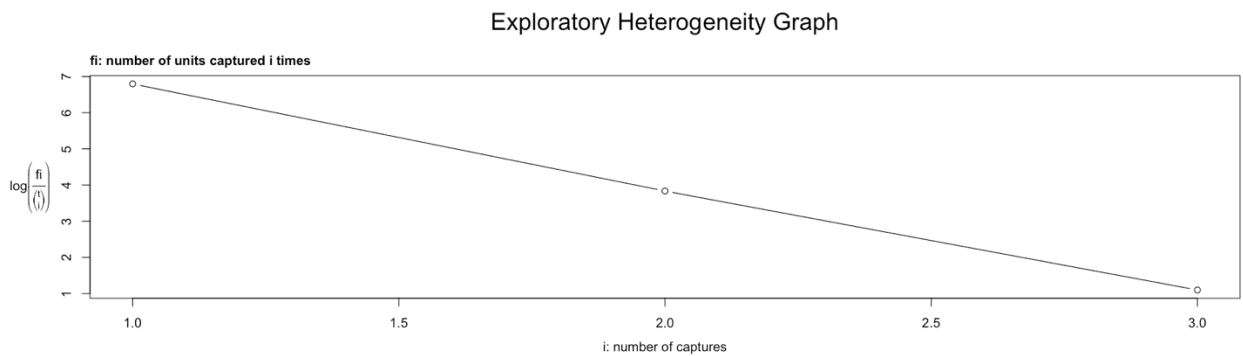
Finally, Table 2.12 shows how the expected number of individuals with a certain capture history is parametrized in log-linear models for closed and open population models<sup>104</sup>. For simplicity, three (instead of four in our context) source models are shown.

**Table 2.12** - Parametrization for the expected number of individuals with certain capture histories for various three source capture-recapture models. The probability of being captured in a survey is allowed to vary across the 3 surveys as  $p_1, p_2$  and  $p_3$  but is constant in  $M_0$  model. Other parameters include  $\phi_i$ , the probability that an individual survives from the  $i^{\text{th}}$  to  $(i+1)^{\text{th}}$  sample;  $\chi_i$ , the probability that an individual is not seen after the  $i^{\text{th}}$  sample;  $1/\psi_i$ , the probability that an individual alive and unmarked in the population at the time of the  $(i+1)^{\text{th}}$  sample was in the population at the time of the  $i^{\text{th}}$  sample; and  $1/\lambda_i$ , the probability that an individual alive in the population at the time of the  $i^{\text{th}}$  sample but not observed thereafter, is still alive in the population at the time of the  $(i+1)^{\text{th}}$ . Note the substitution  $\lambda_i = \chi_i / (\phi_i(1 - p_{i+1})\chi_{i+1})$ .

Capture history	Expected number of individuals with certain capture histories for various models		
	Closed population (M0)	Closed population with temporal dependence (Mt)	Open population
111	$N \times p \times p \times p$	$N \times p_1 \times p_2 \times p_3$	$N \times p_1 \times \phi_1 p_2 \times (1 - \chi_2)$
011	$N \times (1 - p) \times p \times p$	$N \times (1 - p_1) \times p_2 \times p_3$	$N \times (1 - p_1) \times \phi_1 \psi_1 p_2 \times (1 - \chi_2)$
101	$N \times p \times (1 - p) \times p$	$N \times p_1 \times (1 - p_2) \times p_3$	$N \times p_1 \times \phi_1 (1 - p_2) \times (1 - \chi_2)$
001	$N \times (1 - p) \times (1 - p) \times p$	$N \times (1 - p_1) \times (1 - p_2) \times p_3$	$N \times (1 - p_1) \times \phi_1 \psi_1 (1 - p_2) \times \psi_2 (1 - \chi_2)$
110	$N \times p \times p \times (1 - p)$	$N \times p_1 \times p_2 \times (1 - p_3)$	$N \times p_1 \times \phi_1 p_2 \times \chi_2$
010	$N \times (1 - p) \times p \times (1 - p)$	$N \times (1 - p_1) \times p_2 \times (1 - p_3)$	$N \times (1 - p_1) \times \phi_1 \psi_1 p_2 \times \chi_2$
100	$N \times p \times (1 - p) \times (1 - p)$	$N \times p_1 \times (1 - p_2) \times (1 - p_3)$	$N \times p_1 \times \lambda_1 \phi_1 * (1 - p_2) \times \chi_2$

## 2.6.9. S2.9: Diagnostic test for heterogeneity in log-linear models

Figure 2.8 shows a diagnostic test confirming heterogeneity between individuals in terms of their capture probability does not need to be included in the model. Work from Lindsey (1986) and Rivest (2007) shows that this plot for  $f_i$ , the number of units captured on  $i$  different occasions, should be concave upward in the presence of heterogeneity. On the other hand, linearity indicates heterogeneity does not need to be accounted for in the model.



**Figure 2.8** - Diagnostic test for heterogeneity: linearity indicates that individual heterogeneity in terms of their capture probability is not needed in the models.

## 2.6.10. S2.10: subdividing the PSE by type of risk

### behavior/activity

Based on individuals' responses to the criteria in Table 2.1, we split the definition of HRP based upon forest-related activities or agriculture-related activities. For instance, in this secondary analysis, criterion A was split between the following 2 sub-criteria:

- A\_Agriculture: During the past month, stayed overnight away from home AND reason for the absence was working in the *rice field or plantation* in this province or another province
- A\_Forest: During the past month, stayed overnight away from home AND reason for the absence was working in the *forest* in this province or another province

Criterion C could not be split between forest and agriculture activities because of how the question was framed. In a sensitivity analysis, individuals meeting criterion C were all allocated to either the forest sub-category or the agriculture sub-category, producing liberal and conservative estimates for the population size of HRP in the two sub-categories. Also note that criterion F only pertains to HRP individuals being identified because of their forest activities. Last, the FTAT HRP criteria could not be split either because of the eligibility criteria in the FTAT intervention.

Table 2.13 and 2.14 show the results from the household survey methods when splitting our definition of HRP between those that were identified as HRP because of forest-related activities or because of agriculture-related activities. Results from the sensitivity analysis are also reported.



**Table 2.13** - Results for the population-based survey method for population size estimation of agriculture-related HRP individuals.

	Estimate	% HRP [95% CI]	PSE [95% CI]
Baseline	HRP Agriculture (Conservative)	7.2 [5.9; 8.4]	2,164 [1,794; 2,539]
Baseline	HRP Agriculture (Liberal)	15.6 [14.7; 17.1]	4,702 [4,246; 5,156]
MTAT	HRP Agriculture	6.6 [4.8; 8.5]	2,006 [1,436; 2,568]
Endline	HRP Agriculture (Conservative)	3.8 [3.0; 4.6]	1,147 [910; 1,385]
Endline	HRP Agriculture (Liberal)	5.2 [4.3; 6.1]	1,569 [1,308; 1,821]

**Table 2.14** - Results for the population-based survey method for population size estimation of forest-related HRP individuals.

	Estimate	% HRP [95% CI]	PSE [95% CI]
Baseline	HRP Forest (Conservative)	4.9 [4.1; 5.8]	1,496 [1,238; 1,756]
Baseline	HRP Forest (Liberal)	13.8 [12.4; 15.2]	4,168 [3,743; 4,597]
MTAT	HRP Forest	9.0 [7.0; 11.0]	2,727 [2,124; 3,338]
Endline	HRP Forest (Conservative)	0.3 [0.2; 0.5]	100 [50; 152]
Endline	HRP Forest (Liberal)	4.3 [3.5; 5.1]	1,302 [1,058; 1,550]

In conclusion, the wording of surveys questions did not allow for a clear distinction in HRP criteria and the range of estimates was too wide to be informative.

# **Chapter 3: Characterizing mobility patterns of forest goers in southern Lao PDR using GPS loggers**

Francois Rerolle, Emily Dantzer, Toula Phimmakong,  
Andrew Lover, Bouasy Hongvanthong, Rattanaxay Phetsouvanh,  
John Marshall, Hugh Sturrock, Adam Bennett

### **3.1. Abstract**

In the Greater Mekong Sub-region (GMS), engaging in forest activities is a major risk factor for malaria. As countries focus their malaria control and elimination efforts on forest-going populations, a better understanding of their mobility patterns and risk associated with specific types of forest-going trips is essential.

In 2018, we conducted a focal test and treat intervention (FTAT) in Champasak Province, southern Lao PDR, and recruited 2,904 forest-goers in our study. A subset of forest-goers were offered to carry a “i-Got-U” GPS logger for roughly two months, configured to collect GPS coordinates every 15 to 30 minutes. The utilization distribution (UD) surface around each GPS trajectory was used to extract trips to the forest and forest-fringes. A hierarchical clustering algorithm identified trips with shared mobility pattern characteristics in terms of duration, timing of the trip and forest penetration further enabling classification of high-risk trips because of an increased exposure to dominant malaria vectors in the region. Finally, we used gradient boosting trees to assess which of the forest-goers’ socio-demographic and behavioral characteristics predicted the best their likelihood to engage in trips at higher-risk for malaria.

A total of 122 forest-goers accepted to carry a GPS logger resulting in the collection of 803 trips to the forest or forest-fringes. Six clusters of trips emerged, helping to identify 385 (48%) trips with increased exposure to malaria vectors based on high forest penetration and whether the trip happened overnight. Age, outdoor sleeping structures and number of children were the best predictors of forest-goers’ probability to engage in high-risk trips. The probability to engage in high-risk trips remained high (~33%) in all strata of the forest-going population.

This study characterized the heterogeneity within the mobility patterns of forest-goers and attempted to further segment their role in malaria transmission in southern Lao People's Democratic Republic (PDR). These results are key for national control programs across the region to assess and meet their 2030 malaria elimination goals.

## 3.2. Introduction

The dominant malaria vectors in the Greater Mekong Sub-region (GMS) – *Anopheles dirus* and *Anopheles minimus* – are forest-dwelling mosquitoes<sup>13,14</sup> and malaria transmission in the region is often referenced as “forest malaria”<sup>15</sup>. Forest-going activities, such as wood collection and spending the night in the forest, as well as agriculture in the forest fringe areas<sup>20,62,86</sup>, are major risk factors for malaria in the GMS<sup>16,17,22,25,26,87–89</sup>. Furthermore, recent outbreaks have been attributed to deforestation activities<sup>29,114</sup>. As transmission declines in the region, malaria clusters in these key forest-going populations and national control programs in the GMS now increasingly focus their prevention and treatment efforts on forest-goers<sup>23,24</sup>. Yet, much remains unknown about forest-goers’ mobility patterns and their actual whereabouts in the surrounding forest.

As pointed out in a literature review on malaria and population mobility<sup>64</sup>, population movement is often cited as a barrier to malaria elimination, but there have been very few studies to support the evidence. The authors argued that it resulted in an excessive focus on “mobile populations” as a risk group and encouraged malaria programs to refocus their efforts on mobility itself. Similarly, “forest-going populations”, who often also belong to “mobile populations”<sup>23–25</sup>, is a catch-all term that encompasses a wide range of different risk behaviors<sup>28,62</sup>, and “going in the forest” first needs to be better defined.

Micro-scale movement data of forest-goers is essential to understand their role in the transmission of forest malaria in the GMS. Heterogeneity in mobility patterns likely results in diverse exposures to mosquito vectors and heterogeneous risks for malaria. For instance,

individuals who travel through the forest for days at a time are likely to play a different role in malaria transmission than individuals who cross the forest to reach their rice field everyday but return home every night. Data on forest-goers' mobility patterns could also be leveraged to better access these population if geographical or temporal bottlenecks can be identified.

The recent advent of portable global positioning system (GPS) logging devices offers unprecedented opportunities to collect fine-scale mobility data on these populations and characterize their movements in and out of the forest. These GPS loggers can provide high resolution data both spatially and temporally and have shown high acceptability in rural settings<sup>115-117</sup>. In previous studies, such devices have successfully been used to assess the importance of individual movement data on the transmission of multiple diseases such as dengue, schistosomiasis, hookworm or filariasis<sup>118-121</sup> but also malaria<sup>122-124</sup>.

