Benjamin Rice
November 29, 2017

An attempt to develop a spatial hierarchical model of predictors of malaria infection in vulnerable communities in Madagascar

Context:
Malaria remains a major, if not the major, economic and health burden for many developing countries, including Madagascar, where recent intervention efforts have largely failed to produce improvements. A better understanding of the factors predicting infection and the spatial structure of variation in infection rate is of interest as a result. Many of the recent studies have focused solely on environmental predictors; however, socio-economic factors that affect the human host's susceptibility (e.g. immune system attenuation due to malnutrition) and ability to respond to infection (e.g. access to medical care) have been long documented. Indeed, the first malaria control campaigns in Europe in the early 1900s often focused on the socio-economic predictors of infection (see the footnote below[1]).

Aim:
Here, I aim to explore some patterns in the malaria infection data I collected during fieldwork this past year. I will analyze a set of malaria infection outcomes paired with 14 variables, mostly socio-economic and demographic (see Supplementary Table 1), in a nested sample design of 5544 individuals in in 1129 households in 24 rural communities in 4 'eco-zones' of Madagascar (see Figure 1 for an overview of the study design, and Table 1 for a summary). The relative importance of different predictors for the outcome (malaria infection) and the changes in the relationships among variables for different spatial and demographic strata of the sample will be explored.

I focus on using conditional probabilities and simple regression to develop a robust argument for the parameters and spatial levels to include in an integrated hierarchical model later. However, I do not attempt to develop that full Bayesian model here. I conclude with potential future analyses to perform based on the results of these preliminary explorations of the data

Table 1: Summary of data set analyzed

Regions	4
Sites	24
Households	1129
Individuals	6293
Mean household size (number of individuals)	5.57
Household size range	(2-19)
Individuals with valid malaria result	5544
Percent of individuals with a malaria result (%)	88.1
Number of malaria positive individuals	776
Malaria prevalence (overall) (%)	13.99
Mean age of individuals (years)	17.34

[1] For e.g. in 1909, Giuseppe Zagari, Italian health commissioner for Sardinia, noted that "I am therefore perfectly convinced that the malaria question is synonymous with the social question and that it is inseparable from the economic development of Sardinia" [Zagari, 1909]

Methods, results, and future directions – Organized by spatial scale

1. Motivation to explore across spatial scales

Figure 1 displays that the study design is highly nested as individuals tested for malaria reside within households that are within sites (villages or clusters of households) that are within regions (geographically and ecologically distinct areas of Madagascar). See Data File 1 for all data (please do not share). First, one can ask if there is evidence that factors acting at higher spatial levels (household, site, or region factors) have an observable effect on the distribution of infections among individuals and thus necessitate a hierarchical model that accounts for these levels.

Evidence supporting the structuring of variation in the probability of infection at multiple spatial scales was observed from the conditional probabilities of infection among households, high variance in infection frequency between very close sites, and from comparing simple regression models between regions.

2. Determining the spatial units to be modeled: Evidence of a nonrandom distribution of infections among ***households***

First, at the household level, one might expect that households have a similar probability of infection given that they are often extremely close to each other physically and the mosquito vectors that transmit malaria are often abundant and well dispersed. If, however, there are household factors that modify the probability of infection, then one would expect the overall probability of infection (*Pr(infection)*) to differ from the probability of infection given information on the household (*Pr(infection|household)*). In fact, it was that observed that the probability of infection for individuals within households where another household member infected was much greater than the unconditional probability of infection for individuals at a site.

Site 204 provides an illustrative example. We can define the probability of observing that an individual is infected at that site, *Pr(infection)*, as the percentage of individuals infected. For site 204 *Pr(infection)* = 0.159 (44 infected out of 276 individuals sampled) (see Supplementary Figure 1). However, *Pr(infection | household)*, if one conditions on infection status of other members of the household, was 0.305. The difference between *Pr(infection)* and *Pr(infection | household)* suggests that infections are not distributed randomly among households and that factors acting at the household level may be relevant predictors of malaria infection.

A possible future way to more rigorously characterize household clustering could be to use simulations where varying relative differences in the probability of infection given household membership are inputted. Distributions of the amount of clustering of infections in households could be derived from simulations that include either (i) no household effect (*Pr(infection)* is fully independent for each individual), (ii) no household effect, but accounts for individual level predictors for e.g. age (*Pr(infection)* for members of a household depends solely on the individual level predictors associated with the individuals in the household), or (iii) a household-specific effect (*Pr(infection)* varies based on household membership) in addition to the individual level predictors. These simulated data could then be compared to the observed levels of household clustering of infections. See Figure 2 for histograms showing the age distribution as compared to the age distribution of infected and uninfected individuals.

