Mid-Project Update

Dylan Bradizza, Miles Abssy, Ofek Gross, Samuel Dumas

November 2nd 2023

Abstract

In this research project we are interested in examining the Ross-Macdonald model, which consists of 7 parameters used to describe the dynamics of Malaria transmission. In our work we would like to determine whether all of the seven Ross-Macdonald model parameters are equally necessary to reliably fit simulated data. We are also interested in optimizing the performance of our code, so that the run time for fitting the model is decreased. Findings from our research would clarify the importance of each parameter in the Ross-Macdonald model. Furthermore, our findings may enhance the computational efficiency for subsequent and similar research.

Mathematical Models

Mathematical models are descriptions of biological processes which describe states under depicted assumptions and can help with an intuitive understanding for how those processes may unfold in nature [1]. Moreover, they can be used to

- clarify assumptions we may have about how complex systems work
- build understanding of how processes in complex systems interact
- generate predictions and hypothesis
- test predictions and hypothesis based on analysis of the model of fitting to data
- determine what data is needed to learn about something (or if it is possible to, given the data that is available, reliably infer parameters of interest)
- determine in what ways certain kinds of data may be problematic or biased

A common application of models is to analyze infected population system dynamics. There are many models which match this characterization, each tailored to specific infections, conditions, and assumptions.

Model Selection

The model of choice for this project is the Ross-MacDonald model, which can be characterized as a SI model (suceptible, infectious) [2]. The model can be devised as seen below [3].

$$\frac{dI_h}{dt} = ab\frac{I_v}{H}(H - I_h) - \gamma I_h \tag{1}$$

$$\frac{dI_v}{dt} = ac\frac{I_h}{H}(V - I_v) - \mu I_v \tag{2}$$

Brief History: Malaria and Mathematical Modelling

Malaria is a deadly mosquito-borne disease caused by Plasmodium spp [4], [5]. Malaria is primarily transmitted by mosquitoes (Anopheles), whose bite allows the entry of Plasmodium into the bloodstream. Once in a human host, infected individuals present with a set of non-specific symptoms, such as diarrhoea, fever, vomiting, and pulmonary complications [4]. Despite continuous eradication efforts, Sub-Saharan Africa bears the brunt of this disease. In many cases, malaria leads to death - in 2021 alone, UNICEF reported more than 600,000 malaria-related fatalities [6]. In an effort to better understand vector-human relationships, many researchers have developed mathematical models, commonly differentialequation-based. One of the most well-known is the Ross-MacDonald model, a simple SI model used to predict disease dynamics over time [2]. The Ross-Macdonald model was first proposed by Sir Ronald Ross in the 1910s, and later extended by George MacDonald in the late 1950s [7], [8]. Interestingly, the model would be finalized toward the ending of the Global Malaria Eradication Programme, which ran from 1955 to 1969 [7]. Since then, others have made contributions to the model, aiming to improve its accuracy. Frequently, this comes in the form of considering fixed variables to be dynamic, such as population size, or by introducing new variables such as quarantined individuals.

While more sophisticated model may exist, the usefulness of the Ross-Macdonald model is not lost upon, and remains a valiant model for initial characterization of human-vector interactions in epidemiology [2].

Assumptions

Like any model, the Ross-Macdonald model has a set of assumptions which must be considered throughout data simulating and analysis. One of the most important assumptions of this model is that the population is closed, that is neither the human population nor the vector population may have new individuals introduced. Furthermore, that infections must strictly occur between a human and vector (i.e., human to vector or vector to human). Lastly, the model assumes that vector bites are evenly distributed throughout the human population and that mosquitoes must always have some opportunity to feed, to which the human is randomly selected [8]. These assumptions form the theoretical framework of the model, however, flexibility can be promoted through the inclusion of implicit assumptions concerning the range of the parameters.

Parameter Analysis

There are seven parameters that are incorporated into the Ross-Macdonald Model, that is a, b, c, H, V, γ, and μ [3]. More specifically, a, b, H, and γ are used within Eq. (1), and a, c, H, V, μ are used within Eq. (2). When fixing the following parameters, it is crucial that the selected values are standardized to the same unit of time (e.g., days, months, years). In this study, the selected unit of time was days. Included below is the range that will be assumed for parameters.

