scientific reports

Check for updates

OPEN

Inference of malaria reproduction numbers in three elimination settings by combining temporal data and distance metrics

Isobel Routledge[1]✉, H. Juliette T. Unwin[2] & Samir Bhatt[2]

Individual-level geographic information about malaria cases, such as the GPS coordinates of residence or health facility, is often collected as part of surveillance in near-elimination settings, but could be more effectively utilised to infer transmission dynamics, in conjunction with additional information such as symptom onset time and genetic distance. However, in the absence of data about the flow of parasites between populations, the spatial scale of malaria transmission is often not clear. As a result, it is important to understand the impact of varying assumptions about the spatial scale of transmission on key metrics of malaria transmission, such as reproduction numbers. We developed a method which allows the flexible integration of distance metrics (such as Euclidian distance, genetic distance or accessibility matrices) with temporal information into a single inference framework to infer malaria reproduction numbers. Twelve scenarios were defined, representing different assumptions about the likelihood of transmission occurring over different geographic distances and likelihood of missing infections (as well as high and low amounts of uncertainty in this estimate). These scenarios were applied to four individual level datasets from malaria eliminating contexts to estimate individual reproduction numbers and how they varied over space and time. Model comparison suggested that including spatial information improved models as measured by second order AIC (ΔAICc), compared to time only results. Across scenarios and across datasets, including spatial information tended to increase the seasonality of temporal patterns in reproduction numbers and reduced noise in the temporal distribution of reproduction numbers. The best performing parameterisations assumed long-range transmission (> 200 km) was possible. Our approach is flexible and provides the potential to incorporate other sources of information which can be converted into distance or adjacency matrices such as travel times or molecular markers.

Individual-level disease surveillance data, collected routinely and as part of outbreak response, capture a wealth of information which could improve measurements of transmission and its spatiotemporal variation, in turn informing the design of epidemiological interventions[1]. Geo-located health facility or residence data are increasingly collected as part of surveillance for diseases such as malaria[2], but could be more effectively utilised to infer transmission dynamics, in conjunction with additional information such as symptom onset time and genetic distance[3]. However, challenges exist in making use of these diverse data sources and leveraging the information they contain within a single inference framework. This is particularly true of endemic diseases such as malaria, where individual level data are increasingly collected in moderate-low transmission or elimination settings, but unlike many epidemic viral diseases[4–8], the development of approaches to analyse such data has been relatively recent[3,9–11].

Malaria transmission is shaped by processes occurring on a wide range of spatial scales. In the absence of human mobility, transmission is limited to the range of the mosquito vector, however human movement, ranging from regular commutes to seasonal or one-off migration events across longer distances can import parasites into new areas provided competent vectors are present[12–15]. As a result, the spatial location of individual cases can provide useful information in inferring transmission dynamics when combined with additional forms of information, such as temporal and molecular data. Furthermore, in elimination settings malaria transmission is thought to take on epidemic dynamics[16], meaning the importance of space and highly dynamic factors such as

[1]University of California, San Francisco, San Francisco, USA. [2]Imperial College London, London, UK. ✉email: isobel.routledge@ucsf.edu

human movement patterns becomes more relevant. However, in the absence of data about the flow of parasites between populations, the spatial scale of malaria transmission (here meaning the distance between locations where one person was infected or detected and the locations where another person received an infective bite) is often not clear. As a result, it is important to understand the impact of varying assumptions about the spatial scale of transmission on key metrics of malaria transmission, such as reproduction numbers. In many contexts, not all cases may be observed within a surveillance system. Missing cases further complicate inference, as without information about how likely cases are to be missing, ambiguity can exist as to whether long-range transmission occurred or whether cases were infected by a closer, unobserved source of infection.

To explore the impact of including distance measures and assumptions about their relationship to transmission likelihood, we developed a flexible framework to incorporate pairwise distances (for example Euclidian distances, travel times, or any quantifiable distance matrix) into our previously published inference framework[11] to estimate individual reproduction numbers (R_c or case reproduction numbers) and explore the impact of varying assumptions about the likelihood of a case having a source of infection not observed in the dataset and the spatial kernel (the function describing the relationship between distance metric and likelihood of transmission occurring) on results, as well as determining the feasibility of inferring the distance kernel and amount of missing cases from surveillance data. Briefly, the model builds upon previous diffusion network approaches[10,11,17–20] which estimate the likelihood of transmission networks using known or inferred times of infection, by also allowing additional distance information to inform inference. They provide a flexible framework to integrate multiple data types, and have been evaluated using real and simulated transmission processes at multiple scales and under varying network structures[21].

We defined twelve scenarios representing different assumptions about the likelihood of transmission occurring over different geographic distances and likelihood of missing infections (as well as high and low amounts of uncertainty in this estimate). These scenarios were applied to four individual level datasets from malaria eliminating contexts to estimate individual reproduction numbers and how they varied over space and time. We used two simple spatial kernels describing the relationship between Euclidian distance between residences and likelihood of transmission occurring, to explore various assumptions about the relationship between locations of cases and likelihood of transmission occurring between them, as well as the impact of unobserved cases. We find the best performing models by second order AIC (ΔAICc)[22] have weakly informative priors on the likelihood of unobserved sources of infection and assume that cases can be linked by transmission across larger spatial scales (over 200 km). However, we find there can be issues of parameter identifiability, which become increasingly relevant when there are not enough data available about key parameters in the model.

Results

We developed a framework to integrate distance information into a previously published inference framework[10,11] which uses the time of symptom onset to infer reproduction numbers and their spatiotemporal variation. We then tested the impact of varying assumptions about the relationship between location of cases and the likelihood of transmission as well as the impact of unobserved infection as modelled by competing edges, ε, considering twelve scenarios (Table 1), and applying them to four line-list datasets from China (*P. vivax* and *P. falciparum*, analysed separately), El Salvador (*P. vivax*) and Eswatini (*P. falciparum*) using Exponential and Gaussian spatial kernels, described in the methods section of this paper.

Results of model comparison by ΔAICc across different scenarios. When ΔAICc scores were used to compare model results, all models which included distance had lower (and therefore better) ΔAICc scores than models which included only time (Table 2 and Supplementary Table 1). In addition, exponential kernels consistently outperformed equivalent scenarios using Gaussian kernels (Supplementary Table 1). Two scenarios consistently performed best as measured by ΔAICc, namely Scenario 9 (El Salvador and Eswatini) and Scenario 11 (China, *P. vivax* and *P. falciparum*). Both scenarios assume longer range human movement is likely and impose a smaller penalty on cases occurring larger distances. These scenarios also allow variation in epsilon edge values and use a very weakly informative prior on epsilon edges, but with a different mean (0.1 for Scenario 9, 0.001 for Scenario 11). These results also return smaller mean R_c results than time-only versions of the model (Figs. 1, 2, 3, 4).

R_c estimates under different scenarios. Across all datasets, large differences in R_c estimates were found depending on boths ε and β parameters. When β is higher, the assumption is that there is little movement of parasites within the country and therefore cases with residential addresses which are far away are unlikely to have infected each other. When this is the case and we assume there are unobserved sources of infection (either through a strongly informative prior on ε with mean 0.1, or an uninformative prior with a lower mean), then R_c values are very low. However if we assume there are little or no unobserved sources of infection, but continue to make restrictive assumptions about space, then most R_c values are very low, but in the localities where there are cases, we estimate much higher R_c values as there are no other possible infectors within a reasonable time and/or spatial area. This is illustrated in Figs. 1, 2, 3 and 4.

When looking at the spatial patterns of R_c estimates under different scenarios several trends are seen across all datasets (Figs. 5, 6, 7, 8). Scenario 4 is particularly interesting to note because this scenario considers the most restrictive assumptions, both about space and unobserved sources of infection. Across datasets, Scenario 4 results in increased focality and higher R_cs within these foci, but in comparison lower R_cs in other areas. All of the best scenarios as measured by ΔAICc resulted in small R_c estimates, but where comparably larger R_c estimates were estimated, they were in localities identified as foci.

Scenario description	Scenario	Beta (fixed)	Epsilon (prior)
Human movement unlikely, most movement under 10 km	1	Gaussian = 0.005	Mean = 0.1
Missing cases more likely (but very uncertain)		Exponential = 0.1	SD = 1
Human movement unlikely, most movement under 10 km	2	Gaussian = 0.005	Mean = 0.1
Missing cases more likely (confident)		Exponential = 0.1	SD = 0.001
Human movement unlikely, most movement under 10 km	3	Gaussian = 0.005	Mean = 0.001
Missing cases less likely (but very uncertain)		Exponential = 0.1	SD = 1
Human movement unlikely, most movement under 10 km	4	Gaussian = 0.005	Mean = 0.001
Missing cases less likely (confident)		Exponential = 0.1	SD = 0.001
Moderate human movement, most movement under 50 km	5	Gaussian = 0.001	Mean = 0.1
Missing cases more likely (but very uncertain)		Exponential = 0.02	SD = 1
Moderate human movement, most movement under 50 km	6	Gaussian = 0.001	Mean = 0.1
Missing cases more likely (confident)		Exponential = 0.02	SD = 0.001
Moderate human movement, most movement under 50 km	7	Gaussian = 0.001	Mean = 0.001
Missing cases less likely (but very uncertain)		Exponential = 0.02	SD = 1
Moderate human movement, most movement under 50 km	8	Gaussian = 0.001	Mean = 0.001
Missing cases less likely (confident)		Exponential = 0.02	SD = 0.001
Longer range human movement likely	9	Gaussian = 0.0001	Mean = 0.1
Missing cases more likely (but very uncertain)		Exponential = 0.01	SD = 1
Longer range human movement likely	10	Gaussian = 0.0001	Mean = 0.1
Missing cases more likely (confident)		Exponential = 0.01	SD = 0.001
Longer range human movement likely	11	Gaussian = 0.0001	Mean = 0.001
Missing cases less likely (but very uncertain)		Exponential = 0.01	SD = 1
Longer range human movement likely	12	Gaussian = 0.0001	Mean = 0.001
Missing cases less likely (certain)		Exponential = 0.01	SD = 0.001

Table 1. Illustrating the different scenarios and corresponding parameter values tested in scenario analysis.