3. Determining the spatial units to be modeled: Evidence of a nonrandom distribution of infections among ***sites***

At the site level, extreme variation in the prevalence of infection was observed between sites, and even between sites within the same region which are separated by short geographic distances (<10km) (See Figure 3). This indicates that site specific factors may be relevant predictors of malaria infection in addition to household specific factors.

Site specific variables with data available include road access, access to medical care, and whether a site is coastal/inland.

A regression model, with no interaction terms, was fit using the site-specific variables listed in Supplementary Table 1 as predictors and malaria prevalence at the 24 sites as a continuous outcome. Coefficient estimates and confidence intervals are shown in Table 2. Data and accompanying R code are provided in Data Files 1 and 2.

Not surprisingly, given just 24 sites and no clear pattern, the confidence intervals for the coefficients for nearly all predictors included zero. Simulated data where the empirical coefficients for travel time to a doctor and distance to a hospital were exaggerated and variance minimized, each by a factor of two, did not differ substantially (See Supplementary Table 2). This could indicate that many more sites, or much larger effect sizes, would be needed to be capable of observing relationships with these predictors.

Illustrating the ability of site factors to confound inference when they are not accounted for, the coefficient estimate for whether a site is coastal or inland differed greatly depending on whether Region 5 sites were included or excluded. Since Region 5 contained the lowest prevalence sites and is inland, the coefficient was large and the confidence interval for the estimate did not include zero when Region 5 sites were included. However, from a basic understanding of malaria, the below-mentioned temperature and elevation limits on malaria transmission in Region 5 are likely the predominant drivers in that region rather than coastal/inland effects. A better test, similar to the 'secret' weapon strategy of Gelman, of whether coastal/inland is a relevant predictor or simply a covariate of temperature/elevation in Region 5, is to see if it the association remains observable in other regions. However, when refitting the model after excluding the Region 5 sites, the strength of the coastal/inland association with prevalence decreased substantially.

This indicates that with the data available it seems that other factors, some of which likely aggregate at the regional level, are as likely or more likely to explain the variation between sites observed. It also seems that given the multitude of variables known *a priori* to differ between regions, but that which are not included in the data available, many other complex relationships, such as the one between inland/coastal sites and prevalence, may exist. As a result, it seems that modeling the 4 regions independently may be necessary to have the opportunity to capture observable relationships within regions.

Table 2: Using a simple generalized linear model of four site-specific predictors of malaria prevalence to demonstrate the need for accounting for inter-regional differences

Table 2A: All regions combined (*n = 24* sites)

Predictor	Coefficient	Confidence interval
Travel time by foot to nearest doctor (hours) (continuous)	-0.73	[-2.31, 0.85]
Road access (binary)	-10.64	[-23.15, 1.87]
Coastal versus inland (binary)	15.56	[2.25, 28.87]
Distance to nearest hospital (km) (integer)	0.31	[-0.099, 0.73]

Table 2B: Excluding Region 5 (*n = 18* sites)

Predictor	Coefficient	Confidence interval
Travel time by foot to nearest doctor (hours) (continuous)	-0.93	[-3.0684, 1.2084]
Road access (binary)	-13.73	[-29.2334, 1.7734]
Coastal versus inland (binary)	11.72	[-5.506, 28.946]
Distance to nearest hospital (km) (integer)	0.28	[-0.215, 0.775]

4. Determining the spatial units to be modeled: Evidence of a nonrandom distribution of infections among ***regions***

At the regional level, large differences between the four geographic areas were apparent (see Supplementary Figure 2 for a map) with Region 5 consistently having very low prevalence and Region 4 having consistently high prevalence. It is worth noting that the regions differ dramatically in terms of precipitation, elevation, demographics and other variables. Notably, mean monthly temperature is much lower and housing practices are much more distinct in Region 5 (located in the elevated Central Plateau), both of which likely contribute to the much lower mean prevalence observed in that region.

However, this introduces the complication that relationships among predictors of malaria infection may vary between regions. For example, temperature may place a strict limit on malaria transmission potential in the High Plateau (Region 5) such that an individual's travel (which inherently requires going from the plateau to higher risk, lower elevation areas) becomes a predominant risk factor. On the other hand, travel history for individuals in Region 4, where prevalence is consistently high in the broad geographic area, is unlikely to increase the probability of infection.