- mosquito biting rate (a): 0.1 - 1 [9],[10],
- transmission probability from infectious mosquito to susceptible human per bite (b): 0.01 - 0.8 [9], [10], [11],
- transmission probability from infectious human to suceptible mosquito per bite (c): 0.072 - 0.64 [9], [10], [11],
- ratio of mosquitoes to human ($m = V/H$): 1 - 10 [12], [13], [14],
- recovery rate of humans (γ): 0.005 - 0.05 [9], [10], [15], [16],
- mortality rate of mosquitoes (μ): 0.05 - 0.33 [10], [15].

Although this value is not a parameter, we may briefly consider the R_0 value for this model, as it is entirely defined by these parameters [3].

$$R_0 = \sqrt{\frac{a^2bcm}{\mu\gamma}} \qquad (3)$$

Notably, when considering if the disease will persist the square root can be omitted. As $\forall n \in \mathbb{R}^+, \sqrt{n} > 1 \implies n > 1$, and similarly $0 < \sqrt{n} < 1 \implies 0 < n < 1$. Although it remains important to consider the quantitative ranges of variables, understanding the nature of these parameters and how they observed is important. Albeit, some of the parameters within this model are easier to measure with sufficient accuracy. Parameters for which this holds may be referred to as knowns, and parameters for which this is contrary, we may address as unknowns.

Knowns

There are two parameters that can be reliably inferred in most contexts, that is the size of the human population and the recovery rate. There are many techniques that can be used to estimate the size of populations where conducting a census of the population is feasible and reliable. Notice that both of these tend to be diminished with respect to mosquito populations. Moreover, it is possible to track how long it takes for a human individual to recover. The conditions required for these to be well known is often contingent on the economic resources available and the spatial range of the population. Although biting rate cannot be measured as efficiently, there are some approaches that would be enable reasonable estimate (e.g., counting the number of individuals on a random sample of humans over time). Hence, the known parameters for this model are a, H, and γ.

Unknowns

The remaining variables b, c, V, and μ are therefore the unknown variables for this model. Focusing on the latter half, consider that mosquitoes are quite small and numerous, which makes estimating their population size difficult. This has been previously approached using mark-recapture experiments, but reported confidence intervals were quite large which reflects the uncertainty present with estimating mosquito populations. Consequently, it can be quite difficult to further determine the rate at which the mosquitoes are dying [17]. A survival function was also considered in the studied mentioned there above, with similar inaccuracies. Furthermore, determining both the probability of an infected mosquito biting a human and the probability of an infected human being bite by a mosquito is quite complex, particularly when taking stochasticity into account. These probabilities are often the consequence of interactions between other parameters within the model (i.e., larger H may increase b, and larger V may increase c). With respect to the difficulties expressed concerning reliably estimating these parameters, it may be of interest to explore the possibility of bypassing taking measurements to derive a value. Inferring parameters can be done by means of likelihood, to which this study aims to research.

Aims and Objectives

As discussed in the previous section, some parameters may be easier to determine, or at the very least approximate. Fortunately, one can use likelihood to see how well certain parameters fit the data, to then estimate what is most likely the true parameter [1]. While likelihood is a formidable tool in inferring parameters, accuracy is not always guaranteed. Ultimately, some parameters may prove to yield weaker predictions, indicating that these values cannot be inferred through likelihood. Conversely, some parameters may prove to be anchors to predictions made, indicating that these values are crucial to accurate estimation, and consequently, the model as a whole. From this, we may delineate the following null hypotheses.

- H_0: All (seven) of the parameters are required to reliably fit the data.
- H_A: There exists some subset of parameters, $S \subset \{a, b, c, H, V, \gamma, \mu\}$, which can be used to reliably fit the data.

Observing subsets of any size at most seven can be quite demanding computationally $\sum_{n=1}^{7} \binom{7}{n}$, and thus it may be better to observe subsets up to a smaller size; as there were four unknown parameters, this will be the

largest size considered in this study. This yields a total of 98 distinct subsets. Therefore, using likelihood, the hypothesis aims to verify how many parameters are required to explain the data.