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

Table 2. Summary of ΔAICc results.

For the line-list dataset from El Salvador, within the range of values explored in the sensitivity analysis (Table 3), regardless of how informative the prior was for either β, the distance shaping function, or for ε, the epsilon edge, β, was always estimated as whatever the mean of the prior was set as between the prior mean values of 1e−4 and 1e−2 (Fig. 9). However, when the mean value was set at 0.1, the estimated parameter converged at a slightly lower value of 0.075, with the exception of when the prior for ε was very low (all priors with mean ε of 1e−10 and also the more informative priors with mean 1e−5, when standard deviation was 1e−4). R_c is strongly shaped by the value of ε, with higher values of ε returning lower values of R_c, however R_c also declined with increasing values of β.

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

For both *P. vivax* and *P. falciparum* datasets from China, within the parameter range explored in the sensitivity analysis, regardless of how informative the prior was for either β, the distance shaping function, or ε, the epsilon edge, β was always estimated as whatever the mean of the prior was set as (Figs. 11, 12), suggesting a lack of identifiability or information within the data. When estimating R_c, and interesting interacting effect of ε (missing or unobserved infections) and β (distance) was seen. When β is low, although lower values of ε produce slightly higher mean R_c values, the difference in R_c estimates with varying prior values for ε is much smaller

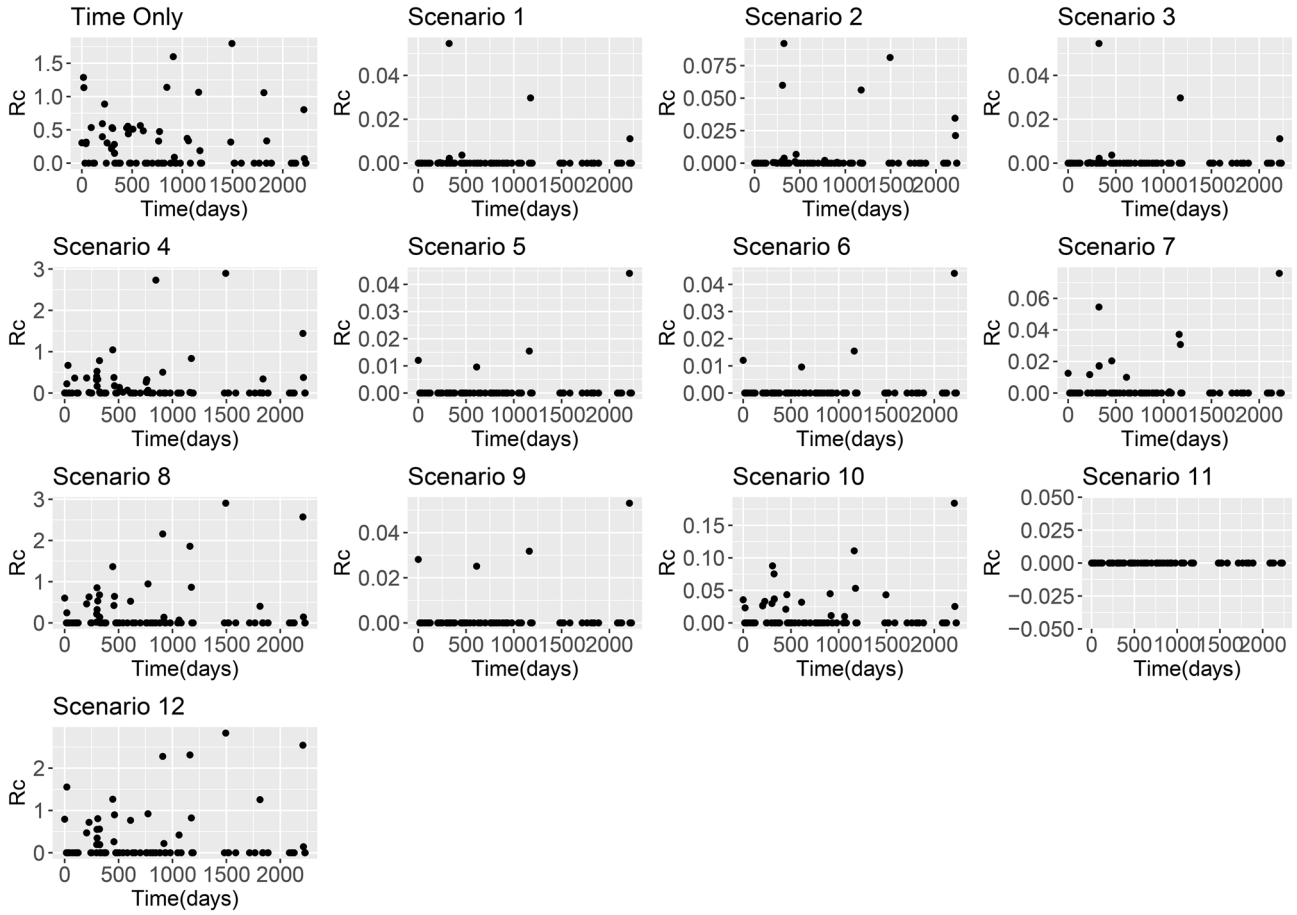

Figure 1. R_c estimates from El Salvador line list based on using the time-only scenario and Scenarios 1–12 with an exponential kernel.

than when β is a higher value. In other words, when the prior for ε is low, 1e−10, R_c estimates do not vary as β changes, however when the prior for ε is much higher, then increasing β from 1e−4 to 0.1 reduces R_c estimates (from 0.21 to 0.01 for *P. vivax)*.

Discussion

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

Approaches such as the one presented here can be useful tools for making the most of available surveillance data in near elimination and elimination settings. As more countries collect detailed individual level data, and reach lower case counts which make measures such as parasite prevalence less informative, there is more potential for approaches which leverage the detailed information contained in individual level data. Reproduction numbers can be used to measure the impact of elimination interventions, seasonality and identify areas of higher receptivity to transmission, if importation were to occur there, and when incorporated into geostatistical models with spatially resolved covariates can be used to create maps of receptivity[9–11].

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

Exponential Kernels consistently outperformed Gaussian kernels as measured by ΔAICc. Although classic models of dispersion are as a diffusion process with Gaussian displacement, more leptokurtic or "fatter-tailed"