Another example can be found with household size (the number of individuals cohabitating in a household), which is known to be a risk factor for infectious disease (due to increased contact from crowding, its correlation with poverty, etc). Due to the cultural custom of the predominant ethnic group in Region 5, the *Betsileo*, large, multi-generational households are common. Indeed, mean household size in Region 5 (6.49 individuals, *n* = 243 households) was noticeable higher than the overall mean (5.58, *n* = 1129) or that observed in other regions (5.34, 5.45, and 5.19 for Regions 2, 3, and 4, respectively).

As expected from this observation, if one regresses mean household size on malaria infection for a site, without considering regional differences, increased household size is negatively associated with malaria infection due the lower malaria prevalence in the larger household high plateau sites (see Table 3). This goes against the *a priori* expectation that increased household size should predict disease risk generally. However, a future, more informative test would be to include household size as a household specific variable for each individual as the wide confidence intervals seen in Table 3 demonstrate that sample size at the site level (*n* = 24 or 18) does not provide much power.

Table 3: Mean household size as a site predictor – not enough sites to have sufficient statistical power

Model: *Pr(infection)* = $\alpha + \beta$*(*mean household size*) + error

Model parameter estimates	All regions included*	Region 5 excluded*
α	38.089 [1.20, 74.97]	7.44 [-76.01, 90.89]
β	-4.301 [-10.749, 2.14]	2.04 [-13.58, 17.66]
Residual sd	15.03	15.65

*95% confidence intervals shown in brackets next to parameter estimates

This provides further evidence that the direction or strength of many relationships may change between regions and that modeling them independently may be the best strategy.

Conclusions:

The primary aim was to begin exploring patterns in the spatial units to be modeled in this nested data set on malaria infection outcomes in Madagascar. Evidence of a nonrandom distribution of infections among households, sites, and regions was demonstrated and this exercise provided insight onto how those higher scale levels can be accounted for when attempting to understand the variation among individuals. While developing the full hierarchical model will be a much more important step, it was somewhat useful in that these analyses produced information that can be used to rigorously defend the decisions used when constructing the hierarchical model. For example, while there were *a priori* expectations that some variables are possible significant predictors, it was not clear *a priori* if they would explain the large-scale differences between sites and regions that were seen or would be consistent within regions. I described a few instances where differing relationships were observed between regions (e.g., in household size, importance of coastal/inland). This indicates that regions should be modeled independently perhaps. It was also observed that some variables that could be applied at the site level (e.g. mean household size) would be better explored at lower spatial scales such as at the level of individual factors. While disappointing that a full Bayesian hierarchical model isn't ready (partially because I struggled to assemble the data set in sufficient time) I do appreciate the opportunity to begin exploring the data using some simple regression analyses that I would not have been able to do prior to taking this class.

Supplementary Table 1: Variables and descriptions

	Data column or Variable name	Notes	Variable type
	Malaria infection	Outcome/response variable: Malaria infection status as determined by a rapid diagnostic test (RDT) on site (1 = infection, 0 = no infection	binary
	region_code	Code specifying which of the 4 regions where the sampling site is locatied (2 = Southest (Mananjary), 3 = Southwest (Toliara), 4 = West Coast (Morombe), 5 = Amoroni Mania (High Plateau)) *note that it is not relevant that the region code starts at 2	NA
	s_code	Code specifying the site	NA
1	**s_doc_walk_hours**	Hours by foot to nearest doctor as reported by village elders	continuous
2	**s_road_access**	1 = site is reachable by passable roads at the time of sampling, 0 = no road access	binary
3	**s_coastal**	1 = Site is within 3km of coast or estuary, 0 = site is three or more km inland	binary
3	**s_ distance_hospital**	Straight line distance to nearest district hospital (in km)	continuous
	hh_code	Code specifying the household, defined as a unit of people that cohabitate and regularly share meals	NA
4	**hh size**	Number of individuals cohabitating in the household	integer
5	**hh_income**	If income has been received through cash crop sales, livestock sales, or wage labor by the household in the last month (1 = received income, 0 = no income)	binary
6	**hh_hoh_school_yrs**	Total number of years of education completed by the head of household	integer
7	**hh_wall_material**	1 = Walls of the household are made out of non-porous materials (e.g. brick or cement); 0 = walls made from traditional porous materials (e.g. bamboo)	binary
8	**hh_hamlet_ownership**	1 = The household has a temporary agricultural shelter and has spent 1 or more months living in that temporary shelter in the past year, 0 = no hamlet ("lasy" in Malagasy)	binary
	ind_id	Code specifying the individual	NA
9	**ind_sex**	1 = male, 0 = female	binary
10	**ind_age**	Approximate age calculated from 2017 minus year of birth (note that individual's access to documentation is limited and many birthdays cannot be estimated at a finer scale than the year of birth)	integer
11	**ind_food_insecure**	Data available for adults (age > 16) only: 1 = The individual reports skipping eating for entire days due to a lack of food, 0 = Does not report skipping eating	binary
12	**ind_travel_mos_away**	1 = Individual spent 1 or more months living in another location in the last year, 0 = permanent resident at the sampling site	binary
13	**ind_travel_recent**	1 = Individual reported spending more than 2 nights away from home community in the last month, 0 = remained at home	binary
14	**ind_anemia**	Hemoglobin reading (g / dL) indicative of anemia status	continuous