Model Setup

Firstly, we must define the Ordinary Differential System (ODE) system within R using the package `deSolve` [18]. Here, we are using `dIh` to denote the change in infectious humans, and `dIv` to denote the change in infectious vectors.

```
rossMacOde <- function(times, state, params){
  with(as.list(c(state,params)),{
    dIh <- a * b * Iv / H * (H - Ih) - gamma * Ih
    dIv <- a * c * Ih / H * (V - Iv) - mu * Iv

    return(list(c(dIh, dIv)))
  })
}
```

As indicated in the *Model Selection* section, the Ross-MacDonald model comprises of two differential equations. While the code above will correctly implement the desired model, it remains important to ensure that initial conditions and time intervals are compatible with the parameters. Notably, it must be that $I_{v0} \leq V, I_{h0} \leq H$, and that the time intervals be in units of days. It is unlikely that initial infected populations are close to the entire population size in the instances that are tested, as there tends to be minimal variation, but it may be worth giving slight overview for a complete understanding of the model logistics when addressing required data.

Now that our ODE system has been setup, data can be inputted into our model to simulate infected population numbers through time. Below is a template delineating which data should be stored and where to store it; these are denoted using the angle brackets ($< >$). This requires fixing seven parameters, two initial values, and an overall time frame. As time should be considered using days as units and it is frequently of interest to consider data over some number of years (n), `times` should be defined by `seq(1,365*n)`.

```
params <- c(a, b, c, H, V, gamma, mu)

ivInt <- <initial infected vectors>; ihInt <- <initial infected humans>
times <- <time stamps to be considered>

state <- c(Ih = ihInt, Iv = ivInt)

out <- as.data.frame(ode(state, times, rossMacOde, params)) %>%
  pivot_longer(! time)
```

There are two stable states that occur within this model, that is the number of infections decipitates to zero, or the number of infected individuals humans and vectors reaches some capacity (bounded at least by H and V, respectively). This can be seen in the following equations.

$$I_h* = 0 = I_v* \Longrightarrow \frac{dI_h}{dt} = ab\frac{0}{H}(H-0) - \gamma \cdot 0 = 0, \ \frac{dI_v}{dt} = ac\frac{0}{H}(V-0) - \mu \cdot 0 = 0 \tag{4}$$

It may be the case that certain sets of parameters and initial values require significant time to reach their equilibrium, however, the time frame shall hopefully allow for most models to reach this state.

Parameter Data

The seven parameters that are used within the Ross-Macdonald model occupy the largest role within the model with respect to outcomes over time. Consider that R_0 is entirely defined by these seven parameters, and thus, the parameters alone whether the number of infections will grow within the model. While initial values are also important as subsequent data is generated based on the initial state, these computations are done with respect to the parameters.

Within this study, a total of 1728 combinations of parameters will be tested. These combinations were generating using 2, 3, 3, 4, 4, 3, and 2 values for a, b, c, H, V, γ, and μ, respectively. This can be imported by running the following code chunk.

```
parData <- read_csv("EEB313_Project_Data.csv")
```

The majority of the parameter combinations within this data set (41/48 to be specific) have an R_0 greater than 1. While it remains within the scope of this study to observe models where the infection dies down, there is a stronger emphasis on situations where the infection prevails. An initial viewing of the data can be seen below.

```
#str(parData)
head(parData)
```

```
## # A tibble: 6 x 11
##   paramN     a     b     c     H     V     m gamma    mu     R0 Persist
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <lgl>
## 1      1   0.3   0.1  0.08   100   100     1 0.25    0.1 0.170  FALSE
## 2      2   0.3   0.1  0.08   100   100     1 0.25    0.3 0.0980 FALSE
## 3      3   0.3   0.1  0.08   100   100     1 0.01    0.1 0.849  FALSE
## 4      4   0.3   0.1  0.08   100   100     1 0.01    0.3 0.490  FALSE
## 5      5   0.3   0.1  0.08   100   100     1 0.005   0.1 1.2    TRUE
## 6      6   0.3   0.1  0.08   100   100     1 0.005   0.3 0.693  FALSE
```

Although this has been made evident thus far, this data set does not contain initial values nor time intervals. However, this data set provides a great foundation for introducing these quantities, to then begin the implementation of likelihood.