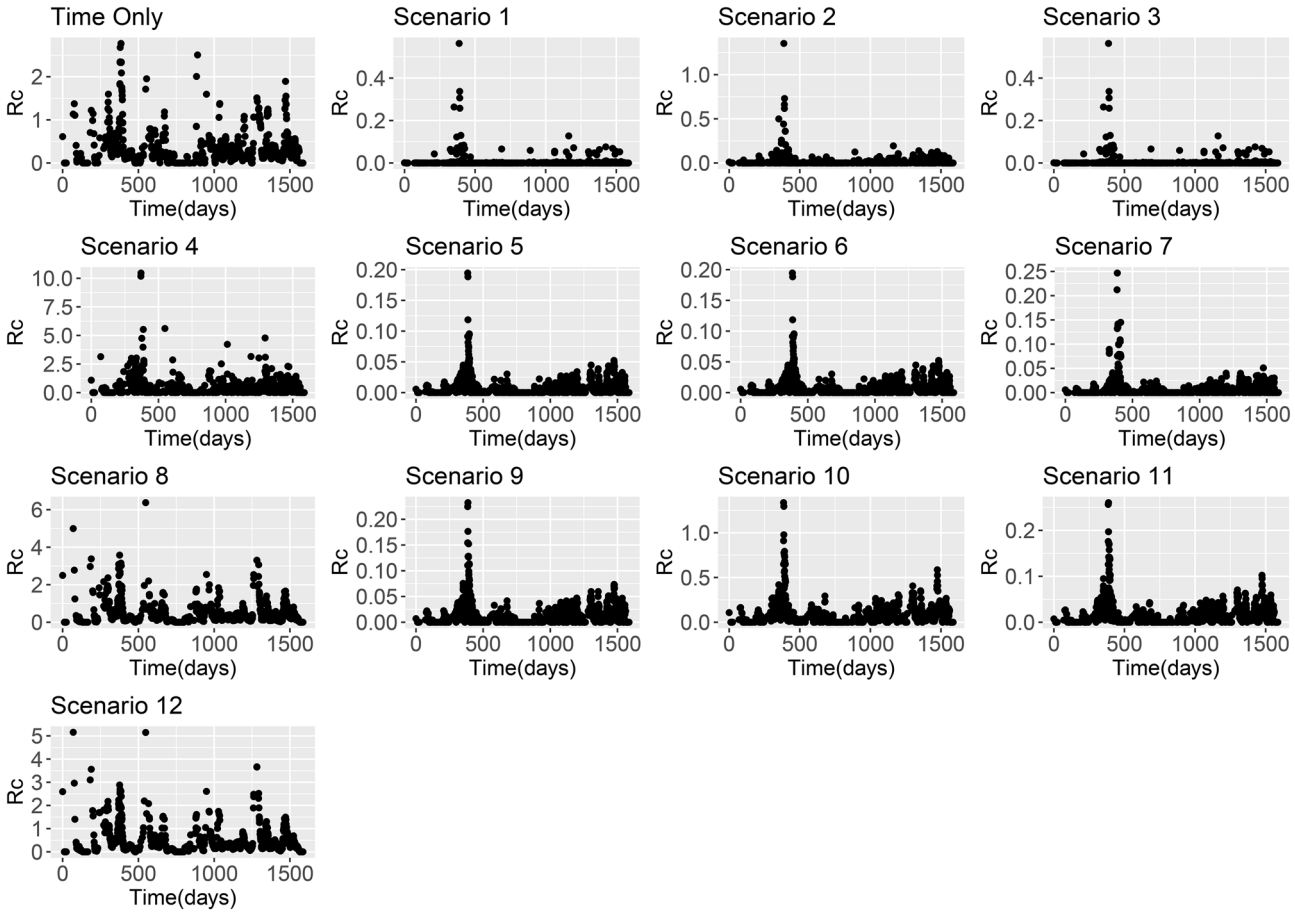

Figure 2. R_c estimates from Eswatini line list based on using the time-only scenario and Scenarios 1–12 with an exponential kernel.

probability distributions, where more of the probability density is concentrated in the tails of the function, are often found to better represent empirical dispersal patterns than traditional Gaussian kernels[23]. This "fatter-tail" in the exponential can be seen in Figs. 13, 14 and 15.

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

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

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

There are several limitations to this approach and analysis. Firstly, there is a potential lack of identifiability between ε, the epsilon edge, and β, the shaping parameter of the spatial kernel. To give an intuitive example, say two cases occurred 50 km from each other in space within a reasonable timeframe of symptom onset times for transmission to have occurred. Without strong prior information about what the spatial kernel may be, and/

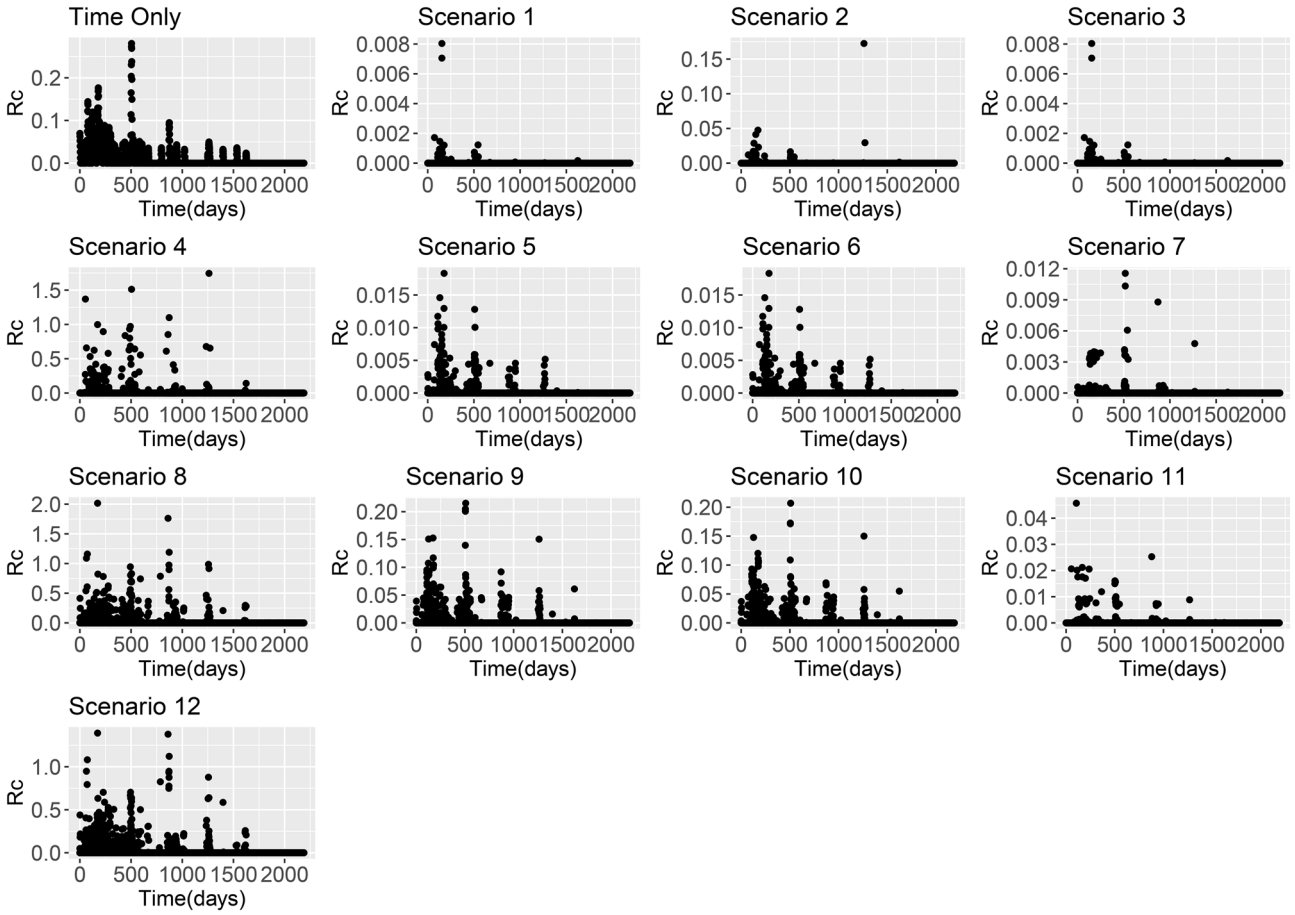

Figure 3. R_c estimates from China *P. falciparum* line list based on using the time-only scenario and Scenarios 1–12 with an exponential kernel.

or how likely cases are to have an external source of infection, it is not clear whether these cases are linked by transmission (and there is some human travel/parasite movement, modelled by a less restrictive spatial kernel) or whether there are unobserved source(s) of infection leading to both cases. This is also exemplified in the results of the sensitivity analysis, where the mean of the prior for beta strongly shapes the final estimate of beta, and the epsilon value also shapes beta.

In the absence of reliable information about either of these values, strong assumptions must be made about either/both the likelihood of cases being infected by unobserved sources of infection and the relationship between distance and. Similar approaches analysing the diffusion of twitter hashtags, it was recommended to fix the parameter beta, and the authors acknowledged potential challenges in estimating this parameter[17], and indeed approaches from others have also noted problems with unconstrained distance kernels in space–time diffusion modelling (Swapnil Mishra, personal correspondence). One potential way to address this is to divide the epsilon edge by the distance parameter $\frac{\varepsilon}{\beta}$, thereby linking the two parameters and thereby penalising increases in β.

Whilst the temporal aspect is not fixed, we view the utility in this method in excluding or penalising improbable transmission links between far away cases, rather than as a way of trying to determine what the spatial relationship between cases is for malaria transmission, or determining the relative contribution of space to malaria transmission. Recent simulations[24] using temporal, spatial, and travel history to infer malaria reproduction numbers also found that travel history must be highly accurate or infections must be focal in time to be able to accurately estimate reproduction numbers.