Supplementary Table 2: Parameter values: Empirically derived, chosen simulation inputs or simulation outputs for the intercepts, coefficients, and residual standard deviations show

	Empirical estimate	Simulation (1X)* Inputs	Simulation (1X) Outputs (mean)	Simulation (2X)* Inputs	Simulation (2X) Outputs (mean)
Intercept	14.09	14	5.64	14	18.72
Travel time by foot to nearest doctor (hours)	-0.72	-0.70	-0.71	-1.50	-1.49
Distance to nearest hospital (km)	0.14	0.13	0.13	0.26	0.26
Residual sd	15.61	15	14.84	7.30	7.18

*In the “2X” simulation scenario the coefficients (i.e. slopes) were increased approximately by a factor of two and the residual sd was decreased by approximately a fact of two (intercept unchanged). For both simulation scenarios, the proportion of repetitions in which the simulation returned parameter estimates with confidence intervals that covered the inputted parameter values were recorded. For both coefficients in both simulation scenarios, the proportions were above 0.93.

Figure 1: Schematic of nested study design where individuals were sampled within households at 6 sites per region for 4 regions

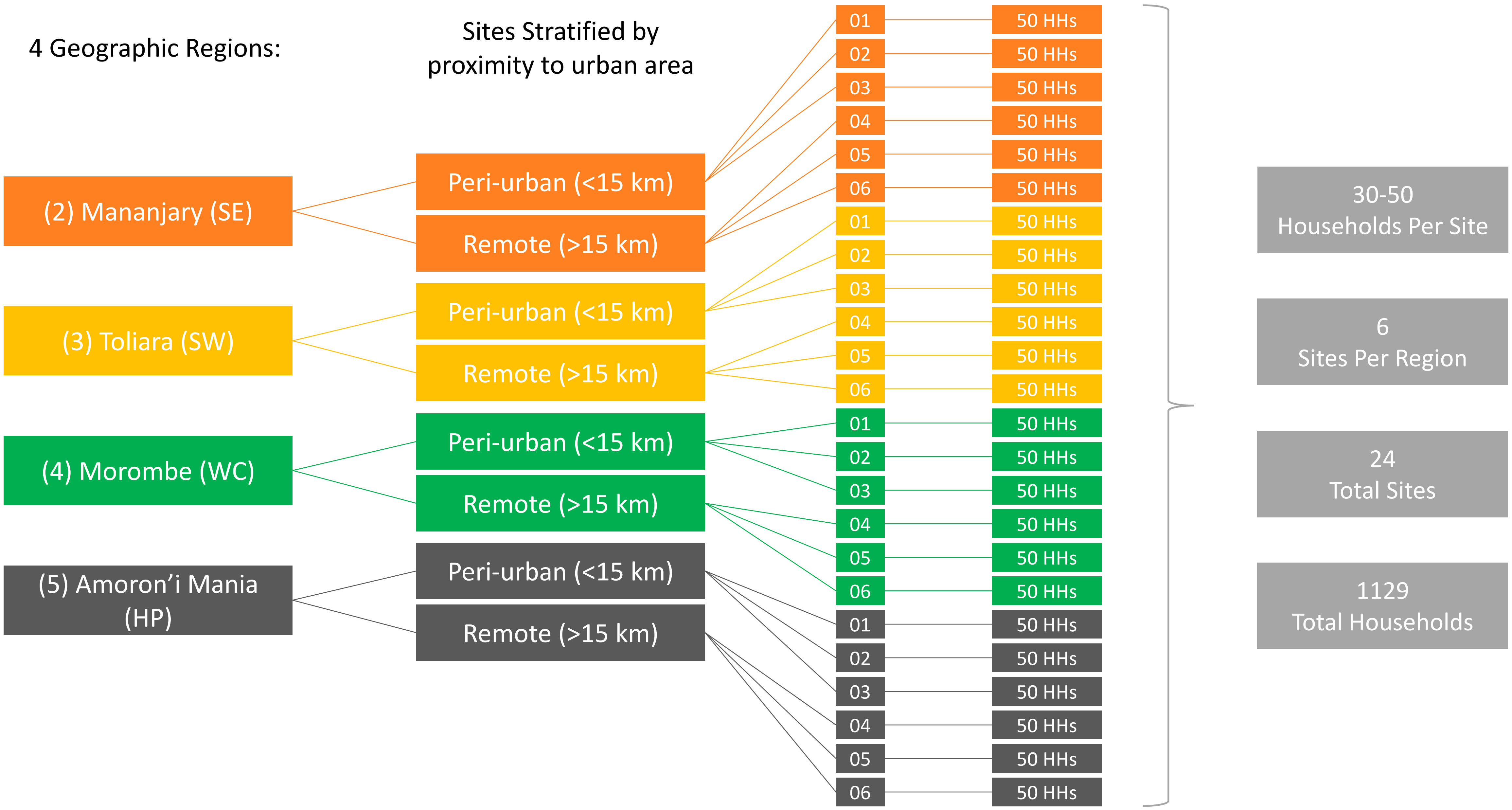


Figure 2: Variation in the percentage of individuals infected (prevalence) between sites within regions and between regions