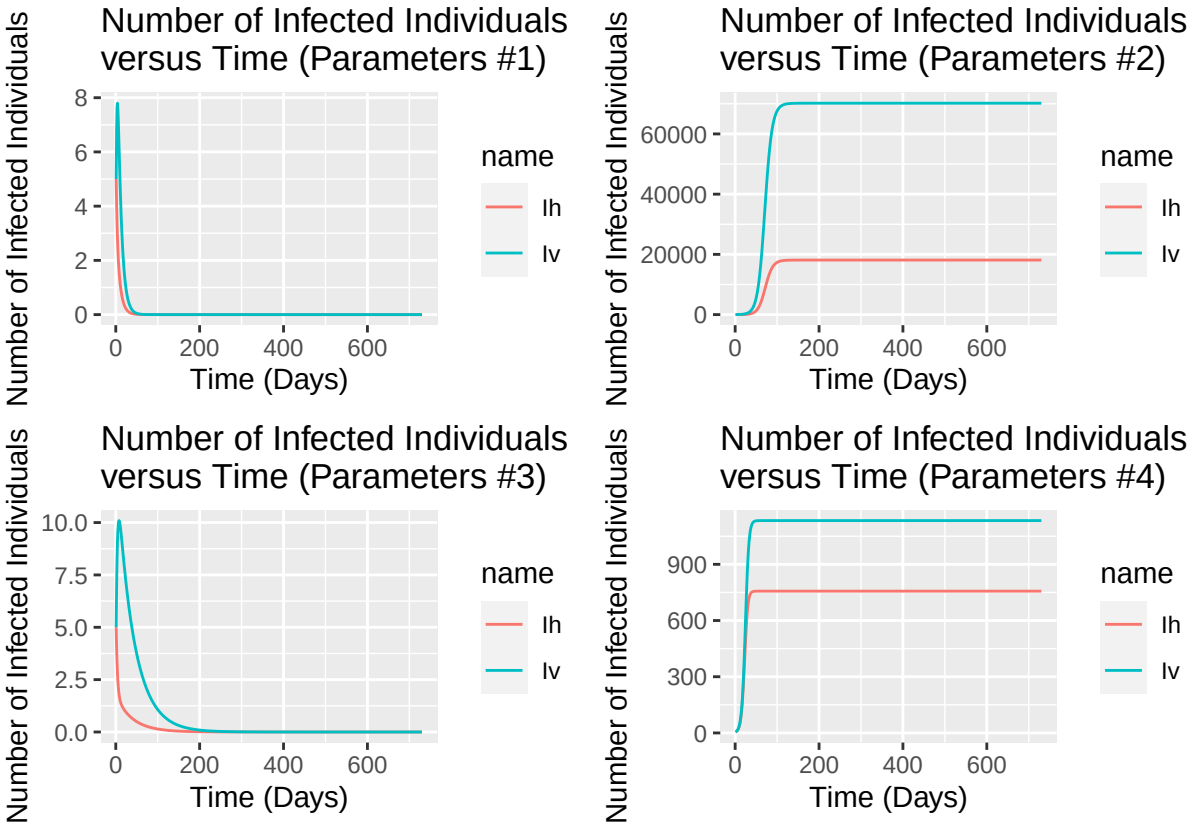
Initial Values and Time

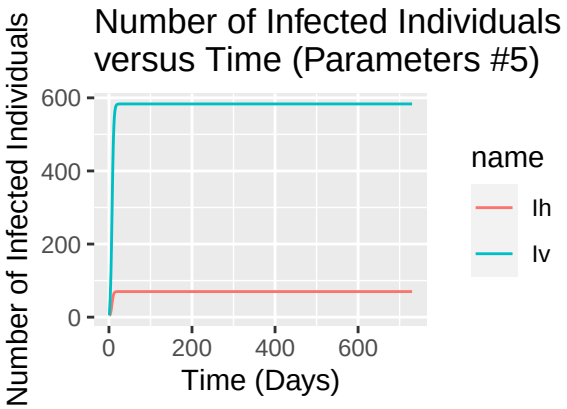
The Ross-Macdonald model requires two initials values, that is the initial number of infected humans (I_{h0}) and the initial number of infected vectors (I_{v0}). Unlike the parameter m, there are no constraints on the ratio of infected individuals from each group (aside from being bounded by H and V respectively). Consequently, selecting pairs of initial infected populations should be done according to those parameters. By slightly changing initial values, observing how likelihood varies on a smaller scale can be thoroughly done. Moreover, time should be considered such that the model reaches equilibrium for any set of parameters, at least as an initial exploration. It might be useful/interesting to observe the strength of likelihood even if the model has yet to reach equilibrium. Consider that even in real data (ignoring dynamic influence), the data may not have yet reach a stable/unstable state. After observing simulated data from a sample of plots, 10 years seems to be roughly appropriate for a time frame.