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

Secondly, our approach was designed for application to near elimination and elimination settings, where surveillance and case management is very strong, numbers of cases are small, and therefore there is less overlap in potential infector/infectees, and changes in transmission are more apparent. If applying these approaches to contexts which are less far along the journey to elimination, the issue of identifiability may be even more of an issue as one cannot reasonably assume/fix epsilon edges to be a very small number. Asymptomatic infection will likely be more important to consider, more sophisticated methods to deal with missing cases will be required. There also will likely be a weaker signal in space and time, which may require the integration of additional

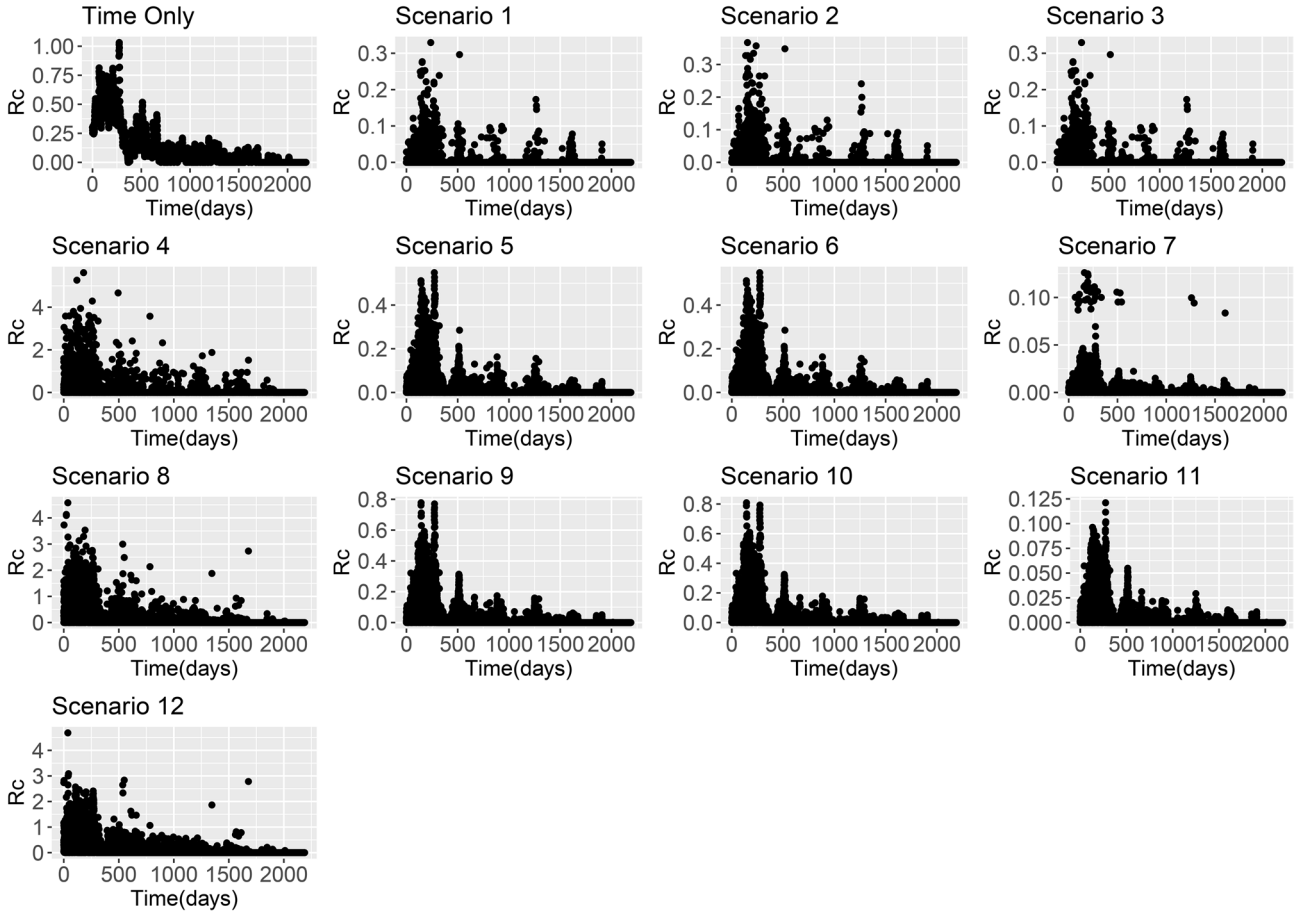

Figure 4. R_c estimates from China *P. vivax* line list based on using the time-only scenario and Scenarios 1–12 with an exponential kernel.

information such as genetic distance. There also will be a transmission level above which these methods will no longer be useful, although we do not know what this exact level is. The spatial granularity at which it is possible to estimate Rc at different transmission intensities also remains unknown.

Finally, due to there being no "ground truth" it is challenging to rigorously compare model performance. ΔAICc and Akaike Weights are standard measures for model comparison, however as mentioned previously, there are important limitations in using these metrics for model comparison. We did test the model inferred Rc values on a series of simulated line lists (Supplementary Text 1, Supplementary Fig. 1) which were spatially explicit, where the "true" expected Rc was known. However, a useful future step would be to carry out more in-depth analysis to investigate the impact of varying parameter values and the interaction between the shaping parameter of the spatial kernel and epsilon. Further simulations may also aid in exploring how tolerant the method is to missingness.

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

Many methods used to model and represent space and mobility have not been tested here due to the issues of identifiability seen even in simple models of space. Gravity, radiation[25], and friction surfaces[26] are all potentially useful models of how space may affect the likelihood of transmission. As mosquitoes have a limited range and lifespan, developing better data and models of human movement, and how it varies in different cultural contexts and between different demographic groups, will provide useful information to appropriately parameterise and design the spatial component of the model[13,27–29].

Although the prior for the shaping parameter of the serial interval was selected under the assumption that the majority of cases are treated in a timely manner, In this analysis we have not explicitly utilised information about the time and location of treatment, although this is available in some contexts. This may be useful information to constrain the potential time window of infection occurring, as detailed information about infectivity and gametocyte carriage following treatment with anti-malarials is available[30], although sub-optimal dosage, compliance

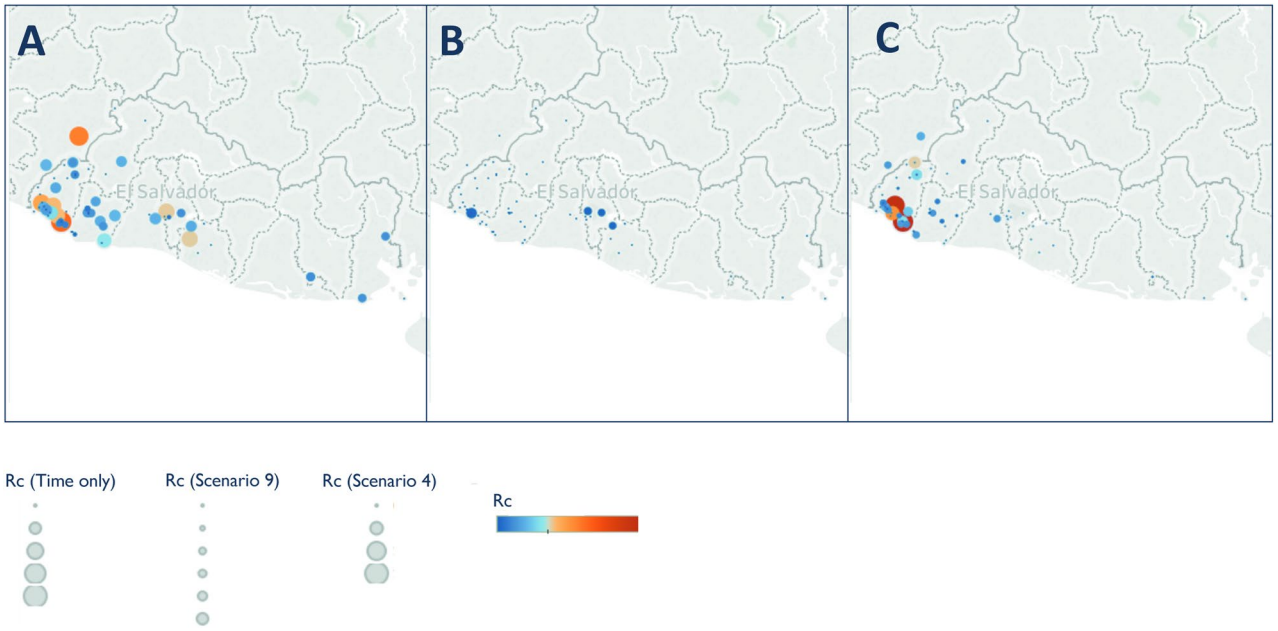

Figure 5. Map of Rc estimates for El Salvador Map of (**A**) time-only (**B**) best scenario by AIC (Scenario 9) and (**C**) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases. Note the increasing focality in (**C**), with higher Rc values estimated on the Pacific Coastal area of the Ahuacapan and Sonsonate municipalities, where the NMCP have long identified as the remaining foci of risk.

and resistance have been associated with differing outcomes and therefore having additional information about treatment and prevalence of resistance would also be useful.

Another avenue for future work would be to adapt the approach to incorporate further sources of information, such as genetic markers of similarity between parasites. For our approach to be useful in contexts which are not at or within a few years of elimination, incorporation of additional information into the inference framework will be required. This could be carried out either directly by incorporating an additional term or function in the likelihood or indirectly through informing the value of parameters and allowing them to vary between individuals. Previous work within the machine learning and network analysis community has successfully integrated diverse sources of information about texts such as language and similarity of context into very similar algorithms to the one presented here[17].

Conclusion

Increasingly, line-list data contain spatial and other forms of information. Developing rigorous approaches to leverage the information contained within these diverse datasets will increasingly be useful in malaria surveillance and epidemiology[2,3,13] and developing a framework which flexibly takes on different forms of data within an integrated inference framework is a key aspect of this. There may be more useful information contained in genetic, and or travel, mobility data. However, as we have seen there can be issues of identifiability, which becomes increasingly relevant when there is not enough data available about key parameters in the model. Finding ways for leveraging multiple datasets, understanding their relationships, how they can enhance info contained in others, or used to build consensus is important.