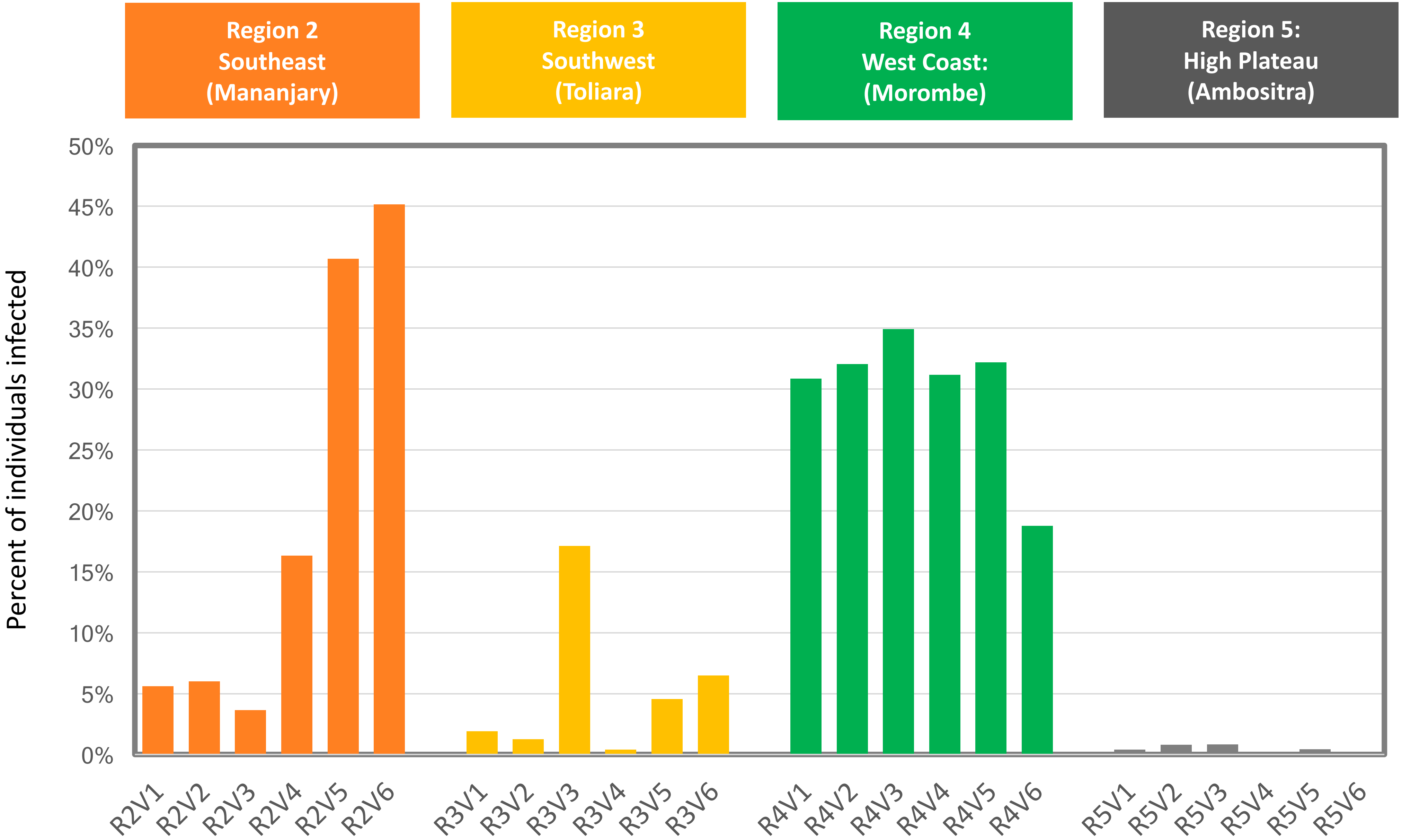