In this analysis, we collected fine-scale movement data from forest-goers recruited in a focal test and treat (FTAT) intervention conducted in southern Lao People's Democratic Republic (PDR). To our knowledge, this study is the first to describe the mobility patterns of forest-going populations in the GMS using GPS loggers. We conducted a clustering analysis to characterize the heterogeneity within these mobility patterns and a regression analysis to attempt to further segment forest-going populations in terms of their potential exposure to malaria vectors.

### **3.3. Methods**

#### **3.3.1. Study area**

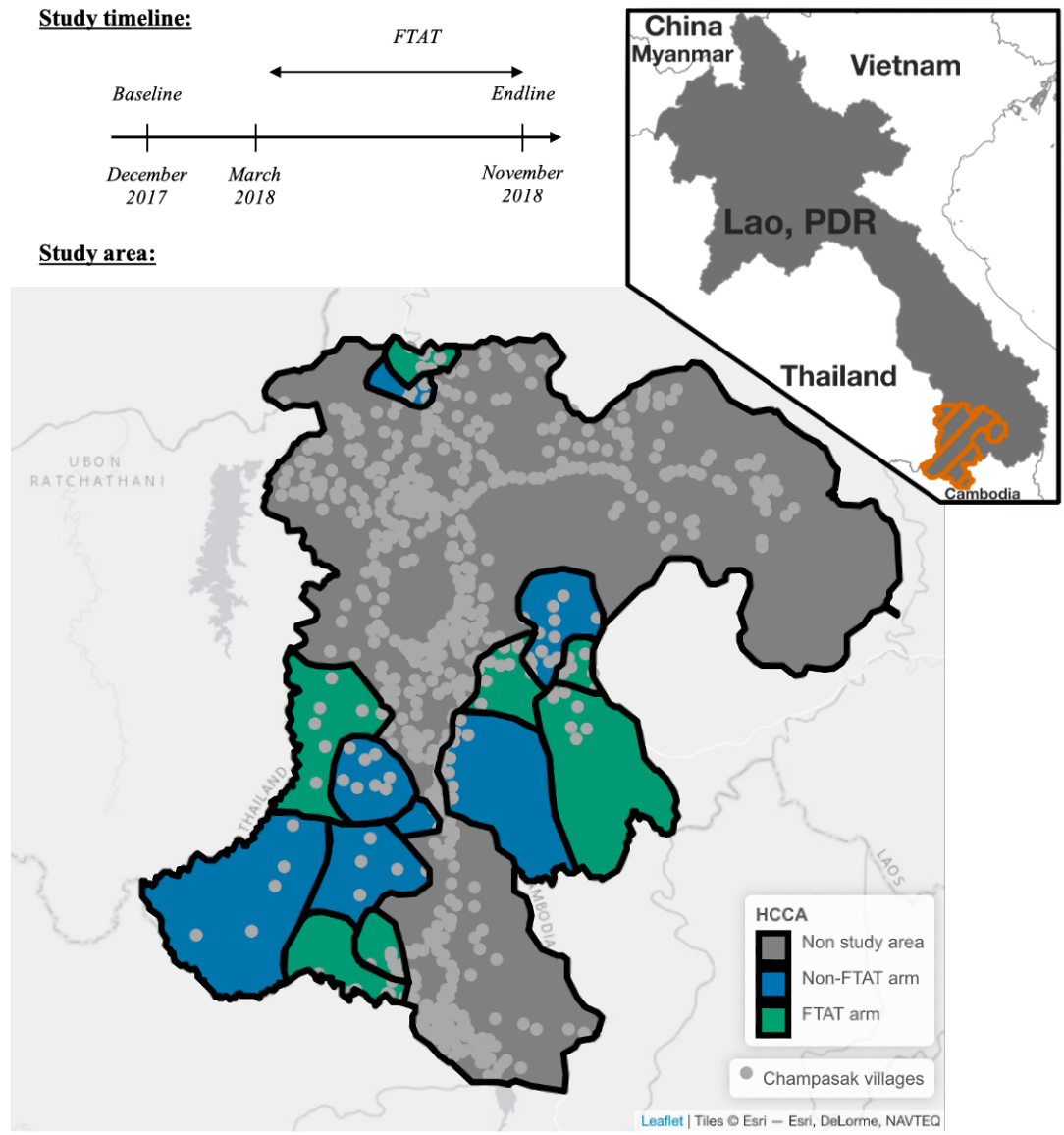
This study was conducted in Champasak Province, one of the five southernmost provinces in Lao PDR, where 95% of the country's malaria burden concentrates<sup>71</sup>. The data were collected as part of a randomized controlled trial designed to evaluate the effectiveness of forest-based active case detection<sup>66</sup>. Across four districts, seven of 14 health center catchment areas (HCCA) were randomly assigned to a Focal Test-And-Treat (FTAT) arm, an intervention specifically targeting forest-goers and conducted continuously between March and November 2018. The study area was selected in consultation with the national malaria program based on malaria burden (highest API in 2016). Figure 3.1 shows the study timeline and a map of the study area.

#### **3.3.2. Data sources**

##### ***3.3.2.1. FTAT survey***

Fifteen teams of two peer navigators (PNs) were employed in FTAT HCCAs to conduct test-and-treat activities amongst members of their communities presumed to be “forest-goers” because of their activities in or near the forest. PNs were themselves forest-goers recruited from the local communities via health authorities and trained to conduct continuous surveillance by testing for malaria using Rapid Diagnostic Tests (RDTs)<sup>66</sup>. PNs were instructed to actively target forest-goers in forest fringe areas and to enroll anyone meeting the FTAT eligibility criteria: aged 15 or older and having spent at least one night outside a formal village in the past 30 days.

Upon recruitment of forest-goers in FTAT, PNs conducted an epidemiological survey covering the demographic, behavioral, occupational, malaria knowledge and practice domains. To understand the mobility patterns of this population of forest-goers, PNs offered a subset of them, conveniently sampled, to carry a GPS logger that would record GPS coordinates as they carry it.



**Figure 3.1** - Top left: Study timeline with a rolling FTAT survey between March and November 2018. Bottom: Study area with 7 of 14 health center catchment areas (HCCA) randomly assigned to FTAT. The study was conducted in Chamapasak province in southern Lao PDR neighboring Thailand and Cambodia (see upper right indent).



### **3.3.2.2. *GPS data***

In May, 53 GPS loggers (I-gotU 120) were dispatched across the 15 PN teams to be offered to interviewed forest-goers and carried for about two months. During that first cycle, loggers were configured to collect GPS coordinates every 30 minutes and were retrieved in July/August by the PN teams for data downloading. A second cycle of data collection was started in September with 69 GPS loggers configured to collect GPS coordinates every 15 minutes. Loggers were retrieved in November for data downloading. Recruiting PNs teams also carried GPS loggers, configured to collect GPS coordinates every 30 minutes over the two cycles.

In order to simplify instructions, the loggers were configured so that they could not be turned off by forest-goers or PNs and the logging intervals selected, 15 to 30 minutes, afforded an estimated 7 to 12 days of battery life. Loggers could be charged on outlets with regular phone chargers, which most forest-goers possessed. Yet, to avoid battery depletion while on forest trips or off the grid, external charging devices (Verbatim®) and two sets of four individual AA lithium batteries were additionally provided to recruited forest-goers. Forest-goers were instructed to carry the GPS loggers at all times, to frequently charge them (at least once a week) and to meet again after two months for GPS loggers' retrieval. PNs demonstrated all aspects of the GPS loggers' utilization, including charging, to recruited forest-goers.

### **3.3.2.3. *GPS logger retrieval questionnaire***

After roughly two months, PNs met again with forest-goers to collect the GPS loggers in exchange for a \$10 monetary incentive. Upon retrieval, a short questionnaire was conducted to assess acceptability and feasibility of using GPS loggers to record mobility patterns of forest-

goers. In particular, the survey asked about forest-goers' charging practices and logger utilization over the two-month study period.

### **3.3.3. GPS data processing**

#### ***3.3.3.1. Data cleaning***

The advertised precision of the I-gotU GPS loggers used in this study is 10m. Yet, the makers warn of possible large errors in the GPS coordinates collected, notably when the logger stay indoor for long periods of time and cannot connect with the satellites. To remove those erroneous GPS points, we used a filtering algorithm that identifies GPS points unusually far away from both the previous and next GPS points. See supplemental materials S3.1 in appendix 3 for details.

#### ***3.3.3.2. Significant locations***

The data collected by a GPS logger is a time series of GPS points forming a trajectory (Figure 3.2A). If several GPS points cluster together, it indicates a location visited frequently or for long periods of time by the HRP carrying the GPS logger (or a location where the GPS logger was left behind). Using a method developed by Barraquand and Benhamou<sup>125</sup> and implemented in the *adehabitatLT*<sup>126</sup> package in R<sup>127</sup>, we computed the residence time spent within a moving 50m-radius circle window centered on every GPS point of the trajectory. Then, we used the biased random bridge kernel method<sup>128</sup> implemented in the *adehabitatHR*<sup>126</sup> R<sup>127</sup> package, to estimate the utilization distribution (UD) 30m per 30m surface around the trajectory. The UD is a concept widely used in animal movement ecology that measures the utilization of space via the intensity of the GPS points occurrence on the map. A significant location was defined as a 100m-radius

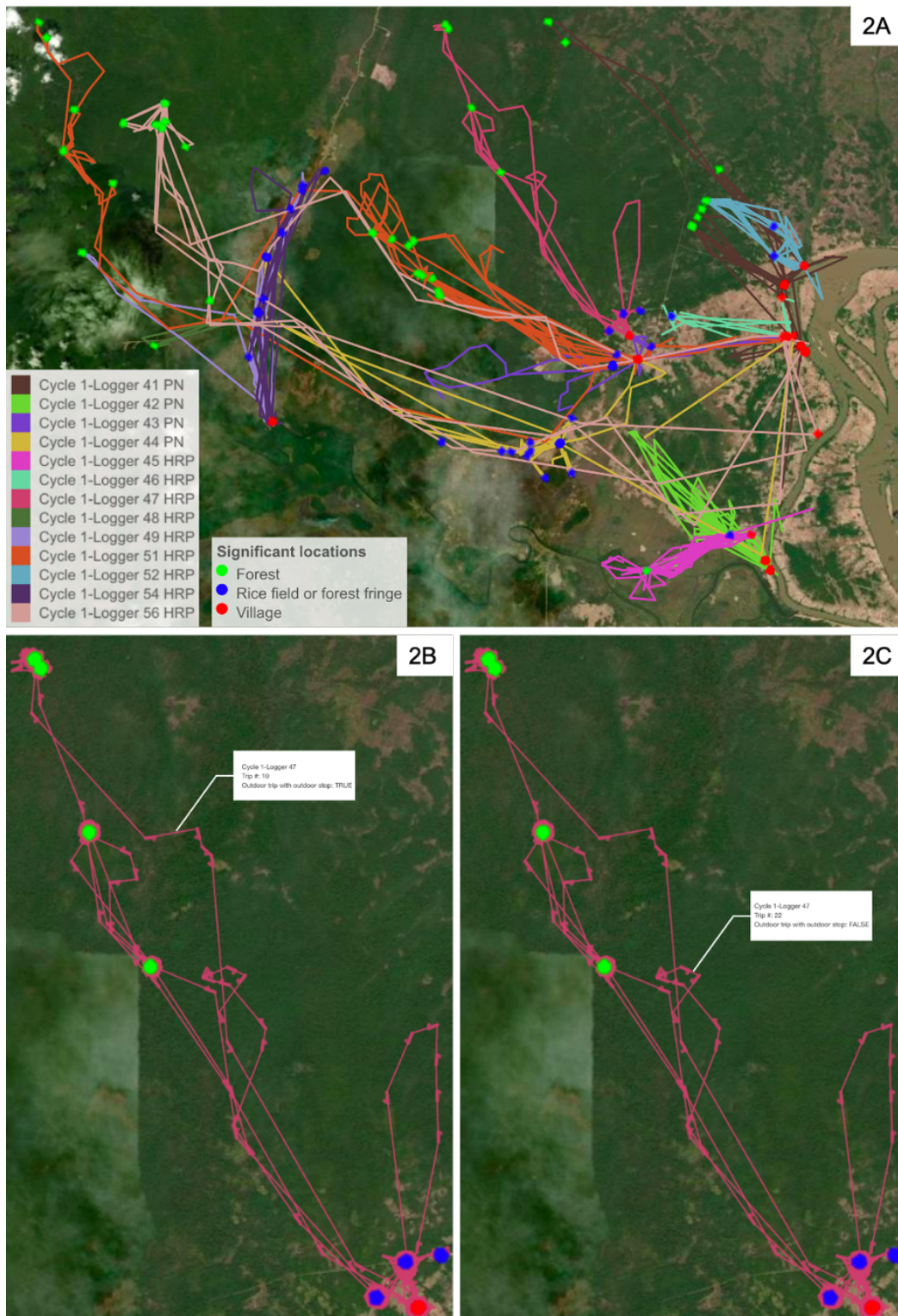
circle centered on a local maximum of the UD surface that contains at least one GPS point of the trajectory with a residence time above 2h. Simply put, a significant location is a 100m-radius circle where the GPS logger stayed for more than 2 hours at least once along the trajectory.