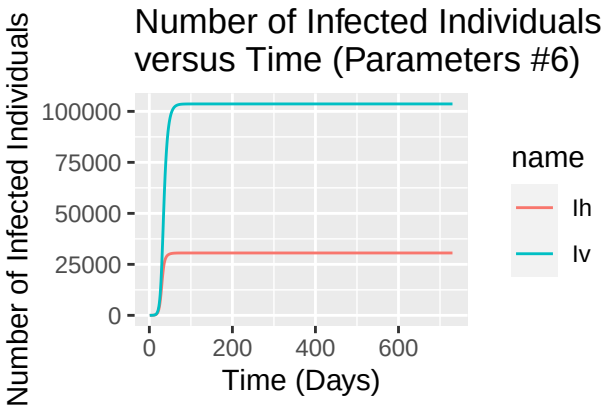
```
buildN <- function(ih,
                   iv,
                   m,
```

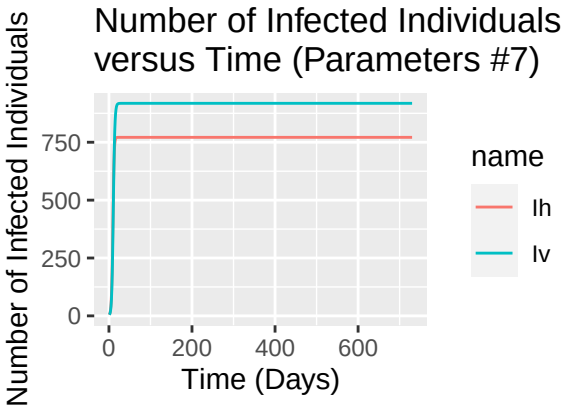
```
                    n = 10){
  dflist <- list()
  times <- seq(1, 365*m)
  state <- c(Ih = ih, Iv = iv)
  parI <- sample(1:1728, 10, replace = F)
  for (i in 1:n){
  params <- parData[parI[i],]
  out <- as.data.frame(ode(state, times, rossMacOde, params)) %>%
     pivot_longer(! time)
  dflist[[i]] <- out
  }
  return(dflist)
}
```

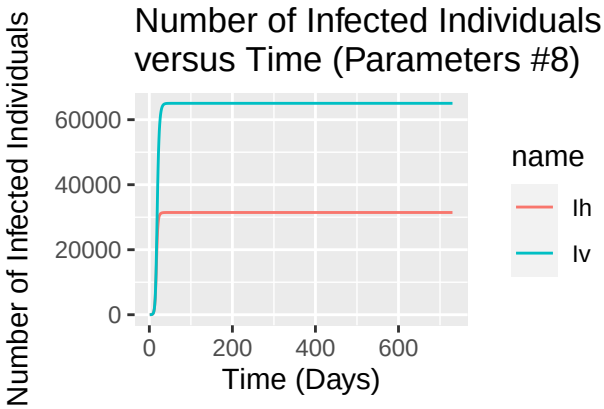
The function `buildN` creates a list containing ten random sets of parameters (without replacement) from `parData`. More or less draws can be made, however ten is the default. Notably, the data frames contained within the elements of this list have 3 columns; time (t), infected humans (I_h), and infected vectors (I_V). These can then be plotted, comparing infected abundance against time.