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

Methods

Data. *The Kingdom of Eswatini.* This dataset, previously analysed by Reiner et al.[31] captures malaria cases recorded by the National Malaria Elimination Programme in the Kingdom of Eswatini (formally known as Swaziland) between January 2010 and June 2014. For each case detected during this time (N = 1373), case investigation was carried out. For each case the following were collected: GPS coordinates of household location,

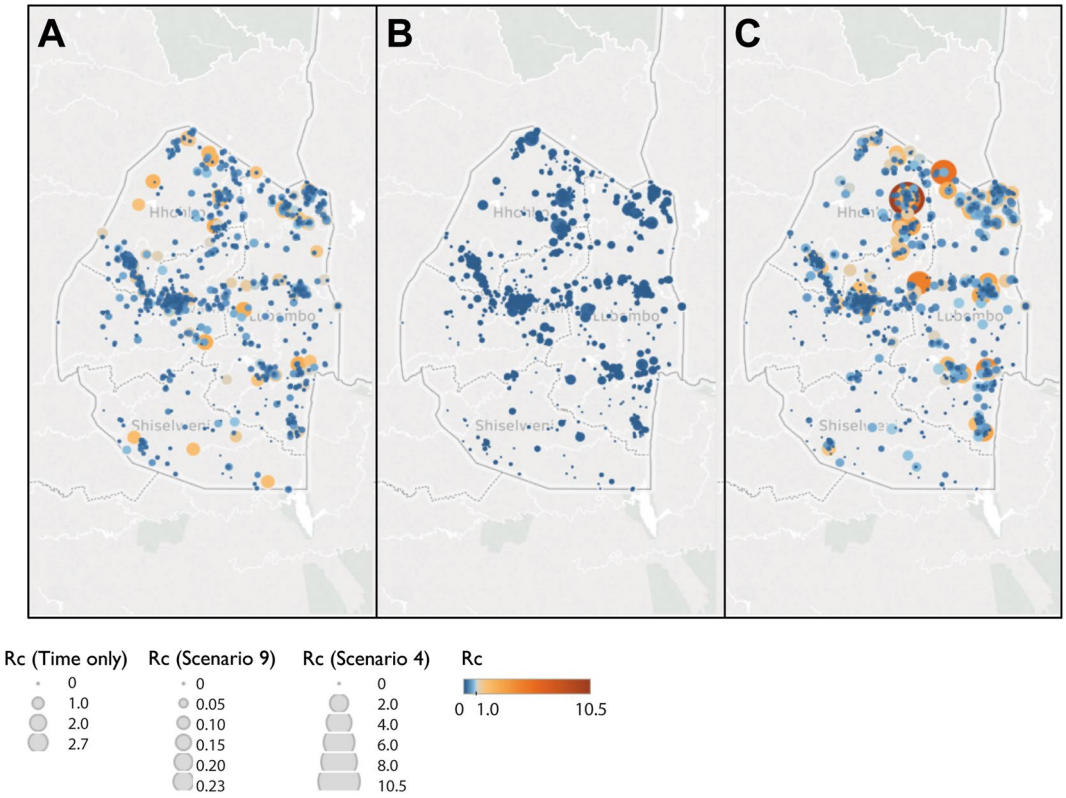

Figure 6. Map of Rc estimates for Eswatini Map of (**A**) time-only (**B**) best scenario by AIC (Scenario 9) and (**C**) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases. Note the increasing focality in (**C**), with higher Rc values estimated around the northern corner of the country which borders Mozambique.

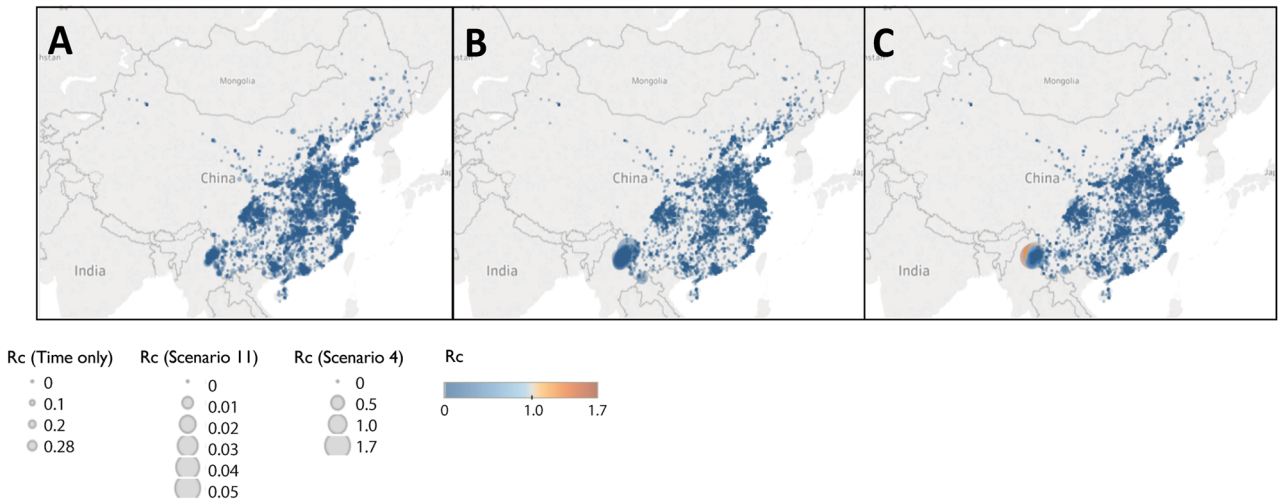

Figure 7. Map of Rc estimates for *P. falciparum* in China Map of (**A**) time-only (**B**) best scenario by AIC (Scenario 11) and (**C**) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases.

demographic information (age, occupation and sex), use of malaria prevention interventions such as long-lasting insecticide treated bednets (LLINs), and date of symptom onset, diagnosis and treatment, as well as travel history. Based on travel history cases were defined as locally acquired, imported. For a small number of cases (N = 58) the local/imported status was determined "unknown". For the purposes of this analysis, these cases were

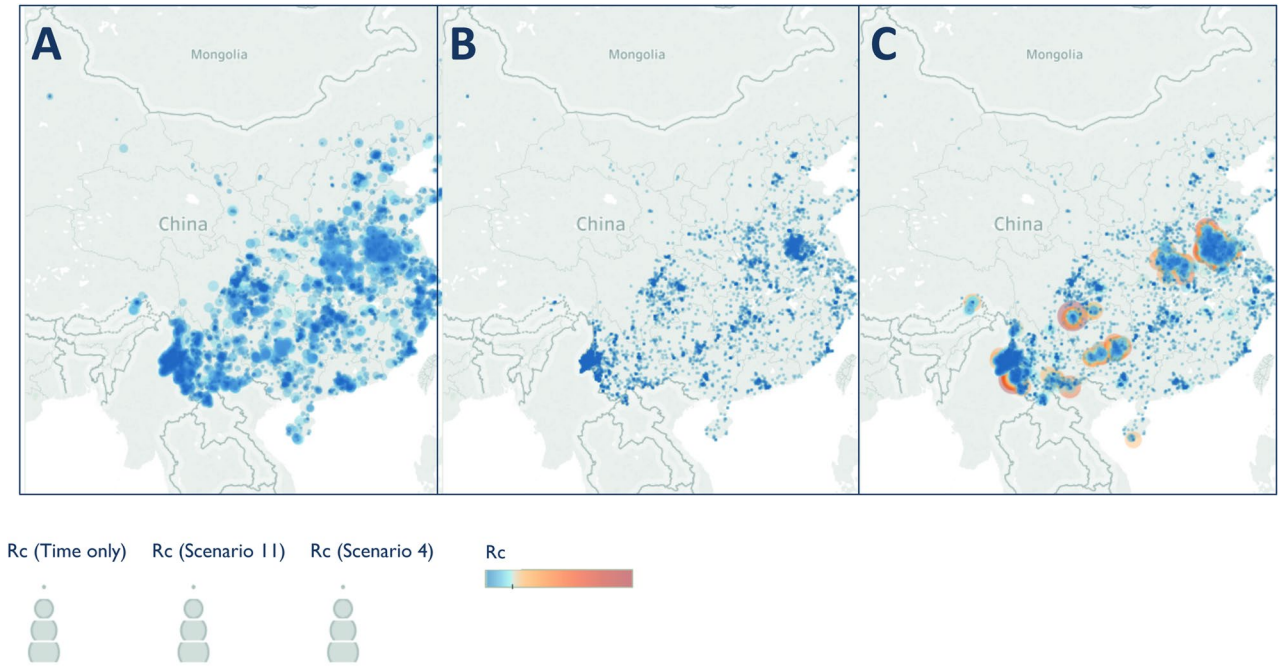

Figure 8. Map of Rc estimates for *P. vivax* in China Map of (**A**) time-only (**B**) best scenario by AIC (Scenario 11) and (**C**) Scenario 4, representing an assumption of little long-distance transmission and few unobserved cases.

∈ mean	∈ SD	β mean (Gaussian)	β mean (exponential)	β SD
1e−10	0.0001	0.00001	0.0001	0.0001
1e−5	0.001	0.0001	0.001	0.001
1e−3	0.01	0.001	0.01	0.01
1e−2	0.05	0.01	0.1	0.05
1e−1	0.1			0.1
0.5				

Table 3. Different parameters considered in sensitivity analysis. Note all combinations of each parameter were considered.

treated the same as local cases, i.e. they were assumed to have potentially been infected by other cases in the dataset and/or been infectors themselves.