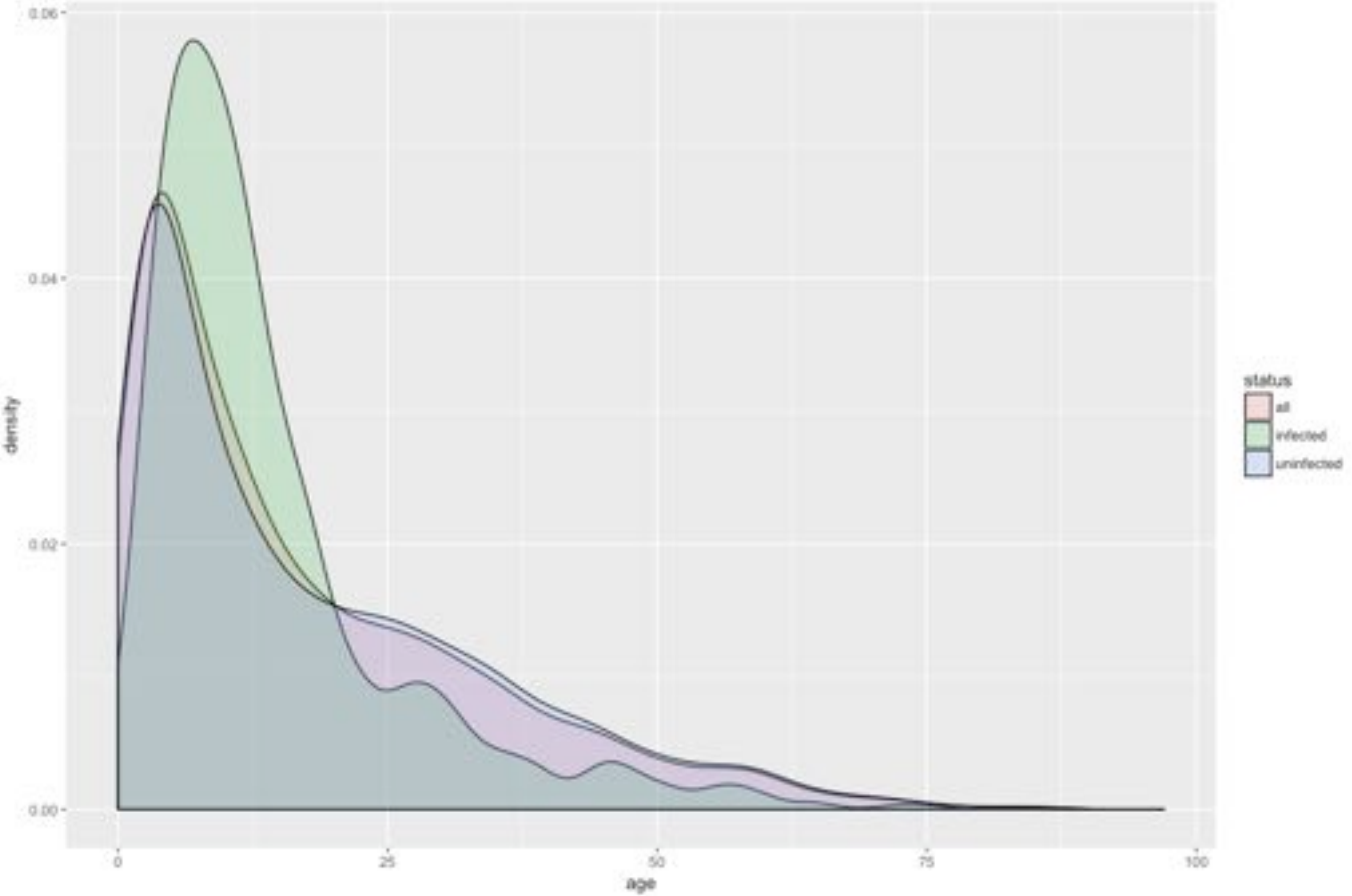