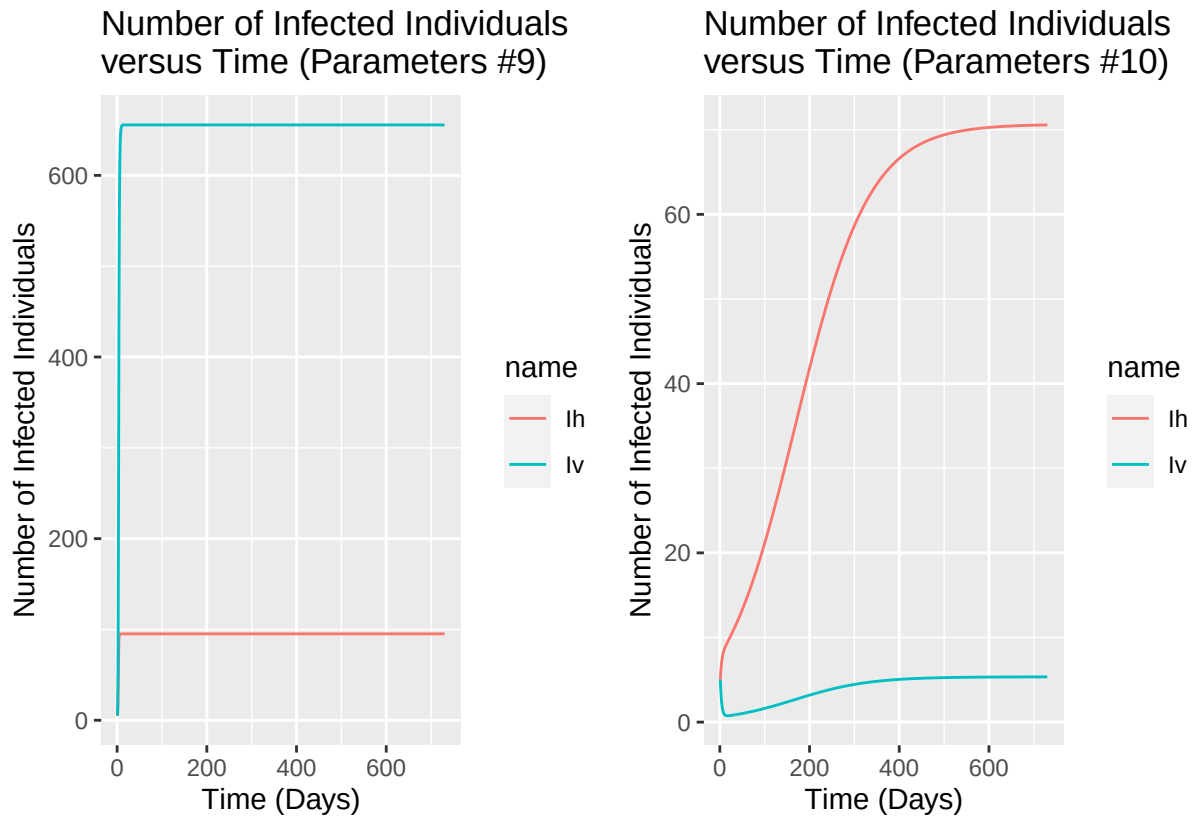



Number of Infected Individuals versus Time (Parameters #5)
Number of Infected Individuals
600
400
200
0
0
200
400
600
Time (Days)
name
Ih
Iv


Number of Infected Individuals versus Time (Parameters #6)
Number of Infected Individuals
100000
75000
50000
25000
0
0
200
400
600
Time (Days)
name
Ih
Iv


Number of Infected Individuals versus Time (Parameters #7)
Number of Infected Individuals
750
500
250
0
0
200
400
600
Time (Days)
name
Ih
Iv


Number of Infected Individuals versus Time (Parameters #8)
Number of Infected Individuals
60000
40000
20000
0
0
200
400
600
Time (Days)
name
Ih
Iv



Within the ten plots above, there is already some noticeable variation between their outcomes. While fixing the y-axis across the plots would ensure a more appropriate analysis, the main purpose of these plots if to introduce how some of the population dynamics appear within this model.

It should be noted that during the creation of data, it was briefly conjectured that picking a unique m may be more valuable than unique H and V. While the general shape may be preserved for similar m, the graphs tend to be scaled with respect to time. Moreover, if introducing stochasticity at any point within this study, smaller populations will be subjected to greater effects than the larger counterparts. Now that all the fundamental aspects of the model, data, and restrictions have been covered, likelihood can be considered.