Significant locations were mapped on top of earth terrain layers, using ESRI imagery in the leaflet R package, along with the GPS tracks and classified as forest, forest-fringe/rice field or village-based locations by visual inspection. Residence time at village-based significant location as well as self-reported home village by forest-goers in the FTAT questionnaire were additionally used to identify forest goers' house location. Finally, we used PNs' GPS tracks as well as their self-reported home village (from employment contracts) to identify significant locations related to our study such as follow-up meetings at PNs' house. GPS coordinates of forest-goers' and PNs' home village were extracted from a list of geo-referenced villages in the province provided by the national malaria control program.

### ***3.3.3.3. Outdoor trips***

A trip was defined as a series of consecutive GPS points outside forest-goers' house, between two GPS points recorded at forest-goers' house location. Every trip including GPS points that formed an outdoor-based significant location (forest or forest-fringe/rice field) qualified as an outdoor trip (Figure 3.2B). Trips where a forest-goer tours the forest for hours but never really stops in a single location long enough (Figure 3.2C) should also be classified as outdoor trips. To identify those other outdoor trips, we first learned the relationship between our classification of outdoor vs village-based significant location and the following covariates using a random forest algorithm: number of Open Street Map<sup>82</sup> buildings or places, total 2015 population and average

2018 tree crown cover within 100m and distance to closest village in the province. Tree crown cover layers came from Hansen<sup>60</sup> and population from WorldPop<sup>129</sup>. We then used the predicting algorithm to classify non-significant location GPS points as outdoor or village-based. Finally, outdoor trips were defined as trips that include an outdoor-based significant location or a series of consecutive GPS points adding up to more than two hours outdoor. Simply put, an outdoor trip is a trip where the forest-goer spent more than two hours consecutively outdoor. Trips including a significant location related to our study were discarded as unrepresentative of the forest-goers' routine.



**Figure 3.2** - Trajectories for GPS loggers collected during the first cycle in Moonlapamok district for PNs and HRPs (High-risk populations), i.e forest-goers (2A). Figures 2B and 2B respectively show examples of an outdoor trip with an outdoor stop (trip #10 for GPS logger 47 in 2B) and without (trip #22 for GPS logger 47 in 2C). Significant stop locations are shown as circles, colored according to their terrain class. Directional arrows were added on top of each GPS points in 2B and 2C to represent the movement flow.

### 3.3.4. Cluster analysis

For each outdoor trip, we computed the mobility pattern characteristics listed in Table 3.1. They were selected to translate the GPS trajectories in terms of exposure to the dominant malaria vectors in the GMS, *An. dirus* and *An. minimus*<sup>13,14</sup>. Four domains were covered. Two domains, forest surroundings and timing of the trips, pertained directly to the ecology of these mosquitoes, which thrive in a forested environment and bite during nighttime and around twilight and dawn hours (6 pm and 6 am). The two other domains, pace and fragmentation of the trips, reflect the possible organization and habits of those trips and can influence vector control options. For instance, it may be easier to carry hammocks bed nets over short distances and trips with numerous and frequently visited stop locations may offer higher than average mosquito protection such as forest huts.

**Table 3.1** - *Mobility patterns variables computed for each of the outdoor trips and used as features in the clustering algorithm (after normalization, standardization and projection onto the principal components).*

Domain	Forest	Pace	Fragmentation	Timing
Variables	Average 2018 tree crown cover	Duration	Number of different significant location	Overnight trip
	Max 2018 tree crown cover	Distance	Proportion of trip spent at significant location	Trip around twilight and/or dawn hours (6 am and/or 6 pm)
	Proportion of trip where 2018 tree crown cover > 50%	Max speed	Population density	

Variables in Table 3.1 were standardized by subtracting the mean and dividing by the standard deviation and right-skewed variables (pace and population density) were log-transformed. Then, we used principal component analysis to project the variables onto the principal components (PC) that captured 95% of the variability in the dataset. Then, hierarchical clustering with the complete distance method, was applied on the selected PCs to explore the clustering structure of

the data. The hierarchical clustering algorithm starts with one observation per “leaf” (=cluster) and progressively groups similar observations together one at a time until they are all grouped together in a single cluster. An advantage of hierarchical clustering over other clustering algorithm such as k-means, is that the number of desired clusters,  $k$ , does not need to be set in advance. Instead, the resulting dendrogram tree represents the clustering structure for all  $k$  from 1 to  $n$ , the number of observations. The length of the tree branches quantifies the dissimilarity between the leaves and can be used to assess how many clusters should represent the structure of the data. The intra-class correlation coefficient (ICC) for input variables in Table 3.1 was also computed for different choices of  $k$  to evaluate how many clusters would best capture the variability in the dataset.

Finally, mobility pattern characteristics in Table 3.1 were summarized for each of the clusters identified and plotted to expose the heterogeneity between the clusters, describe their distributions across the trips and attempt to label the type of trips identified in each of the clusters.

### **3.3.5. Regression analysis**

Nighttime outdoor trips in clusters with high forest penetration translate in an increased exposure to malaria vectors and were classified as “high-risk” trips. Then, we used gradient boosting trees to assess which of the forest-goers’ socio-demographics and behavioral characteristics collected in the FTAT survey best predicted their likelihood to engage in such high-risk trips for malaria. Gradient boosting was selected as one of the most advanced supervised learning algorithms that can accommodate missing values and model non-linearities. Importantly, its implementation in

the GPboost<sup>130</sup> R<sup>127</sup> package allows for random effects at forest-goers' levels to correctly account for the correlation structure with multiple outdoor trips per forest-goers. Automated grid search and 4-fold cross validation were used to select the best fitting tuning parameters.

Results are presented using SHAP (SHapley Additive exPlanations) values<sup>131</sup>, an innovative tool increasingly used for interpretation of machine learning models. SHAP values attribute for each feature and each prediction, importance values. It enables to rank the different features in their ability to predict the outcome but also to visualize the adjusted non-linear relationship between the predictors and the outcome.



## **3.4. Results**

### **3.4.1. Data description**

#### **3.4.1.1. *FTAT survey***

Over the course of 8 months, 2,904 forest-goers were recruited in FTAT and 122 carried a GPS logger. Using their answers in the FTAT survey, Table 3.2 shows how forest-goers recruited in the GPS component of the study differed from those that did not carry a GPS logger. Overall, the two groups were similar although some differences emerged. Forest-goers that carried a GPS logger were older (39.2 vs 36.4 years) and tended to travel in smaller groups (3 vs 4) and for less nights (4.1 vs 7.2) than the forest-goers that did not carry a GPS logger. They were also more likely to be male (95% vs 65%), to report forest work as their primary activity (46% vs 28%) and no sleeping structure in the previous night (51% vs 30%) than the forest-goers that did not carry a GPS logger.

**Table 3.2** - Comparison between forest-goers that carried a GPS logger and those that did not in terms of their answers to FTAT variables.

FTAT variable	Mean among HRP that		p-value
	Carried a GPS logger	Did not carry a GPS logger	
Number of forest-goers in group	3	4.14	< 0.01
Age in years	39.2	36.36	0.01
Number of children	1.79	1.63	0.24
Nights away from home on trip	4.12	7.36	< 0.01
Km away from home	6.63	7.58	0.38
Number of people working/traveling with on trip	2.7	3.62	< 0.01
Ever spent night in forest in rainy season	0.92	0.9	0.62
Ever spent night in forest in dry season	0.94	0.89	0.14
Ethnic minorities	0.07	0.1	0.47
Married	0.87	0.8	0.11
Rice farming is main source of income	0.89	0.92	0.24
Male	0.95	0.65	< 0.01
Education less than primary school	0.43	0.49	0.3
Wood collection is primary reason to visit forest in rainy season	0.32	0.34	0.67
Wood collection is primary reason to visit forest in dry season	0.43	0.48	0.4
Forest work is primary activity this week	0.46	0.28	< 0.01
Motorized main mode of transportation	0.69	0.68	0.81
Relationship to people on trip is family	0.6	0.63	0.55
No sleeping structure last night	0.51	0.3	< 0.01

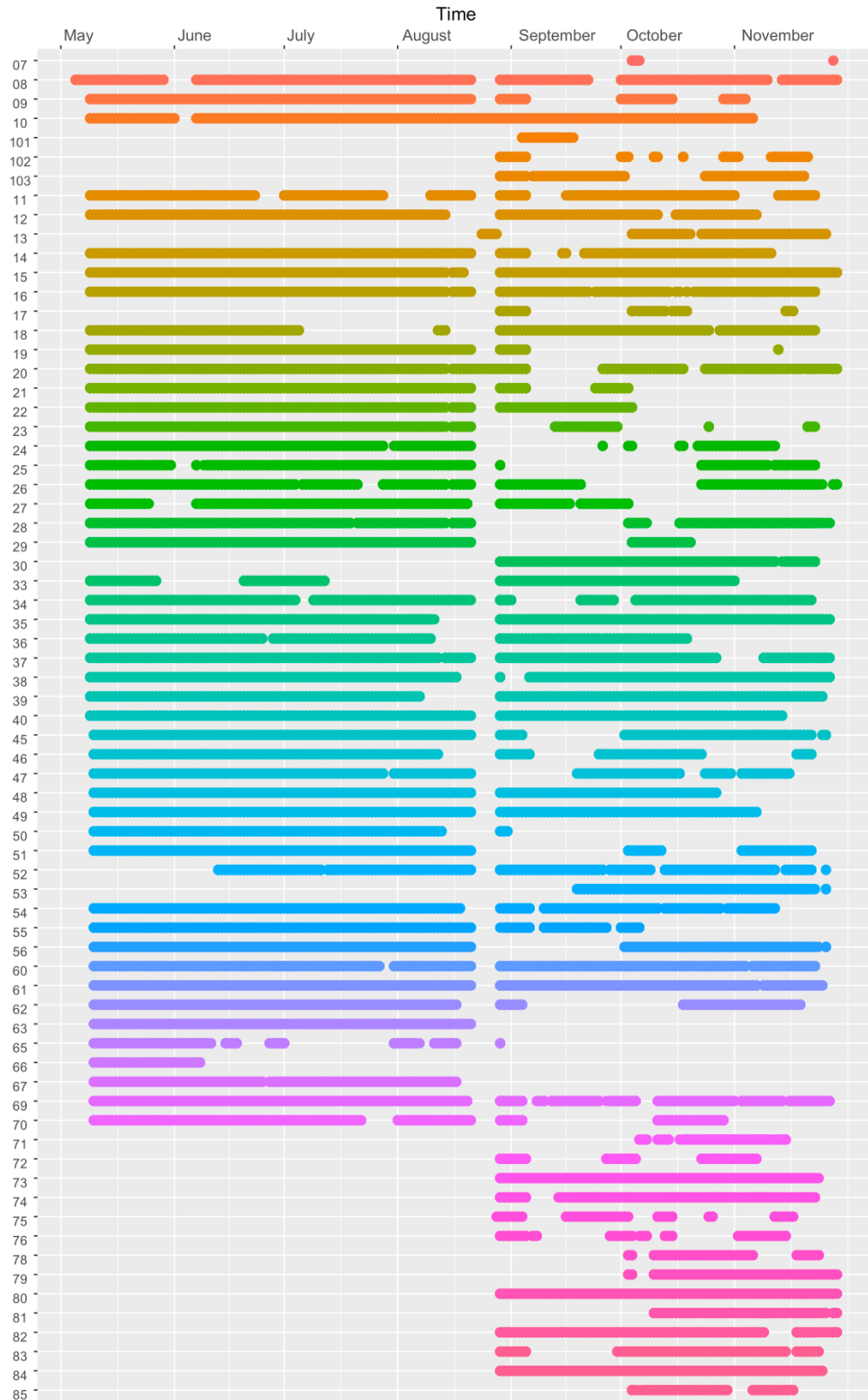
### 3.4.1.2. GPS data

Two (1.6%) GPS loggers were not returned and data downloading from 5 (4.2%) others failed, resulting in a total of 472,751 GPS points collected from 115 (94.2%) GPS loggers. Figure 3.3 shows time series of when GPS coordinates were collected for each of the loggers. The plot only shows a few gaps in the time series indicating that the forest-goers generally kept their GPS loggers charged. We can see the clean demarcation between the two cycles of data collection at the end of August where the loggers were with the field team for data download and configuration. For the first cycle, on the left-hand side of the plot, there are almost no data gaps. This motivated us to decrease the logging interval from 30 min to 15 min in the second cycle,

which resulted in more gaps. Also note that an additional 15 GPS loggers were purchased and handed out in the second cycle.