China. This dataset consists of individual-level case data for all confirmed and probable cases reported in China between 2011 and 2016[11,32] (Tables 4, 5). The data consist of an individual identifier, date of symptom onset, date of diagnosis and date of treatment, as well as the geolocated address of residence and health facility. If the suspected location of infection was in China and not in the same district, then the presumed location of infection was also included in the dataset. Demographic information such as age and sex were also collected. For the analysis, data were separated into *P. falciparum* and *P. vivax* cases. *P. malariae* (N = 252) and *P. ovale* (N = 822) were reported but excluded from the analysis due to the lower public health concern of these species.

El Salvador. This dataset consists of all confirmed cases of malaria recorded by the Ministry of Health in El Salvador between 2010 and the first two months of 2016[10] (N = 91 cases, of which 30 imported, 6 *P. falciparum*, 85 *P. vivax*). For each case, the date of symptom onset was recorded. Residential address was available for all but two cases. For these cases, the location was available at the *municipio,* or municipality level, and the coordinates of the centroid of the municipality (which for both were cities) were used as the geo-location. Two cases had addresses listed outside of El Salvador, both of which were located in Guatemala. All cases within El Salvador with full addresses (N = 85) were georeferenced by latitude and longitude to *caserío/lotificación* level, which is approximately neighbourhood or hamlet level.

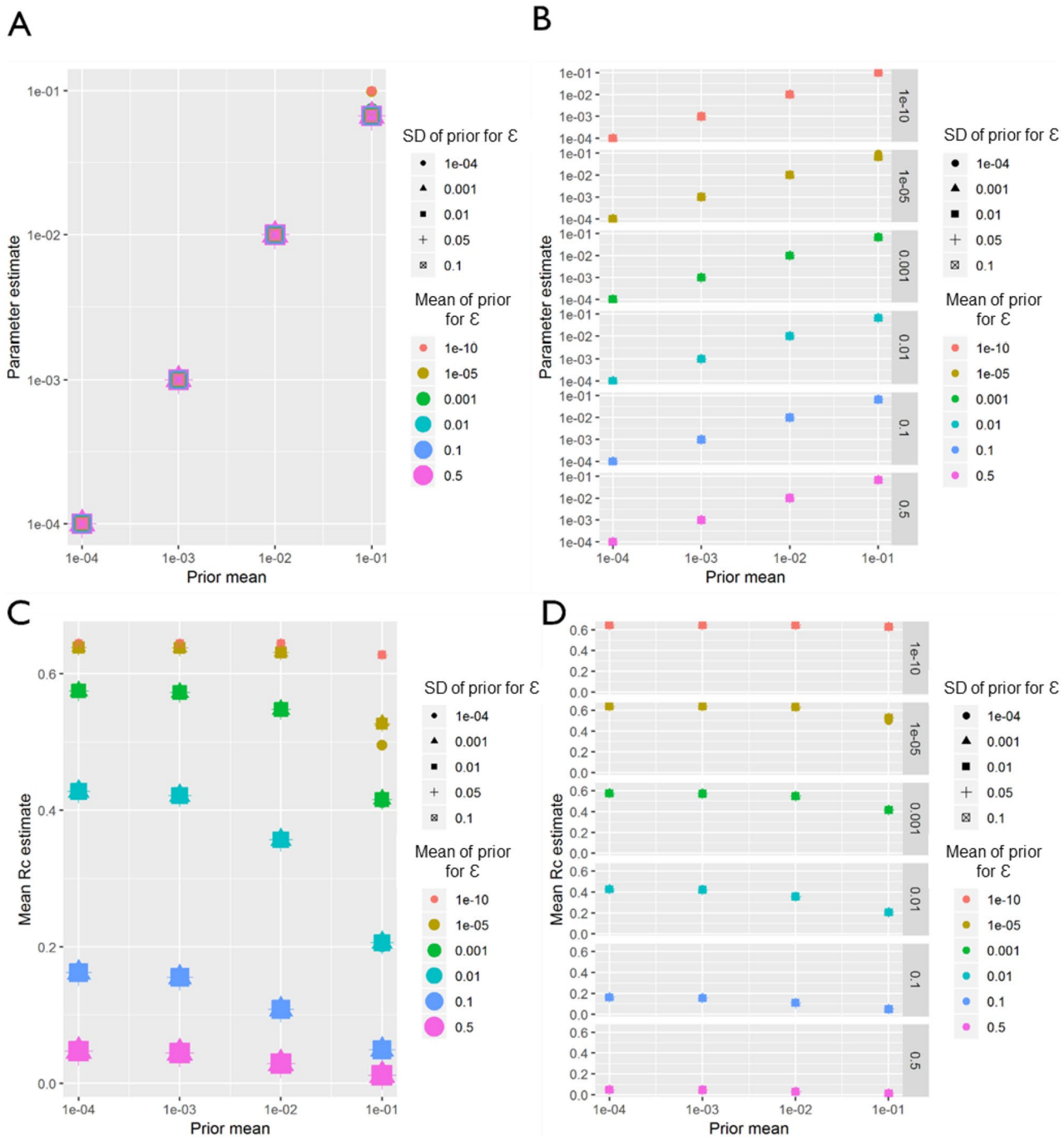

Figure 9. El Salvador sensitivity analysis. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε, which represents shapes represent different hazards of infection by an external, unobserved source. For A-D, the x-axis represents the prior mean used for β. (**A**) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. (**B**) shows the same results, stratified by the prior mean of ε for clarity. (**C**) Shows the impact of priors for β and ε on the mean Rc estimate, and again (**D**) shows the same result, stratified by the prior mean of ε.

Transmission model specifics. In order to incorporate pairwise distance metrics, we extended our previously published algorithm applied to Yunnan Province, China[11] by introducing a second function, f_2, which describes the relationship between space (or distance of any kind) and likelihood of transmission. An appropriate function such as a Gaussian kernel is defined and the parameter(s) shaping that distribution, β, are either fixed, or given a prior distribution and estimated from the data. Multiplied, together, this returns a single function:

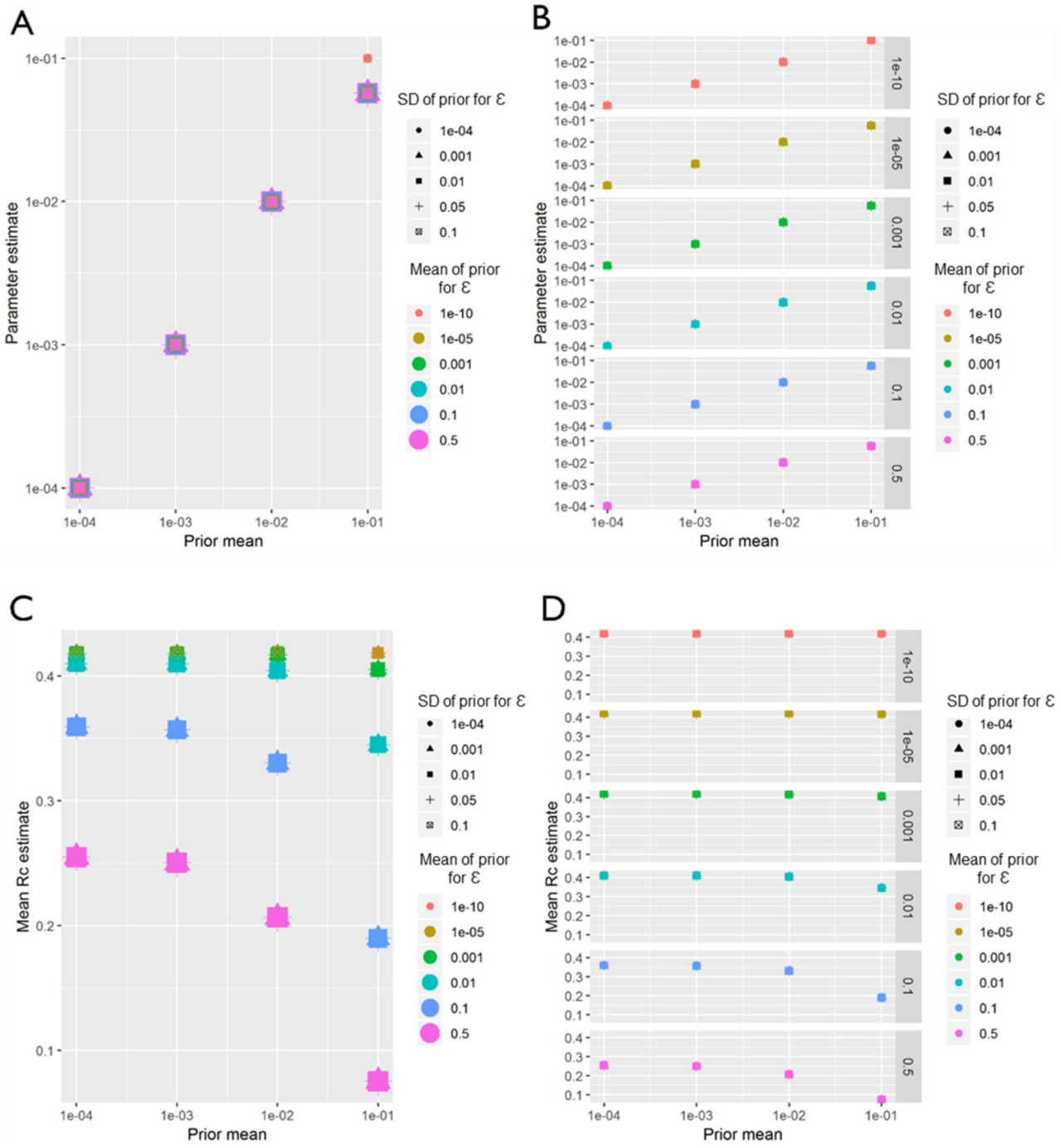

Figure 10. Eswatini sensitivity analysis. Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε, which represents shapes represent different hazards of infection by an external, unobserved source. For (**A**)–(**D**), the x-axis represents the prior mean used for β. (**A**) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. (**B**) shows the same results, stratified by the prior mean of ε for clarity. (**C**) Shows the impact of priors for β and ε on the mean Rc estimate, and again (**D**) shows the same result, stratified by the prior mean of ε.