Likelihood

As stated previously, the intent of this project is to fit simulated data using some finite subset of parameters to fit the data. Using likelihood assumes that data are normally distributed and there are equal variance among data points. The pseduo code below can be used to solve the differential equation for a fixed set of parameter values and some given range for parameters of interest. Using this information, the code then returns the parameters which optimize for likelihood.

```
ihInt <- <number of initial infected humans>
ivInt <- <number of initial infected vectors>
times <- seq(1, 365*m)
state <- c(Ih = ih, Iv = iv)

params <- expand.grid(a = <constant> or <seq(start, end, steps)>,
                      b = <constant> or <seq(start, end, steps)>,
```

```
                      c = <constant> or <seq(start, end, steps)>,
                      H = <constant> or <seq(start, end, steps)>,
                      V = <constant> or <seq(start, end, steps)>,
                      mu = <constant> or <seq(start, end, steps)>,
                      gamma = <constant> or <seq(start, end, steps)>,
                      kappa = <seq(start, end, steps)>,
                      p = <seq(start, end, steps)>
                     )

return_LL_at_specific_combo_params <- function(params_to_use, data){

  reporting_times <- data$time

  out <-
    as.data.frame(ode(state, times, rossMacOde, params_to_use)) %>%
    subset(time %in% reporting_times)

  LLh <- c()
  LLv <- c()

  for (i in 1:length(reporting_times)){
    LLh[i] <- dbinom(data$Ih[i], size = round(params_to_use$kappa*out$I[i]),
                     prob = params_to_use$p)
    LLv[i] <- dbinom(data$Iv[i], size = round(params_to_use$kappa*out$I[i]),
                     prob = params_to_use$p)
  }

  LogLik <- sum((log(LLh) + log(LLv))/2)

  return(data.frame(params_to_use, LogLik = LogLik))
}

LogLikihoods <- NULL
outALL <- NULL

for (i in 1:nrow(params)){
  LogLikihoods <- rbind(LogLikihoods,
                        return_LL_at_specific_combo_params(params[i,], data)
                        )
  outALL[[i]] <- data.frame(ode(state, times, rossMacOde, params[i,1:7]), index = i)
}

MLE <- LogLikihoods %>% subset(is.finite(LogLik)) %>%
  subset(LogLik == max(LogLik)); MLE
```

Noticeably, there are two distinct sets of data (human and vector) which must be fitted using likelihood. It will be interesting to observe how likelihood determines parameters (i.e., the existence of any biases, preference of variables in both equations), given the number of parameters and their important role throughout the model. Currently, the overall log-likelihood is computed using an avarege of the likelihood across the two groups; a standard average as opposed to weighted is more appropriate for this context.

Moreover, the set of parameters optimzied for likelihood can be thereafter contrasted against the *true* parameters that were used to simulate the data. Plotting the set of parameters in the same plot enables deductions to be made regarding how well the data are fitted, particularly which segments may be better fitted.

```r
outALL <- do.call(rbind, outALL)

best_solution <- outALL %>%
  subset(index == which(LogLikihoods$LogLik == max(LogLikihoods$LogLik))) %>%
  group_by(time) %>%
  mutate(expected_measurement = MLE$kappa*MLE$p*I)

best_solution %>% ggplot() +
  geom_line(aes(x = time, y = expected_measurement), color = "black") +
  geom_point(data = data, aes(x = time, y = Ih), size = 2) +
  geom_point(data = data, aes(x = time, y = Iv), size = 2) +
  geom_line(aes(x = time, y = Ih), color = "red") +
  geom_line(aes(x = time, y = Iv), color = "blue")
```

Next Steps

The material explored in the sections above depict the necessary framework for this study, though there may be slight modifications made to procedures if deemed necessary. This may come in the form of additional data, even altering the current values, or changes to code. Although the likelihood code has been proposed, actuall analysis on this aspect have yet to be observed, and thus this is a focal point of this study going forward. Moreover, while likelihood is currently average across the two groups, it may be of interest to observe how frequently the best solution set of parameters is also the highest optimized set for likelihood within the two individual groups. This could be useful when observing the existence of any biases. Therefore, the required tools are all put into place, and now, it is simply a matter of applying them.