Data visualization exposed a few GPS points obviously logged incorrectly and our filtering algorithm discarded 1,973 (0.4%) data points. Most of the time, these errors occurred while the GPS logger was sitting at forest-goers' house location, most likely beneath some type of roof that disabled connection with the GPS satellites.

Plotting the GPS trajectories also highlighted that forest-goers did not always carry their GPS logger with them. Indeed, some GPS loggers obviously were left at home for weeks at a time. The incentive to give the GPS logger back to the study team after two months may have discouraged forest-goers to take the risk to carry them all the time. Importantly, our instructions insisted primarily on the importance of accurately recording trips to the forest, forest-fringes and rice fields. That is why we decided to focus our analysis on outdoor trips rather than on the whole mobility patterns over the two-month study period. In the process, our analysis discarded 95% of the GPS points to focus on the 21,668 (5%) collected along 803 outdoor trips from 96 (79%) forest-goers. The out of the bag (OOB) error rate for our terrain classification algorithm trained on 1,068 significant locations was 8.6%.



**Figure 3.3** - Time series plot of when GPS loggers were on and collected GPS coordinates. One row per GPS logger. Gaps indicate times when loggers ran out of battery.

### 3.4.1.3. *GPS logger retrieval questionnaire*

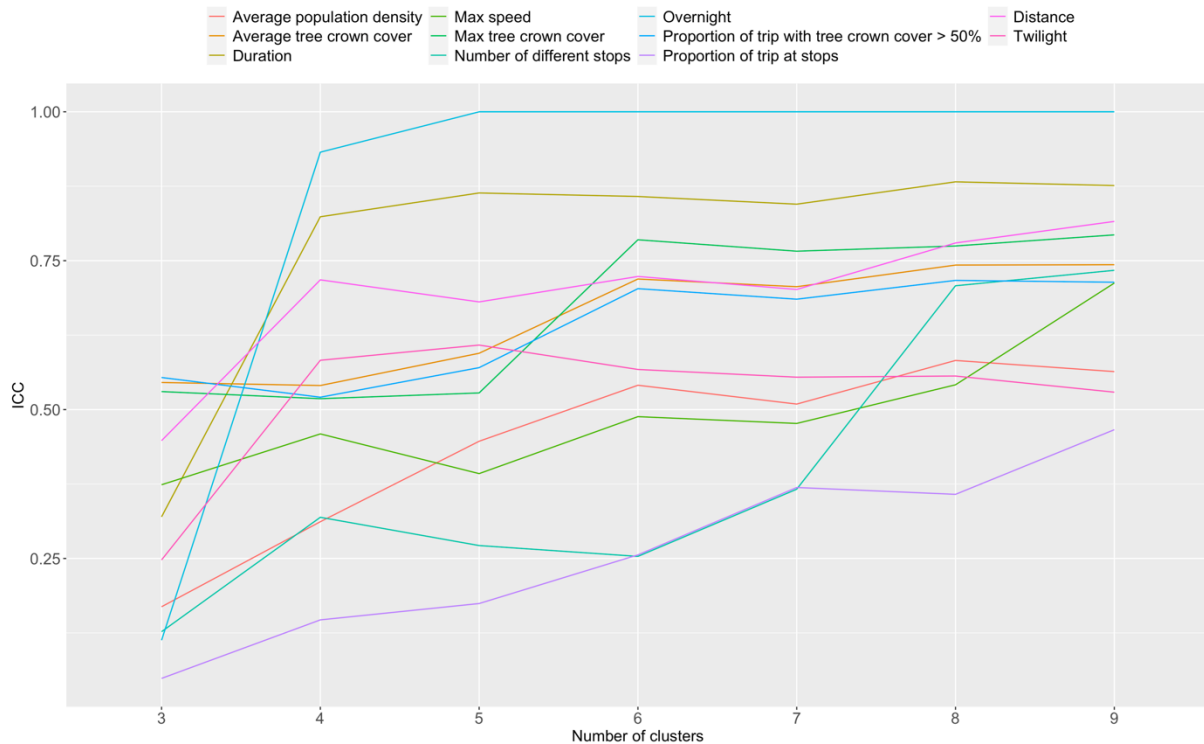
Table 3.3 summarizes forest-goers' answers to the retrieval questionnaire conducted when they gave the GPS logger back to the study team. The majority (93.3%) of forest-goers respected the instructions to charge their GPS logger at least once a week. According to the forest-goers, their GPS logger ran out of battery fairly rarely, with 77.5% of them never doing so. Surprisingly, 61.7% of forest-goers shared their GPS logger with another household member, although that happened mostly (80%) for no more than a few days only. Only 39.3% of the forest-goers reported carrying their GPS logger every day, which supports our decision to restrict our analysis to outdoor trips only. In terms of acceptability, the field team also reported informally that most forest-goers accepted to carry a GPS logger when offered, with only a few refusals.

**Table 3.3** - *GPS logger self-reported utilization from retrieval questionnaire after forest-goers gave back their GPS logger. N =120.*

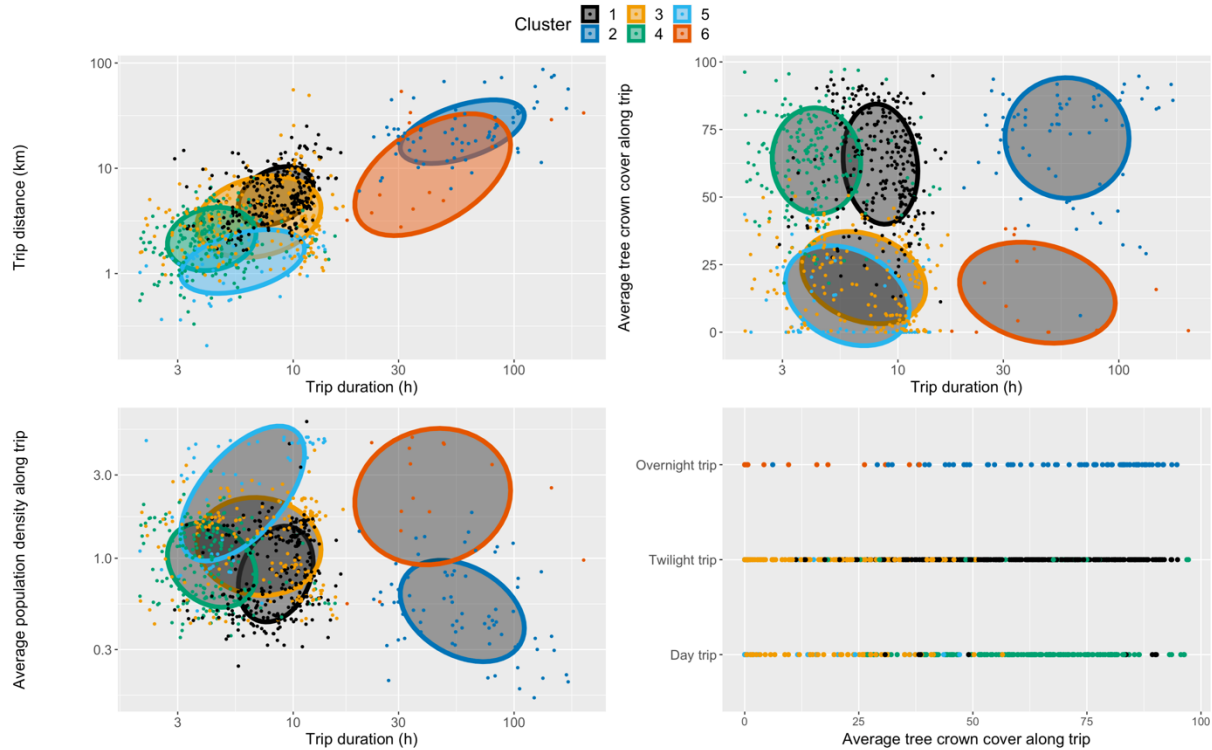
<b>Variable</b>	<b>Levels</b>	<b>%</b>
GPS logger ran out of battery	Never	77.5
	1-4 times	15.5
	More than 5 times	7
Charging practice	At least once a week	93.3
	Less than once a week	6.7
Carried GPS	Every day	39.3
	Most of the time	58.1
	Rarely	2.6
Anyone else carried logger	Yes	61.7
Who else	Household member	100
For how long	A few hours	24
	A few days	56
	A few weeks	20

### 3.4.2. Cluster analysis

The first seven PC accounted for 96% of the variability in the data and were therefore extracted to summarize the outdoor trips data. The dendrogram tree (Figure 3.9 in the appendix), resulting from the hierarchical clustering algorithm, is well-balanced and the distribution of large branches suggest cutting down the tree with 6 clusters (horizontal red line). To support our decision, we also looked at how the ICC for mobility variables in Table 3.1 evolved as the number of selected clusters varied. For most of these variables, Figure 3.4 shows an improvement in the ICC all the way until 5, 6 clusters but then levels off. Our interpretation of these plots oriented us to select 6 clusters to summarize the outdoor trips data.



**Figure 3.4** - Plot of how the ICC for mobility patterns variables in Table 3.1 vary with the number of clusters selected. Except for the proportion of trips at stops, the number of different stops and max speed whose ICC continue to improve beyond 9 clusters, for most variables, the ICC increases up to 5 or 6 clusters and then levels off.



**Figure 3.5** - Bi-plots of the clustering structure in the feature space. Points are colored by cluster assignment and ellipse capturing 50% of the clusters' points, assuming bivariate normal distribution, are superimposed. Features represented were selected for their ability to separate the data and highlight the clustering structure of the data.

Figure 3.5 presents biplots of the resulting clustering structure in the feature space. In combination with Table 3.4, where each of the input mobility variables is summarized by clusters, we can attempt to label the 6 types of clusters identified. For instance, the darkblue dots of cluster 2 corresponds to outdoor trips with high forest penetration and that lasted overnight. As a result, we propose to label this cluster as “overnight forest trips”. Doing so similarly with the other clusters, we found that the recorded forest-goers’ outdoor trips are best differentiated along 3 dimensions (bolded in Table 3.4): forest penetration, duration/distance and whether the trip happened overnight. Six clusters of outdoor trips emerged: overnight forest trip, overnight non-forest trip, short forest trip, short non-forest trip, day forest trip, day non-forest trip (Table 3.4).