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

Determined by times t, spatial locations x, transmission rates α, spatial parameter(s) β.

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

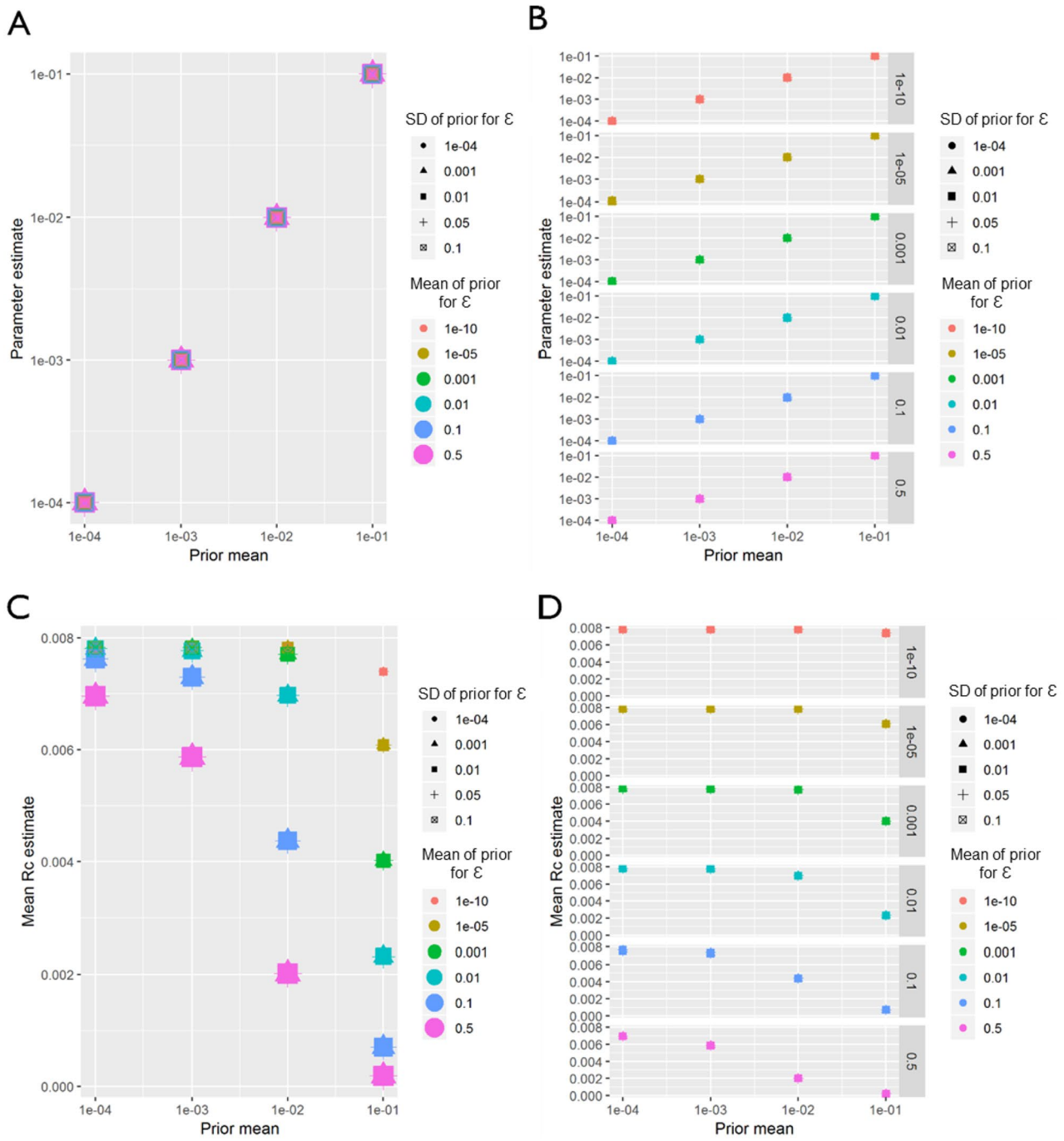

Figure 11. China (*P. falciparum*) Sensitivity Analysis. Sensitivity analysis showing the impact of varying the prior means for *P. falciparum* in China. Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε, which represents shapes represent different hazards of infection by an external, unobserved source. For (**A**)–(**D**), the x-axis represents the prior mean used for β. (**A**) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. (**B**) shows the same results, stratified by the prior mean of ε for clarity. (**C**) Shows the impact of priors for β and ε on the mean Rc estimate, and again (**D**) shows the same result, stratified by the prior mean of ε.

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

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

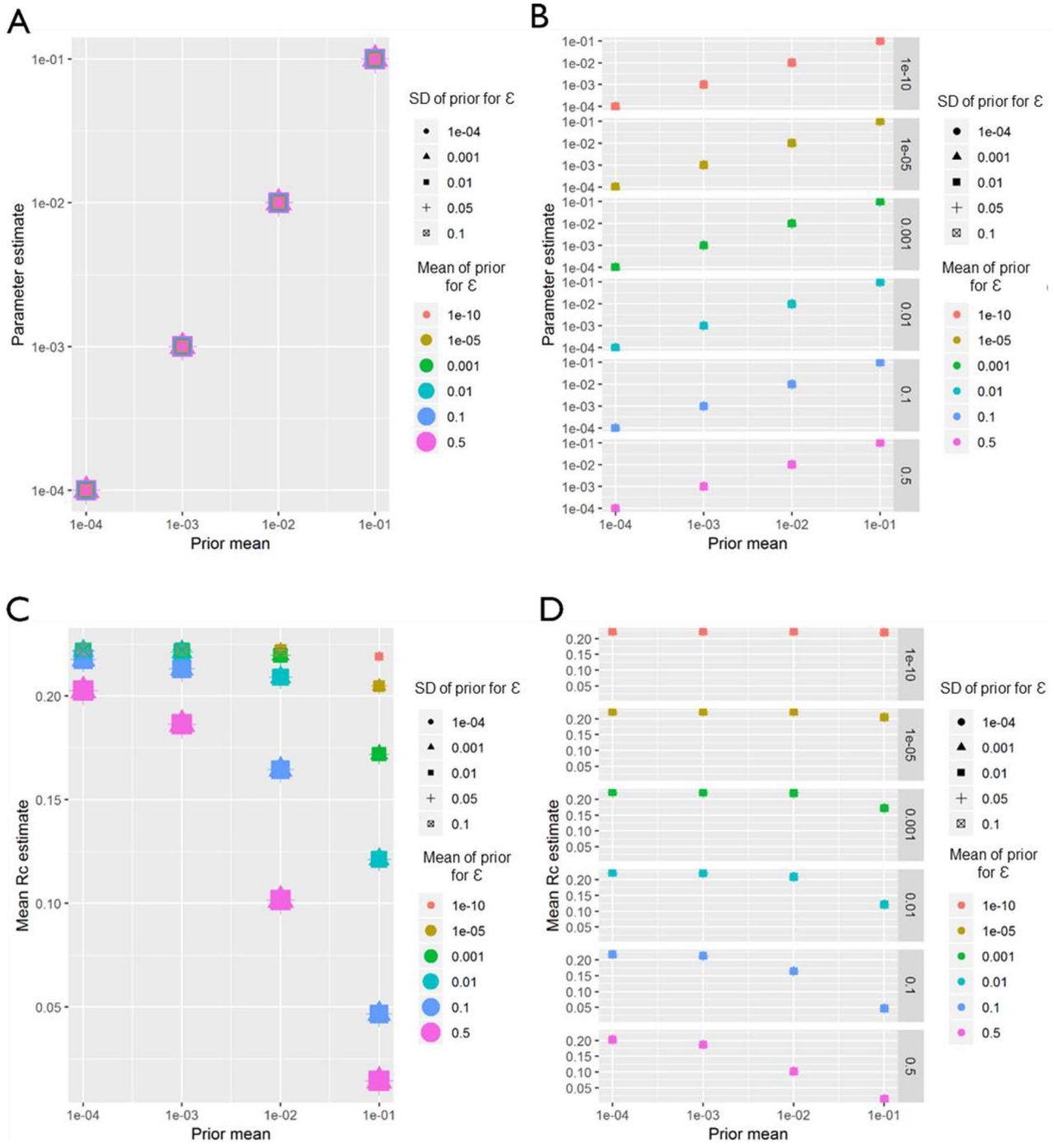

Figure 12. China (*P. vivax*) Sensitivity Analysis. Sensitivity analysis showing the impact of varying the prior means for *P. vivax* in China. Sensitivity analysis showing the impact of varying the prior means for Eswatini. Sensitivity analysis showing the impact of varying the prior mean for the distance kernel shaping parameter, β. The different colours and shapes represent different means and standard deviations respectively of the normally-distributed prior of epsilon, ε, swhich represents shapes represent different hazards of infection by an external, unobserved source. For (**A**)–(**D**), the x-axis represents the prior mean used for β. (**A**) the y-axis shows the maximum a posteriori parameter estimate for the parameter β. (**B**) shows the same results, stratified by the prior mean of ε for clarity. (**C**) Shows the impact of priors for β and ε on the mean Rc estimate, and again (**D**) shows the same result, stratified by the prior mean of ε.