Delegation of Tasks

- Abstract: Ofek
- Mathematical Models: Sam
- Model Selection: Dylan
  - Brief History: Dylan, Miles, Ofek, Sam
  - Assumptions: Miles
  - Parameter Analysis: Miles, Sam
    - Known: Sam
    - Unknown: Sam
- Aims and Objective: Sam
- Model Setup: Sam
  - Parameter Data: Sam
  - Initial Values: Sam
- Likelihood: Sam
- Next Steps: Sam
- Review & Edits: Dylan, Miles, Ofek, Sam

References

[1] V. M. Zhang and M. K. Yuksel, "Quantitative methods in r for biology," *EEB313*. Nov. 2023. Available: https://eeb313.github.io/

[2] S. Mandal, R. R. Sarkar, and S. Sinha, "Mathematical models of malaria - a review," *Malaria Journal*, vol. 10, no. 1, 2011, doi: 10.1186/1475-2875-10-202.

[3] X. Jin, S. Jin, and D. Gao, "Mathematical analysis of the ross–MacDonald model with quarantine," *Bulletin of Mathematical Biology*, vol. 82, no. 4, 2020, doi: 10.1007/s11538-020-00723-0.

[4] M. A. Phillips, J. N. Burrows, C. Manyando, R. H. van Huijsduijnen, W. C. Van Voorhis, and T. N. Wells, "Malaria," *Nature Reviews Disease Primers*, vol. 3, no. 1, 2017, doi: 10.1038/nrdp.2017.50.

[5] J. C. Tavares, *Malaria*. Morgan & Claypool, 2013.

[6] UNICEF, *UNICEF*. May 2023. Available: https://data.unicef.org/topic/child-health/malaria/#:~:text=Nearly%20every%20minute%2C%20a%20child%20under%20five%20dies%20of%20malaria,to%20619%2C000%20deaths%20in%20total

[7] D. L. Smith, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie, "Ross, MacDonald, and a theory for the dynamics and control of mosquito-transmitted pathogens," *PLoS Pathogens*, vol. 8, no. 4, 2012, doi: 10.1371/journal.ppat.1002588.

[8] K. Magori and J. M. Drake, "The population dynamics of vector-borne diseases," *Nature Education Knowledge*, vol. 4, Jan. 2013.

[9] N. Chitnis, J. M. Hyman, and J. M. Cushing, "Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model," *Bulletin of Mathematical Biology*, vol. 70, no. 5, pp. 1272–1296, 2008, doi: 10.1007/s11538-008-9299-0.

[10] G. G. Mwanga, H. Haario, and V. Capasso, "Optimal control problems of epidemic systems with parameter uncertainties: Application to a malaria two-age-classes transmission model with asymptomatic carriers," *Mathematical Biosciences*, vol. 261, pp. 1–12, 2015, doi: 10.1016/j.mbs.2014.11.005.

[11] W. Gu, G. F. Killeen, C. M. Mbogo, J. L. Regens, J. I. Githure, and J. C. Beier, "An individual-based model of plasmodium falciparum malaria transmission on the coast of kenya," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 97, no. 1, pp. 43–50, 2003, doi: 10.1016/s0035-9203(03)90018-6.

[12] R. M. Anderson and R. M. May, *Infectious diseases of humans: Dynamics and control*. Oxford university press, 1992.

[13] J. L. Aron, "Mathematical modelling of immunity to malaria," *Mathematical Biosciences*, vol. 90, no. 1–2, pp. 385–396, 1988, doi: 10.1016/0025-5564(88)90076-4.

[14] D. Gao *et al.*, "Optimal seasonal timing of oral azithromycin for malaria," *The American Journal of Tropical Medicine and Hygiene*, vol. 91, no. 5, pp. 936–942, 2014, doi: 10.4269/ajtmh.13-0474.

[15] R. Anderson, *The population dynamics of infectious diseases: Theory and applications*, 1982, doi: 10.1007/978-1-4899-2901-3.

[16] S. Ruan, D. Xiao, and J. C. Beier, "On the delayed ross–MacDonald model for malaria transmission," *Bulletin of Mathematical Biology*, vol. 70, no. 4, pp. 1098–1114, 2008, doi: 10.1007/s11538-007-9292-z.

[17] D. Cianci *et al.*, "Estimating mosquito population size from mark–release–recapture data," *Journal of Medical Entomology*, vol. 50, no. 3, pp. 533–542, 2013, doi: 10.1603/me12126.

[18] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer, "Solving differential equations in R: Package deSolve," *Journal of Statistical Software*, vol. 33, no. 9, pp. 1–25, 2010, doi: 10.18637/jss.v033.i09.