Unsurprisingly, trip duration and trip distance are positively correlated while population density and forest cover are negatively associated. Most outdoor trips tend to stop on at least one occasion and forest-goers spend on average between 30% and 80% of their trip time at a stop location. About two thirds (66%) of the outdoor trips collected were classified as forest trips and just over 10% of outdoor trips happened overnight. Overnight trips are also the longest both in duration and distance covered.



**Table 3.4** - Distribution of input mobility patterns parameters for each of the six identified clusters. Along with Figure 3.6, these numbers are suggestive of what the best labels would be for the clusters.

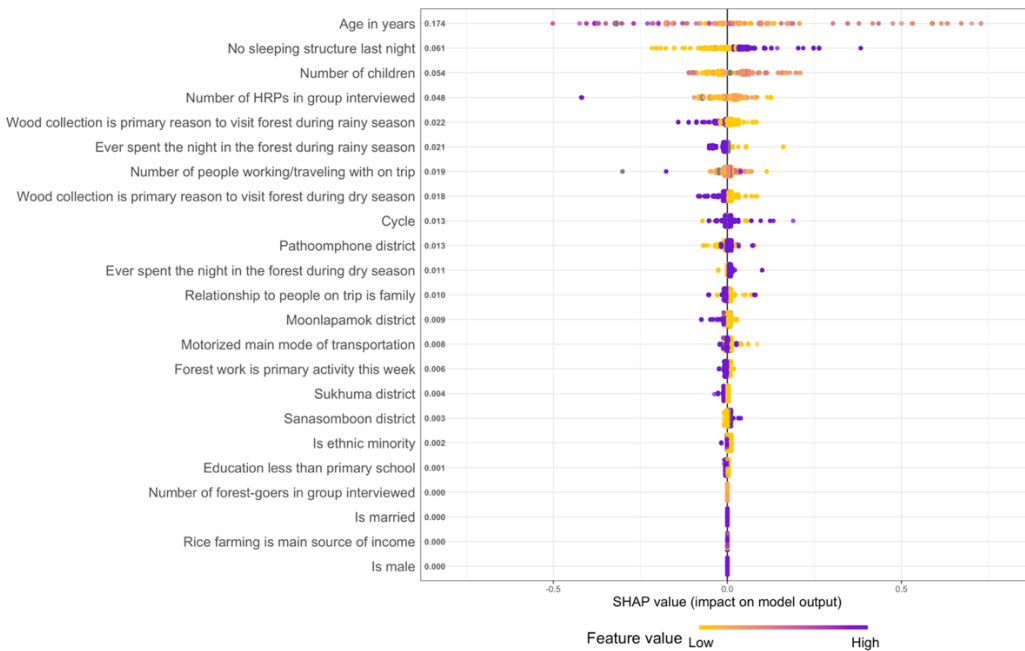
Cluster	1	2	3	4	5	6
<i>Proposed label</i>	<i>Day forest trips</i>	<i>Overnight forest trips</i>	<i>Day non-forest trips</i>	<i>Short forest trips</i>	<i>Short non-forest trips</i>	<i>Overnight non-forest trips</i>
Count (%)	275 (34%)	75 (9%)	183 (23%)	197 (25%)	58 (7%)	15 9 (2%)
<b>Percent of overnight trips (%)</b>	<b>0</b>	<b>100</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>100</b>
Percent of twilight/dawn trips (%)	97.1	0.0	71	21.8	46.6	0.0
Mean average tree crown cover along trip [IQR]	62.2 [49.1; 77.5]	71.8 [59.6; 85.9]	20.2 [6.2; 31.3]	63.4 [52.6; 75.2]	13.5 [0; 26.3]	14.6 [0.5; 28.5]
Mean max tree crown cover along trip [IQR]	84.7 [79.3; 93.4]	90 [85.6; 95.1]	38.6 [22.4; 55.6]	81.5 [75.4; 91]	16.4 [0; 27.4]	41.9 [6.4; 78.7]
<b>Mean proportion of trip with tree crown cover above 50% [IQR]</b>	<b>0.7</b> [0.5; 0.9]	<b>0.8</b> [0.8; 1]	<b>0.1</b> [0; 0.1]	<b>0.7</b> [0.6; 1]	<b>0</b> [0; 0]	<b>0</b> [0; 0]
<b>Mean trip duration (h) [IQR]</b>	<b>8.8</b> [6.7; 10.7]	<b>67.4</b> [36.4; 83.5]	<b>8</b> [4.5; 11.2]	<b>4.7</b> [3.3; 5.2]	<b>6.8</b> [3.9; 10.4]	<b>55.6</b> [30.8; 48]
<b>Mean trip distance (km) [IQR]</b>	<b>6.4</b> [3.7; 8.1]	<b>26.6</b> [15.8; 31.7]	<b>4.9</b> [1.9; 5.2]	<b>2.5</b> [1.5; 3.2]	<b>1.5</b> [1; 2]	<b>15.2</b> [3.5; 28]
Mean max speed along trip (kmh) [IQR]	3.6 [2.2; 3.9]	6.2 [3.5; 7.4]	3.3 [1.8; 4.1]	1.8 [1.1; 2.4]	1.2 [0.9; 1.7]	3.9 [1.4; 6.7]
Mean proportion of trip at stop location [IQR]	0.5 [0.2; 0.7]	0.7 [0.7; 0.9]	0.6 [0.2; 0.9]	0.3 [0; 0.6]	0.8 [0.7; 1]	0.8 [0.7; 1]
Mean number of stop along trip [IQR]	2 [2; 3]	3.1 [2; 3]	1.7 [1; 2]	1.3 [0; 2]	1.7 [2; 2]	2.3 [2; 2.5]
Mean average population density along trip [IQR]	0.9 [0.5; 1.3]	0.6 [0.3; 0.8]	1.3 [0.7; 1.7]	1 [0.6; 1.4]	3 [1.1; 4.6]	2.8 [1.6; 4.4]

### 3.4.3. Regression analysis

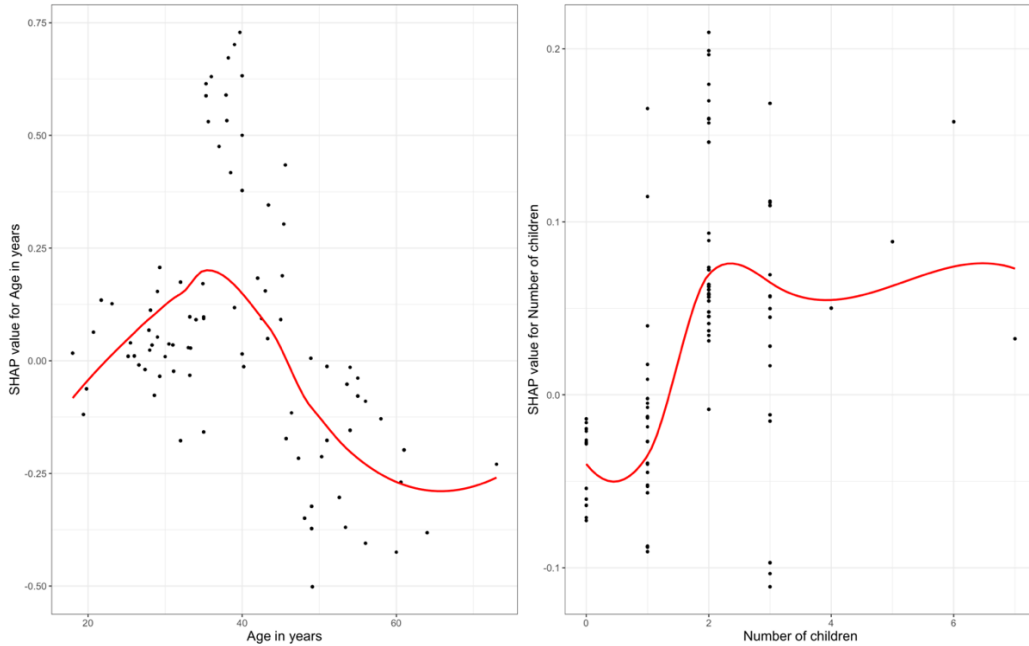
Overnight forest trips as well as forest trips and short forest trips that happened around twilight and/or dawn hours (6 pm and/or 6 am) further defined 385 (48%) high-risk trips because of their presumed higher exposure to malaria vectors. Figure 3.6 presents the results from the regression analysis. Individual-level characteristics of the forest-goers collected in the FTAT survey are ranked in terms of their ability to predict forest-goers' probability to engage in high-risk trips for malaria. Because all the features were collected at the individual level, for each feature, there is one dot per forest-goer, colored by the feature value. The SHAP value represents the change (additive scale) in the forest-goers' probability to engage in high-risk trips. The more positive the SHAP values (right side), the more likely they are to engage in high-risk trips. For instance, forest-goers who reported no sleeping structure the night before their FTAT interview (high feature value, colored in purple) have positive SHAP values. Therefore, they are more likely to engage in high-risk trips. On average, forest-goers' sleeping structure the night before their FTAT interview impacted their probability to engage in high-risk trip by 6.1%. For continuous variables, we can draw the whole SHAP dependence plots (Figure 3.7) for more interpretability. For instance, forest-goers aged between 30 and 45 years have high positive SHAP values. They are therefore more likely to engage in high-risk trips than younger and older forest-goers. On average, forest-goers' age influences their probability to engage in a high-risk trip by 17.4%. For some forest-goers, their middle age increased their probability to engage in high-risk trip by more than 25%.

Together, these results have identified age, lack of outdoor sleeping structure and number of children as the best predictors of high-risk outdoor trips for malaria. Specifically, being 30 to 45

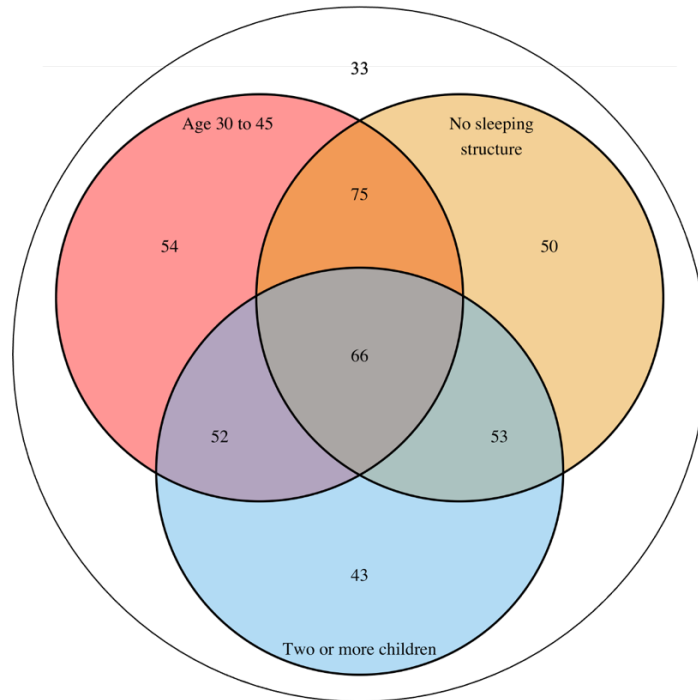
years old, using no structure when sleeping outside and having more than two children all increase the probability for a forest-goers to engage in high-risk trips in terms of their exposure to malaria vectors. All the other features impact forest-goers' probability to engage in high-risk trips by less than 5% on average. As a summary, Figure 3.8 presents the probability of engaging in high-risk trips among forest-goers in the 8 strata defined by to those three main predictors. These predictors, in combination, increase the probability of engaging in high-risk trips up to 75%. We can also see that the reference probability of engaging in high-risk trips among forest-goers not aged between 30 and 45 and who reported sleeping in a structure the night before their FTAT interview and who have less than 2 children is 33%. The average probability to engage in high-risk trips in the seven non-reference strata was 54%, only slightly higher than the unstratified average (48%).