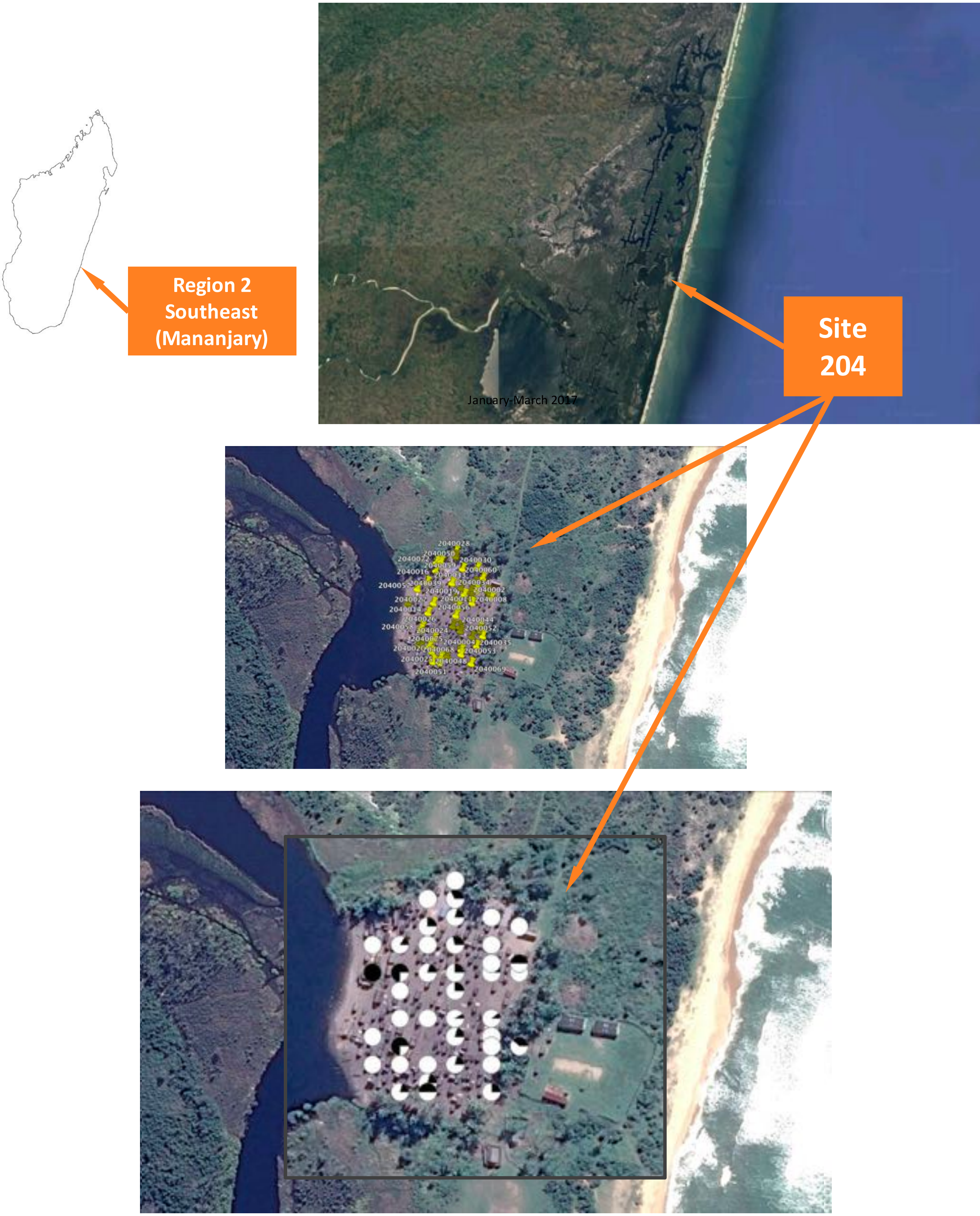