**Figure 3.6 - SHAP importance plot.** Forest-goers' individual features are ranked in terms of their ability to predict the likelihood of forest-goers to engage in high-risk trips. For each feature, there is one dot per forest-goers, colored by the feature value and positioned according to its SHAP value. Larger SHAP values means larger impact on the model predictions. Positive SHAP values result in an increased probability to engage in high-risk trips. The ranking of features is based on the average absolute SHAP value across all forest-goers.



**Figure 3.7** - SHAP dependence plot for the two main continuous predictors of high-risk trips. For each feature, there is one dot per forest-goer. Larger SHAP values means larger impact on the model predictions. Positive SHAP values result in an increased probability to engage in high-risk trips. Super-imposed red lines were modeled using loess with 0.9 span.



**Figure 3.8** - Venn diagram for the raw probability of engaging in high-risk forest trips among the 8 strata of forest-goers defined by the three main predictors identified in the regression analysis: age between 30 and 45, no sleeping structure the night before the FTAT interview and more than two children. Probabilities are expressed in rounded percent. For instance, the baseline probability of engaging in high-risk trips among forest-goers not aged between 30 and 45 and who reported sleeping in a structure the night before their FTAT interview and who have less than 2 children is 33%.

### 3.5. Discussion

Using GPS loggers to capture fine-scale mobility patterns of 122 forest-goers in southern Lao PDR over two-month periods, we extracted data on 803 trips to the forest, forest-fringes or rice-fields. A hierarchical clustering algorithm was used to describe the heterogeneity within these mobility patterns and highlight six major types of outdoors trips. Then, in a regression analysis using gradient boosting trees, forest-goers' age, lack of outside sleeping structures and number of children were identified as the best predictors of their likelihood to engage in trips at higher risk for malaria, in terms of an increased exposure to mosquito vectors. Together, they defined strata of forest-goers with probability as high as 75% and as low as 33% to engage in such high-risk trips.

A key finding from this study is the diversity in forest-goers' mobility patterns highlighted in the cluster analysis. The 803 outdoor trips collected are highly heterogeneous. Some trips lasted no more than 3h when other lasted up to a week. Distance covered ranged from 1 to 100km. Most trips were day trips only but about 10% were overnight. The average tree crown cover along the trip could be above 75% or barely around 5%, even for long trips. Six clusters of outdoor trips were identified with major differences in terms of forest penetration, distance covered, duration and whether the trip happened overnight. These differences likely translate into different exposures to the dominant malaria vectors in the GMS, *An. dirus* and *An. minimus*<sup>13,14</sup>, who thrive in a forested environment and bite during the night and around twilight and dawn hours. This heterogeneity in forest-goers' outdoor trips and exposure to the surrounding mosquito vectors echoes the result from a recent systematic review of the qualitative literature on forest-

goers in the GMS<sup>28</sup>, which calls for a better characterization of the activities that put forest-goers at increased risk for malaria.

We attempted to leverage this heterogeneity in mobility patterns to segment the population of forest-goers and identify sub-groups at higher risk for malaria because of their increased likelihood to engage in high-risk trips. On the one hand, we were able to rank individual level characteristics of forest-goers collected in the FTAT survey in terms of their ability to predict their probability to engage in high-risk trips. The top three individual predictors, number of children, lack of outside sleeping structure and age, would impact, on average, forest-goers' probability to engage in high-risk trips by, respectively, 5%, 7% and 17% on the additive scale and together defined strata of forest-goers with probability as high as 75%. In combination though, these predictors separated the forest-going population in two subgroups with similar probabilities of engaging in such high-risk trips (54% vs 33%). This small difference in risk may be valuable for further targeting resources on high-risk forest-goers but also suggests that some level of risk is ubiquitous among forest-goers. In particular, we failed to identify a very low-risk subgroup and further segmenting this population would imply missing some high-risk forest-goers.

This study also demonstrated how GPS loggers can be used to measure fine-scale mobility patterns of rural and hard to access forest-going populations in the GMS. Thanks to hired PNs, we were able to recruit forest-goers in our study and train them on all aspects of the GPS loggers. Acceptability among forest-goers was high and our study proved its feasibility with very few data gaps thanks to the external charging device and additional batteries that were provided with

the GPS loggers. GPS coordinates every 15 to 30 minutes along forest-going trips represent an incredibly rich dataset about forest-goers' mobility patterns and interaction with their surrounding environment that could not be collected otherwise via surveys or mobile phone data.

On the other hand, data visualization highlighted that forest-goers did not carry the GPS loggers at all times, probably because our instructions insisted too much on the importance of carrying them during forest-going trips. As a result, we restricted our analysis to the 5% of GPS points that were collected along the 803 outdoor trips. This was a necessary step to ensure high-quality input data in our analyses but limits the cost-effectiveness of using such GPS loggers. In addition, these data required substantial processing time and simple steps such as directly collecting the GPS coordinates of forest-goers' house and the exact timing when the GPS logger was handed out would have significantly improved our experience.

Our study has additional limitations. First, our definition of high-risk trips is subjective and based on a simplified version of the malaria ecosystem in the GMS where what matters the most is exposure to mosquito vectors. It is not based on malaria diagnoses. Forest-goers recruited in FTAT were tested for malaria before being given GPS loggers, but reverse causality would have undermined the results from any association analysis and statistical power was low with only six forest-goers in the GPS component of our study testing positive for malaria cases by PCR (polymerase chain reaction). Second, the small sample size of 96 forest-goers in the regression analysis may have lacked enough variation in some individual level features to evaluate their association with high-risk trips. For instance, 95% of forest-goers who carried a GPS logger were male. Third and related, the forest-goers participating in the GPS logger component of the study

happened to be a bit different from those that did not. This may be due to chance or bias in PNs' recruitment of forest-goers. As a consequence, our results may not generalize well to the whole 2,904 forest goers recruited in FTAT or even the 20,000 forest-goers or so estimated to reside in the study area<sup>132</sup>.

In conclusion, this study illustrated how GPS loggers can be leveraged to measure and characterize fine-scale mobility patterns of forest-going populations in southern Lao PDR. The results highlighted the diversity within forest-going trips but did not translate into a clear segmentation of forest-goers' role in malaria transmission in the GMS. These results are key for national control programs across the region to assess and meet their 2030 malaria elimination goals<sup>7,12</sup>.



## **3.6. Appendix 3**

### **3.6.1. S3.1: GPS filtering algorithm**

The advertised precision of the I-gotU GPS loggers used in this study is 10m. Yet, the constructor warns of possible large errors in the GPS coordinates collected, notably when the logger stayed indoor for long periods of time and cannot connect with the satellites. To remove those erroneous GPS points, we used a filtering algorithm that identifies GPS points unusually far away from both the previous and next GPS points, themselves being close by.

Our filtering algorithm first identifies suspect GPS points when the average speed leading to such point from the previously logged one is above 3 km/h (and time difference > 1 min). These suspect points essentially look like suddenly “motorized” departures. Second, the filtering algorithm groups with the identified suspect point, all subsequent points that were recorded really quickly (time difference < 1 min) afterwards. This is because our standard operating procedures documents failed to stress enough that the data logging frequency should not depend on the detected speed of the device. See methods. Third, time difference, distance and elevation between the GPS points starting and ending the sequence of suspect points were computed. Last, the sequence of suspect GPS points (often times comprising only 1 data point) were filtered out according to the following decision rules:

- (Distance between starting and ending suspect GPS points < 100 m) AND (Time between starting and ending suspect GPS points < 30 min)

OR

- (Distance between starting and ending suspect GPS points < 100 m) AND (Time between starting and ending suspect GPS points < 60 min) AND ((Elevation difference with previous GPS point > 500 ft) OR (Average speed between previous and current GPS point > 10 km/h))

OR

- (Distance between starting and ending suspect GPS points < 100 m) AND (Time between starting and ending suspect GPS points < 90 min) AND ((Elevation difference with previous GPS point > 1000 ft) OR (Average speed between previous and current GPS point > 25 km/h))

OR

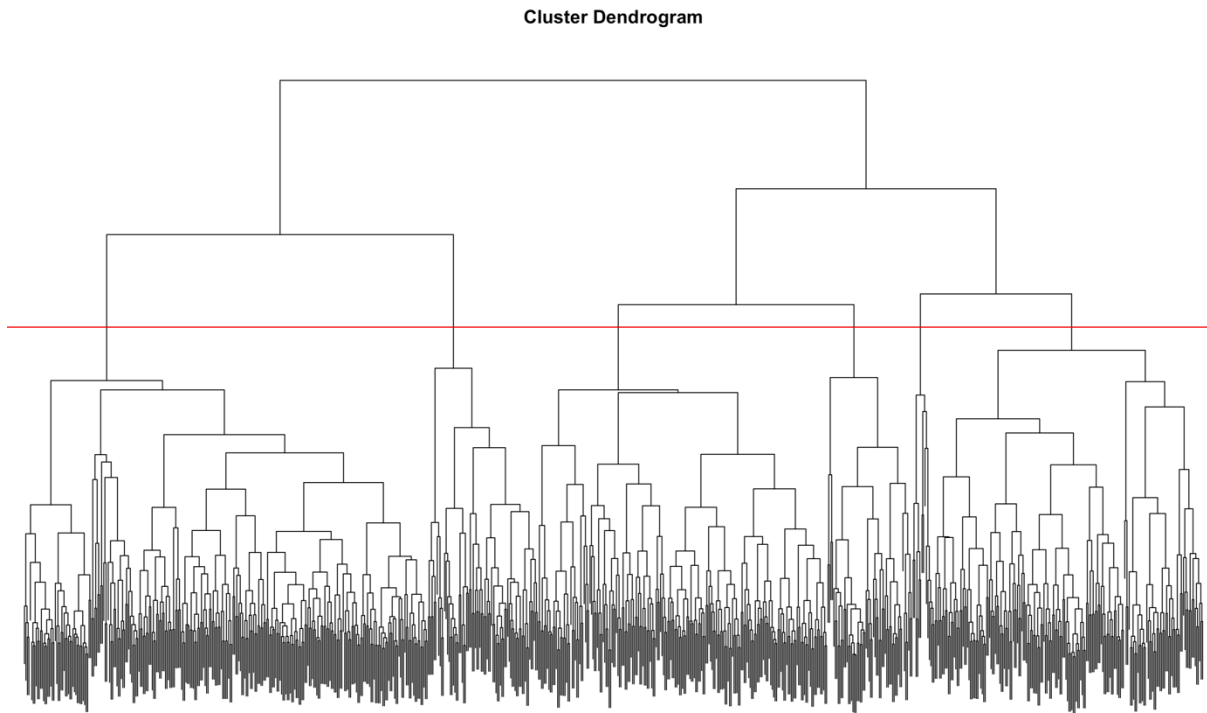
- (Distance between starting and ending suspect GPS points < 3000 m) AND (Time between starting and ending suspect GPS points < 180 min) AND ((Elevation difference with previous GPS point > 3000 ft) OR (Average speed between previous and current GPS point > 40 km/h))

The decision rules were designed to filter out unusual sequence of GPS points between otherwise very close GPS points (before and after). Bumps in elevation difference were also indicative of a temporary dysfunction in the GPS logger and leveraged as such. Unusual speed, even considering the possibility of motor transportation were also used to identify erroneous GPS points. Importantly, these decision rules were refined in an iterative process by visually inspecting the effect of the filtering algorithm on the GPS trajectories.

In addition, our standard operating procedures documents about the configurations of the GPS loggers failed to stress enough that we did not want the data logging frequency to depend on the

detected speed of the device. As a result, for some GPS loggers correctly logging most of their GPS coordinates every 15 to 30 min, the device, when detecting high speed, switched to "motorized" mode and collected GPS coordinates every second. When that happened, we trimmed the data to keep a GPS points every 3 min in order to reduce the computational time in data processing methods described below.

### 3.6.2. S3.2: Additional figures



**Figure 3.9** - Dendrogram from the hierarchical clustering algorithm. Starting from the bottom, every data point, i.e. outdoor trip, is regrouped one a time into “leaves” (=cluster) until they are all in one big and uninformative cluster. The length of the “branches” quantifies the dissimilarity between the leaves. The red horizontal line represents our subjective decision to cut the tree in 6 clusters. We felt selecting 5 clusters would have failed to cut lengthy branches whereas selecting 7 clusters would have started to cut short branches.

## References

1. Organization, W. H. & others. World malaria report 2020: 20 years of global progress and challenges. (2020).
2. Okayas, H. *Mekong Malaria Elimination Programme*. (World Health Organization, Global Malaria Programme, 2018).
3. Organization, W. H. & others. Global report on antimalarial drug efficacy and drug resistance: 2000-2010. (2010).
4. Organization, W. H. & others. Emergency response to artemisinin resistance in the Greater Mekong subregion. Regional framework for action 2013--2015 Geneva: WHO; 2013. (2014).
5. Dondorp, A. M. *et al.* Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med* **361**, (2009).
6. Noedl, H. *et al.* Evidence of artemisinin-resistant malaria in Western Cambodia. *N Engl J Med* **359**, (2008).
7. Organization, W. H. & others. *Eliminating malaria in the Greater Mekong Subregion: united to end a deadly disease*. (2016).
8. Ashley, E. A. *et al.* Spread of Artemisinin Resistance in *Plasmodium falciparum* Malaria. *N. Engl. J. Med.* **371**, 411–423 (2014).
9. Thu, A. M., Phyo, A. P., Landier, J., Parker, D. M. & Nosten, F. H. Combating multidrug-resistant *Plasmodium falciparum* malaria. *FEBS J.* **284**, 2569–2578 (2017).
10. Smith Gueye, C. *et al.* The challenge of artemisinin resistance can only be met by eliminating *Plasmodium falciparum* malaria across the Greater Mekong subregion. *Malar. J.* **13**, 286 (2014).

11. Whittaker, M. A., Dean, A. J. & Chancellor, A. Advocating for malaria elimination - learning from the successes of other infectious disease elimination programmes. *Malar. J.* **13**, 221 (2014).
12. *World Malaria Report 2018*. Geneva: World Health Organization. <https://www.who.int/malaria/publications/world-malaria-report-2018/en/> (2018).
13. Obsomer, V., Defourny, P. & Coosemans, M. The Anopheles dirus complex: spatial distribution and environmental drivers. *Malar. J.* **6**, 26 (2007).
14. Obsomer, V., Dufrene, M., Defourny, P. & Coosemans, M. Anopheles species associations in Southeast Asia: indicator species and environmental influences. *Parasit Vectors* **6**, (2013).
15. Sharma VP & AV, K. Forest malaria in Southeast Asia. in *Proceedings of an Informal Consultative Meeting WHO/MRC, WHO/Malaria Research Centre, New Delhi, 1991* (1991).
16. Durnez, L. *et al.* Outdoor malaria transmission in forested villages of Cambodia. *Malar. J.* **12**, 329 (2013).
17. Chaveepojnkamjorn, W. & Pichainarong, N. Malaria infection among the migrant population along the Thai-Myanmar border area. *Southeast Asian J Trop Med Public Heal.* **35**, 48–52 (2004).
18. Das, N. G., Talukdar, P. K. & Das, S. C. Epidemiological and entomological aspects of malaria in forest-fringed villages of Sonitpur district, Assam. *J. Vector Borne Dis.* **41**, 5 (2004).
19. Lansang, M. A., Belizario, V. Y., Bustos, M. D., Saul, A. & Aguirre, A. Risk factors for infection with malaria in a low endemic community in Bataan, the Philippines. *Acta Trop*

- 63, (1997).
20. Erhart, A. *et al.* Epidemiology of forest malaria in central Vietnam: a large scale cross-sectional survey. *Malar J* **4**, (2005).
  21. Trung, H. D. *et al.* Malaria transmission and major malaria vectors in different geographical areas of Southeast Asia. *Trop Med Int Heal.* **9**, (2004).
  22. Erhart, A. *et al.* Forest malaria in Vietnam: a challenge for control. *Am. J. Trop. Med. Hyg.* **70**, 110–118 (2004).
  23. Guyant, P. *et al.* Malaria and the mobile and migrant population in Cambodia: a population movement framework to inform strategies for malaria control and elimination. *Malar J.* **14**, (2015).
  24. for South-East Asia, W. H. O. *Approaches for mobile and migrant populations in the context of malaria multi-drug resistance and malaria elimination in the Greater Mekong Subregion.* (WHO Regional Office for South-East Asia).
  25. Dysoley, L. *et al.* *Changing patterns of forest malaria among the mobile adult male population in Chumkiri District, Cambodia. Acta tropica* vol. 106 (2008).
  26. Parker, D. M. *et al.* A multi-level spatial analysis of clinical malaria and subclinical Plasmodium infections in Pailin Province, Cambodia. *Heliyon.* **3**, (2017).
  27. Grietens, K. P. *et al.* Low perception of malaria risk among the Ra-glai ethnic minority in south-central Vietnam: implications for forest malaria control. *Malar J.* **9**, (2010).
  28. Nofal, S. D. *et al.* How can interventions that target forest-goers be tailored to accelerate malaria elimination in the Greater Mekong Subregion? A systematic review of the qualitative literature. *Malar J* **18**, 32 (2019).
  29. Lyttleton, C. Deviance and resistance: malaria elimination in the greater Mekong

- subregion. *Soc Sci Med* **150**, (2016).
30. Guerra, C. A., Snow, R. W. & Hay, S. I. A global assessment of closed forests, deforestation and malaria risk. *Ann Trop Med Parasitol* **100**, 189–204 (2006).
  31. Wayant, N. M., Maldonado, D., de Arias, A., Cousiño, B. & Goodin, D. G. Correlation between normalized difference vegetation index and malaria in a subtropical rain forest undergoing rapid anthropogenic alteration. *Geospat Heal.* **4**, 179–190 (2010).
  32. Olson, S. H., Gangnon, R., Silveira, G. A. & Patz, J. A. Deforestation and malaria in Mâncio Lima County, Brazil. *Emerg Infect Dis* **16**, 1108–1115 (2010).
  33. Hahn, M. B. *et al.* Conservation efforts and malaria in the Brazilian Amazon. *Am J Trop Med Hyg* **90**, 591–594 (2014).
  34. Valle, D. & Clark, J. Conservation efforts may increase malaria burden in the Brazilian Amazon. *PLoS One* **8**, e57519 (2013).
  35. Terrazas, W. C. M. *et al.* Deforestation, drainage network, indigenous status, and geographical differences of malaria in the State of Amazonas. *Malar J* **14**, 379 (2015).
  36. Tucker Lima, J. M., Vittor, A., Rifai, S. & Valle, D. Does deforestation promote or inhibit malaria transmission in the Amazon? A systematic literature review and critical appraisal of current evidence. *Philos Trans R Soc L. B Biol Sci* **372**, (2017).
  37. Sawyer, D. R. Frontier malaria in the Amazon region of Brazil: types of malaria situations and some implications for control. *Brasília: PHO/WHO/TDR* (1988).
  38. de Castro, M. C., Monte-Mór, R. L., Sawyer, D. O. & Singer, B. H. Malaria risk on the Amazon frontier. *Proc Natl Acad Sci U S A* **103**, 2452–2457 (2006).
  39. Ilacqua, R. C. *et al.* A method for estimating the deforestation timeline in rural settlements in a scenario of malaria transmission in frontier expansion in the Amazon Region. *Mem*



- Inst Oswaldo Cruz* **113**, e170522 (2018).
40. Singer, B. & de Castro, M. C. Enhancement and suppression of malaria in the Amazon. *Am J Trop Med Hyg* **74**, 1–2 (2006).
  41. Vittor, A. Y. *et al.* Linking deforestation to malaria in the Amazon: characterization of the breeding habitat of the principal malaria vector, *Anopheles darlingi*. *Am J Trop Med Hyg* **81**, 5–12 (2009).
  42. Vittor, A. Y. *et al.* The effect of deforestation on the human-biting rate of *Anopheles darlingi*, the primary vector of *Falciparum* malaria in the Peruvian Amazon. *Am J Trop Med Hyg* **74**, 3–11 (2006).
  43. Moreno, J. E., Rubio-Palis, Y., Páez, E., Pérez, E. & Sánchez, V. Abundance, biting behaviour and parous rate of anopheline mosquito species in relation to malaria incidence in gold-mining areas of southern Venezuela. *Med Vet Entomol* **21**, 339–349 (2007).
  44. Baeza, A., Santos-Vega, M., Dobson, A. P. & Pascual, M. The rise and fall of malaria under land-use change in frontier regions. *Nat. Ecol. & Evol.* **1**, 108 (2017).
  45. Bauhoff, S. & Busch, J. Does deforestation increase malaria prevalence? Evidence from satellite data and health surveys. *World Dev.* **127**, 104734 (2020).
  46. Beck-Johnson, L. M. *et al.* The effect of temperature on *Anopheles* mosquito population dynamics and the potential for malaria transmission. *PLoS One* **8**, e79276 (2013).
  47. Mordecai, E. A. *et al.* Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecol Lett* **16**, 22–30 (2013).
  48. Parham, P. E. & Michael, E. Modeling the effects of weather and climate change on malaria transmission. *Env. Heal. Perspect* **118**, 620–626 (2010).
  49. Parham, P. E. *et al.* Modeling the role of environmental variables on the population

- dynamics of the malaria vector *Anopheles gambiae sensu stricto*. *Malar J* **11**, 271 (2012).
50. Hay, S. I., Snow, R. W. & Rogers, D. J. Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Trans R Soc Trop Med Hyg* **92**, 12–20 (1998).
  51. Teklehaimanot, H. D., Lipsitch, M., Teklehaimanot, A. & Schwartz, J. Weather-based prediction of *Plasmodium falciparum* malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. *Malar J* **3**, 41 (2004).
  52. Busch, J. & Ferretti-Gallon, K. What Drives Deforestation and What Stops It? A Meta-Analysis. *Rev. Environ. Econ. Policy* **11**, 3–23 (2017).
  53. Austin, K. F., Bellinger, M. O. & Rana, P. Anthropogenic forest loss and malaria prevalence: a comparative examination of the causes and disease consequences of deforestation in developing nations. *AIMS Environ. Sci.* **4**, 217–231 (2017).
  54. Pattanayak, S. K., Corey, C. G., Lau, Y. F. & Kramer, R. A. Biodiversity conservation and child malaria: microeconomic evidence from Flores, Indonesia. *Econ. Res. Initiat. Duke Work. Pap.* (2010).
  55. Garg, T. Public health effects of ecosystem degradation: Evidence from deforestation in Indonesia. *Tech. repor t. <https://www.sites.google.com/site/teevrat/research>, last accessed March 8, 2015* (2015).
  56. Fornace, K. M. *et al.* Association between Landscape Factors and Spatial Patterns of *Plasmodium knowlesi* Infections in Sabah, Malaysia. *Emerg Infect Dis* **22**, 201–208 (2016).
  57. Hahn, M. B., Gangnon, R. E., Barcellos, C., Asner, G. P. & Patz, J. A. Influence of deforestation, logging, and fire on malaria in the Brazilian Amazon. *PLoS One* **9**, e85725

- (2014).
58. Valle, D. Response to the critique by Hahn and others entitled ‘Conservation and malaria in the Brazilian Amazon’. *Am J Trop Med Hyg* **90**, 595–596 (2014).
  59. MacDonald, A. J. & Mordecai, E. A. Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. *Proc. Natl. Acad. Sci.* **116**, 22212–22218 (2019).
  60. Hansen, M. C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science (80-. )*. **342**, 850–853 (2013).
  61. *Lao PDR National Strategic Plan for Malaria Control and Elimination 2016-2020*. (2016).
  62. Bannister-Tyrrell, M. *et al.* Forest Goers and Multidrug-Resistant Malaria in Cambodia: An Ethnographic Study. *Am J Trop Med Hyg* **100**, 1170–1178 (2019).
  63. Wen, S. *et al.* Targeting populations at higher risk for malaria: a survey of national malaria elimination programmes in the Asia Pacific. *Malar J* **15**, 271 (2016).
  64. Smith, C. & Whittaker, M. Beyond mobile populations: a critical review of the literature on malaria and population mobility and suggestions for future directions. *Malar J* **13**, 307 (2014).
  65. Bennett, A., Lover, A. A. & Dantzer, E. A. *Unpublished report on formative assessment to identify and characterize mobile and migrant populations in Champasak Province, southern Lao PDR*.
  66. Lover, A. A. *et al.* Study protocol for a cluster-randomized split-plot design trial to assess the effectiveness of targeted active malaria case detection among high-risk populations in Southern Lao PDR (the AcME-Lao study) [version 1; peer review: awaiting peer review].