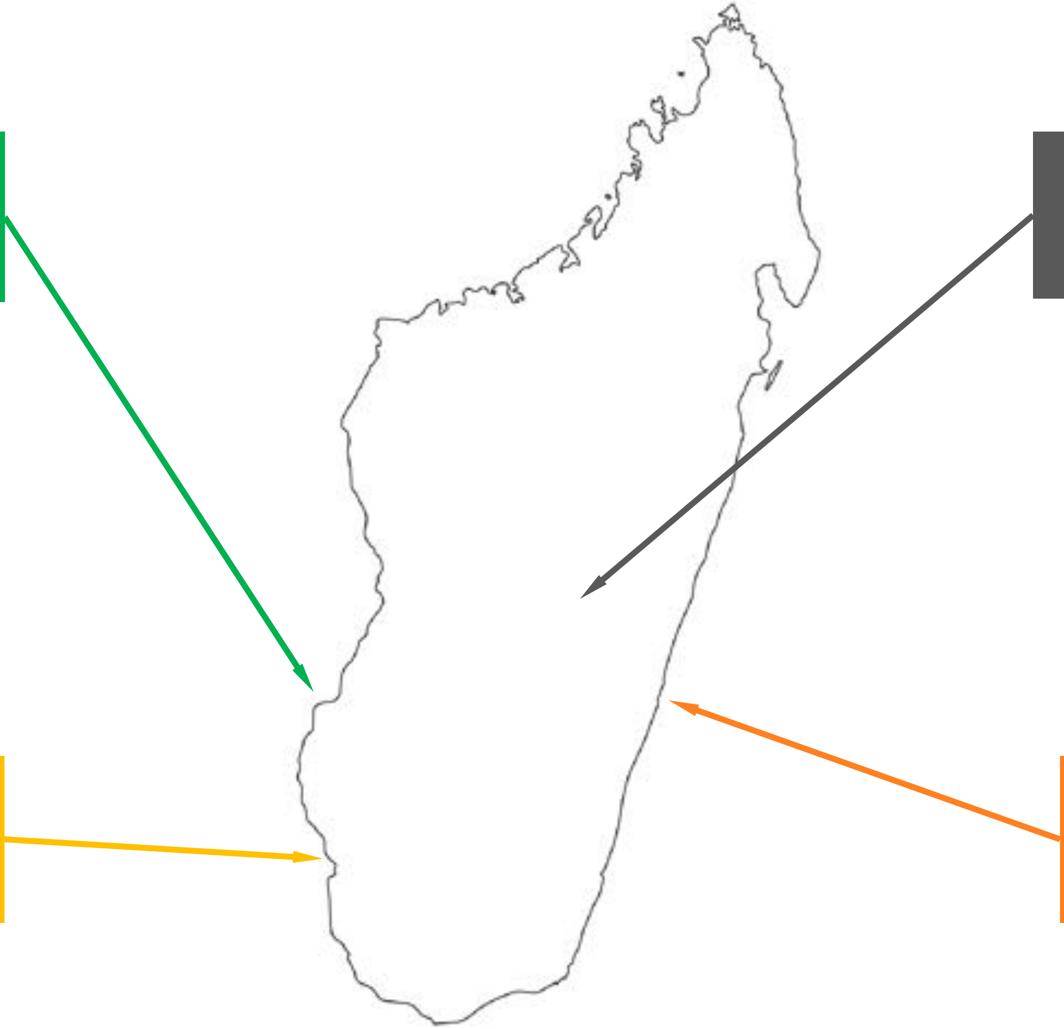
Figure 3: Density plots of age for all individuals, those infected with malaria, and uninfected with malaria (n = 5533)



S Figure 1: Site 204 as an example of infections observed to cluster at the household level. Pie charts show the proportions of infected individuals within households in black



Supplementary Figure 2: Map of Madagascar showing the regions sampled and some climatic characteristics



Region 4
West Coast:
(Morombe)

Low rainfall; high temperature

Region 5:
High Plateau
(Ambositra)

High rainfall; low temperature

Region 3
Southwest
(Toliara)

Very low rainfall; very high temperature

Region 2
Southeast
(Mananjary)

High rainfall; high temperature