- Gates Open Res.* **3**, (2019).
67. Cotter, C. *et al.* The changing epidemiology of malaria elimination: new strategies for new challenges. *Lancet* **382**, 900–911 (2013).
  68. Kaehler, N. *et al.* Prospects and strategies for malaria elimination in the Greater Mekong Sub-region: a qualitative study. *Malar J* **18**, 203 (2019).
  69. Tropek, R. *et al.* Comment on ‘High-resolution global maps of 21st-century forest cover change’. *Science (80-. )*. **344**, 981 (2014).
  70. Hansen, M. *et al.* Response to comment on ‘High-resolution global maps of 21st-century forest cover change’. *Science (80-. )*. **344**, 981 (2014).
  71. Lao National Malaria Database (DHIS2). Vientiane: CMPE/Ministry of Health; 2018.
  72. Lover, A. A. *et al.* Prevalence and risk factors for asymptomatic malaria and genotyping of glucose 6-phosphate (G6PD) deficiencies in a vivax-predominant setting, Lao PDR: implications for sub-national elimination goals. *Malar J* **17**, 218 (2018).
  73. UNFPA. *Lao People’s Democratic Republic: Results of Population and Housing Census 2015 (English Version)*. Accessed 15 July 2019. <https://lao.unfpa.org/publications/results-population-and-housing-census-2015-english-version> (2016).
  74. Lao national census (Lao DECIDE).
  75. Jarvis, A., Reuter, H. I., Nelson, A., Guevara, E. & others. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database. (2008).
  76. Wan, Z., Hook, S. & Hulley, G. MOD11C3 MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006 [Data set]. NASA EOSDIS Land Processes DAAC. (2015).
  77. Funk, C. C. *et al.* A quasi-global precipitation time series for drought monitoring. *US*

- Geol. Surv. Data Ser.* **832**, 1–12 (2014).
78. Wood, S. N. *Generalized additive models: an introduction with R.* (Chapman and Hall/CRC, 2017).
  79. Alegana, V. A. *et al.* Spatial modelling of healthcare utilisation for treatment of fever in Namibia. *Int J Heal. Geogr* **11**, 6 (2012).
  80. Sturrock, H. J. W. *et al.* Fine-scale malaria risk mapping from routine aggregated case data. *Malar J* **13**, 421 (2014).
  81. European Space Agency: GlobCover Land Cover v2 2009 database. (2010).
  82. Open Street Map.
  83. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2008).
  84. Csardi, G., Nepusz, T. & others. The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**, 1–9 (2006).
  85. van Etten, J. R Package gdistance: Distances and Routes on Geographical Grids. *J. Stat. Software, Artic.* **76**, (2017).
  86. Sanann, N. *et al.* Forest work and its implications for malaria elimination: a qualitative study. *Malar. J.* **18**, 376 (2019).
  87. Cui, L. *et al.* Malaria in the Greater Mekong Subregion: heterogeneity and complexity. *Acta Trop.* **121**, 227–239 (2012).
  88. Sluydts, V. *et al.* Spatial clustering and risk factors of malaria infections in Ratanakiri Province, Cambodia. *Malar. J.* **13**, 387 (2014).
  89. Incardona, S. *et al.* Large-scale malaria survey in Cambodia: novel insights on species distribution and risk factors. *Malar J* **6**, 37 (2007).

90. Global, H. I. V. Biobehavioural Survey Guidelines. (2017).
91. Abdul-Quader, A. S., Gouws-Williams, E., Tlou, S., Wright-De Agüero, L. & Needle, R. Key populations in sub-Saharan Africa: population size estimates and high risk behaviors. *AIDS Behav* **19 Suppl 1**, S1-2 (2015).
92. Johnston, L., Saumtally, A., Corceal, S., Mahadoo, I. & Oodally, F. High HIV and hepatitis C prevalence amongst injecting drug users in Mauritius: findings from a population size estimation and respondent driven sampling survey. *Int J Drug Policy* **22**, 252–258 (2011).
93. Shokoohi, M., Baneshi, M. R. & Haghdoost, A.-A. Size Estimation of Groups at High Risk of HIV/AIDS using Network Scale Up in Kerman, Iran. *Int J Prev Med* **3**, 471–476 (2012).
94. Li, L., Assanangkornchai, S., Duo, L., McNeil, E. & Li, J. Risk behaviors, prevalence of HIV and hepatitis C virus infection and population size of current injection drug users in a China-Myanmar border city: results from a Respondent-Driven Sampling Survey in 2012. *PLoS One* **9**, e106899 (2014).
95. Handcock, M. S., Gile, K. J. & Mar, C. M. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics* **71**, 258–266 (2015).
96. Arumugam, E. *et al.* Size Estimation of high-risk groups for hiv infection in india based on data from national integrated bio-behavioral surveillance and targeted interventions. *Indian J Public Heal.* **64**, S39–S45 (2020).
97. Rehle, T., Lazzari, S., Dallabetta, G. & Asamoah-Odei, E. Second-generation HIV surveillance: better data for decision-making. *Bull World Heal. Organ* **82**, 121–127 (2004).

98. Organization, W. H. & others. Guidelines for second generation HIV surveillance: An update: Know your epidemic. (2013).
99. Costenbader, J., Broadhead, J., Yasmi, Y. & Durst, P. B. Drivers affecting forest change in the greater mekong subregion (GMS): an overview. *FAO Rome, Italy* (2015).
100. Jacobson, J. O. *et al.* Surveillance and response for high-risk populations: what can malaria elimination programmes learn from the experience of HIV? *Malar. J.* **16**, 33 (2017).
101. Lumley, T. & others. Analysis of complex survey samples. *J Stat Softw* **9**, 1–19 (2004).
102. Dorfman, R. A. A note on the  $\chi^2$ -method for finding variance formulae. *Biometric Bull.* (1938).
103. Cormack, R. M. & others. Loglinear models for capture-recapture experiments on open populations. (1980).
104. Cormack, R. M. Log-linear models for capture-recapture. *Biometrics* 395–413 (1989).
105. Chao, A. An overview of closed capture-recapture models. *J. Agric. Biol. Environ. Stat.* **6**, 158–175 (2001).
106. for Disease Monitoring, I. W. G. & Forecasting. Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. *Am. J. Epidemiol.* **142**, 1047–1058 (1995).
107. for Disease Monitoring, I. W. G. & Forecasting. Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases. *Am. J. Epidemiol.* **142**, 1059–1068 (1995).
108. Baillargeon, S., Rivest, L.-P. & others. Rcapture: loglinear models for capture-recapture in R. *J. Stat. Softw.* **19**, 1–31 (2007).

109. Venzon, D. J. & Moolgavkar, S. H. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *J. R. Stat. Soc. Ser. C (Applied Stat.* **37**, 87–94 (1988).
110. McCreesh, P. *et al.* Subpatent malaria in a low transmission African setting: a cross-sectional study using rapid diagnostic testing (RDT) and loop-mediated isothermal amplification (LAMP) from Zambezi region, Namibia. *Malar J* **17**, 480 (2018).
111. Schwarz, C. J. & Seber, G. A. F. Estimating animal abundance: review III. *Stat. Sci.* 427–456 (1999).
112. Rivest, L.-P. & Lévesque, T. Improved Log-Linear Model Estimators of Abundance in Capture-Recapture Experiments. *Can. J. Stat. / La Rev. Can. Stat.* **29**, 555–572 (2001).
113. Pollock, K. H. A Capture-Recapture Design Robust to Unequal Probability of Capture. *J. Wildl. Manage.* **46**, 752–757 (1982).
114. Rerolle, F. *et al.* Spatio-temporal associations between deforestation and malaria incidence in Lao PDR. *Elife* **10**, e56974 (2021).
115. Vazquez-Prokopec, G. M. *et al.* Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. *Int. J. Health Geogr.* **8**, 1–11 (2009).
116. Duncan, S. *et al.* Portable global positioning system receivers: static validity and environmental conditions. *Am. J. Prev. Med.* **44**, e19--e29 (2013).
117. Paz-Soldan, V. A. *et al.* Assessing and maximizing the acceptability of global positioning system device use for studying the role of human movement in dengue virus transmission in Iquitos, Peru. *Am. J. Trop. Med. Hyg.* **82**, 723–730 (2010).
118. Vazquez-Prokopec, G. M. *et al.* Using GPS Technology to Quantify Human Mobility, Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban



- Environment. *PLoS One* **8**, 1–10 (2013).
119. Stothard, J. R., Sousa-Figueiredo, J. C., Betson, M., Seto, E. Y. W. & Kabatereine, N. B. Investigating the spatial micro-epidemiology of diseases within a point-prevalence sample: a field applicable method for rapid mapping of households using low-cost GPS-dataloggers. *Trans. R. Soc. Trop. Med. Hyg.* **105**, 500–506 (2011).
  120. Seto, E. Y. W., Knapp, F., Zhong, B. & Yang, C. The use of a vest equipped with a global positioning system to assess water-contact patterns associated with schistosomiasis. *Geospat. Health* 233–241 (2007).
  121. Brant, T. A. *et al.* Integrated risk mapping and landscape characterisation of lymphatic filariasis and loiasis in South West Nigeria. *Parasite Epidemiol. Control* **3**, 21–35 (2018).
  122. Searle, K. M. *et al.* Characterizing and quantifying human movement patterns using GPS data loggers in an area approaching malaria elimination in rural southern Zambia. *R. Soc. open Sci.* **4**, 170046 (2017).
  123. Fornace, K. M. *et al.* Local human movement patterns and land use impact exposure to zoonotic malaria in Malaysian Borneo. *Elife* **8**, (2019).
  124. Hast, M. *et al.* The use of GPS data loggers to describe the impact of spatio-temporal movement patterns on malaria control in a high-transmission area of northern Zambia. *Int. J. Health Geogr.* **18**, 19 (2019).
  125. Barraquand, F. & Benhamou, S. ANIMAL MOVEMENTS IN HETEROGENEOUS LANDSCAPES: IDENTIFYING PROFITABLE PLACES AND HOMOGENEOUS MOVEMENT BOUTS. *Ecology* **89**, 3336–3348 (2008).
  126. Calenge, C. The package ``adehabitat'' for the R software: a tool for the analysis of space and habitat use by animals. *Ecol. Modell.* **197**, 516–519 (2006).

127. Team, R. C. R: A Language and Environment for Statistical Computing. (2014).
128. Benhamou, S. & Riote-Lambert, L. Beyond the Utilization Distribution: Identifying home range areas that are intensively exploited or repeatedly visited. *Ecol. Modell.* **227**, 112–116 (2012).
129. Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS One* **8**, 1–11 (2013).
130. Sigrist, F. Gaussian Process Boosting. *arXiv Prepr. arXiv2004.02653* (2020).
131. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv Prepr. arXiv1705.07874* (2017).
132. Rerolle, F. *et al.* Population Size Estimation of Seasonal Forest-going Populations in Southern Lao PDR. (2021). Under review.

## **Ethics**

The studies conducted as part of this dissertation work were approved by the National Ethics Committee for Health Research at the Lao Ministry of Health (Approval #2016-014) and by the UCSF ethical review board (Approvals #16-19649 and #17-22577). The informed consent process was consistent with local norms, and all study areas had consultation meeting with, and approvals from, village elders. All participants provided informed written consent; caregivers provided consent for all children under 18, and all children aged 10 and above also provided consent directly. These studies were conducted according to the ethical principles of the Declaration of Helsinki of October 2002.

## Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

*Francois Louis Kerolle*

B4EA2B28774A450...

Author Signature

5/26/2021

Date