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

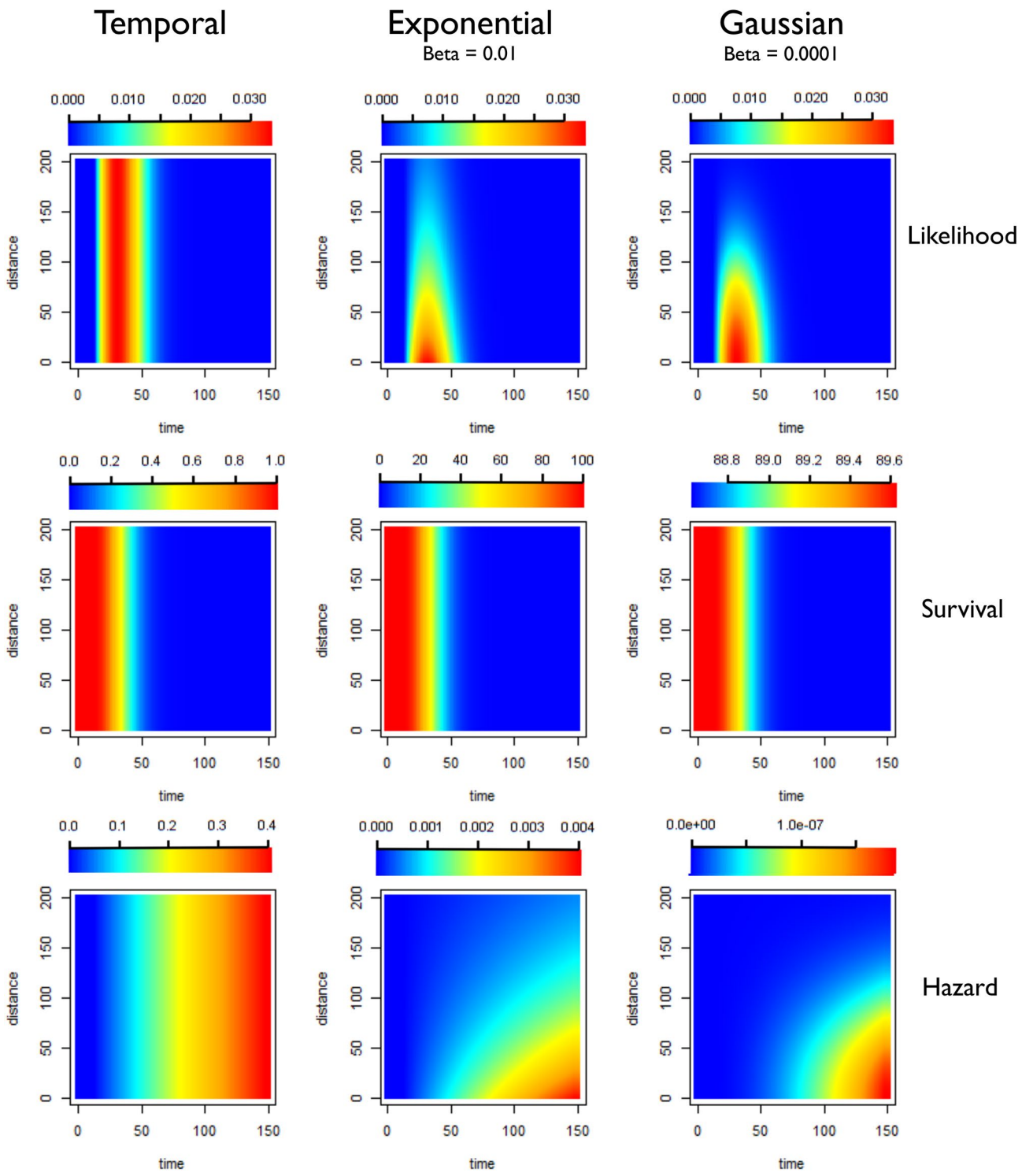

Figure 13. Illustration of likelihoods, hazards and survivals for less restrictive kernels (longer range human movement likely). Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example less restrictive values for beta, the shaping parameter for the distance kernels have been chosen, representing a context where there is more long-range movement of parasites.

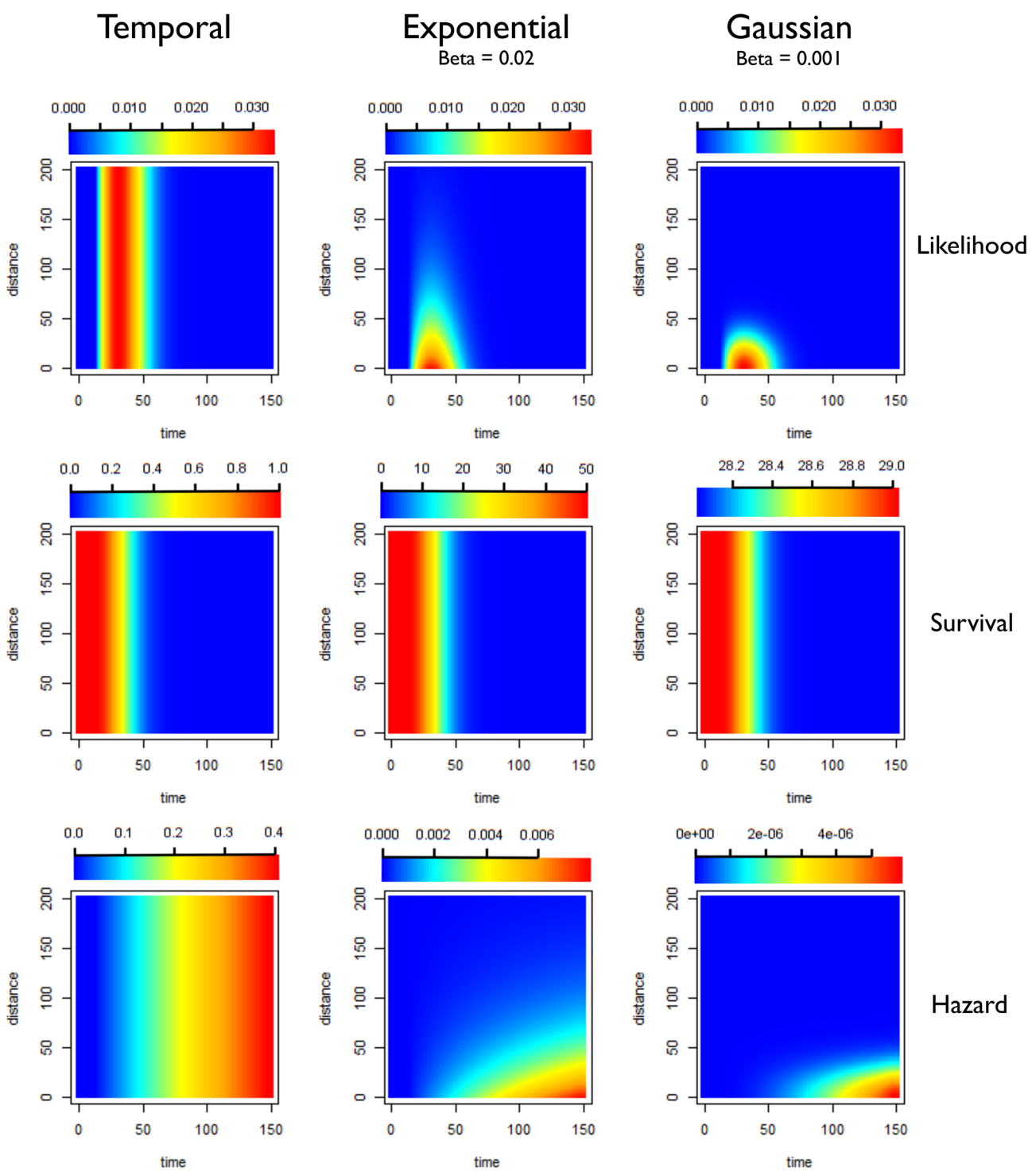

Figure 14. Illustration of likelihoods, hazards and survivals for moderately restrictive kernels (moderate human movement, most movement under 50 km). Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example values for beta, the shaping parameter for the distance kernels have been chosen to represent a context where there is more some movement of parasites, but where little movement is expected beyond 50–75 km. The likelihood for the Gaussian Kernel is more concentrated, which could represent shorter range movement e.g. commutes, whereas the Exponential has a longer tail so could represent a mixture of short and longer range parasite movement.

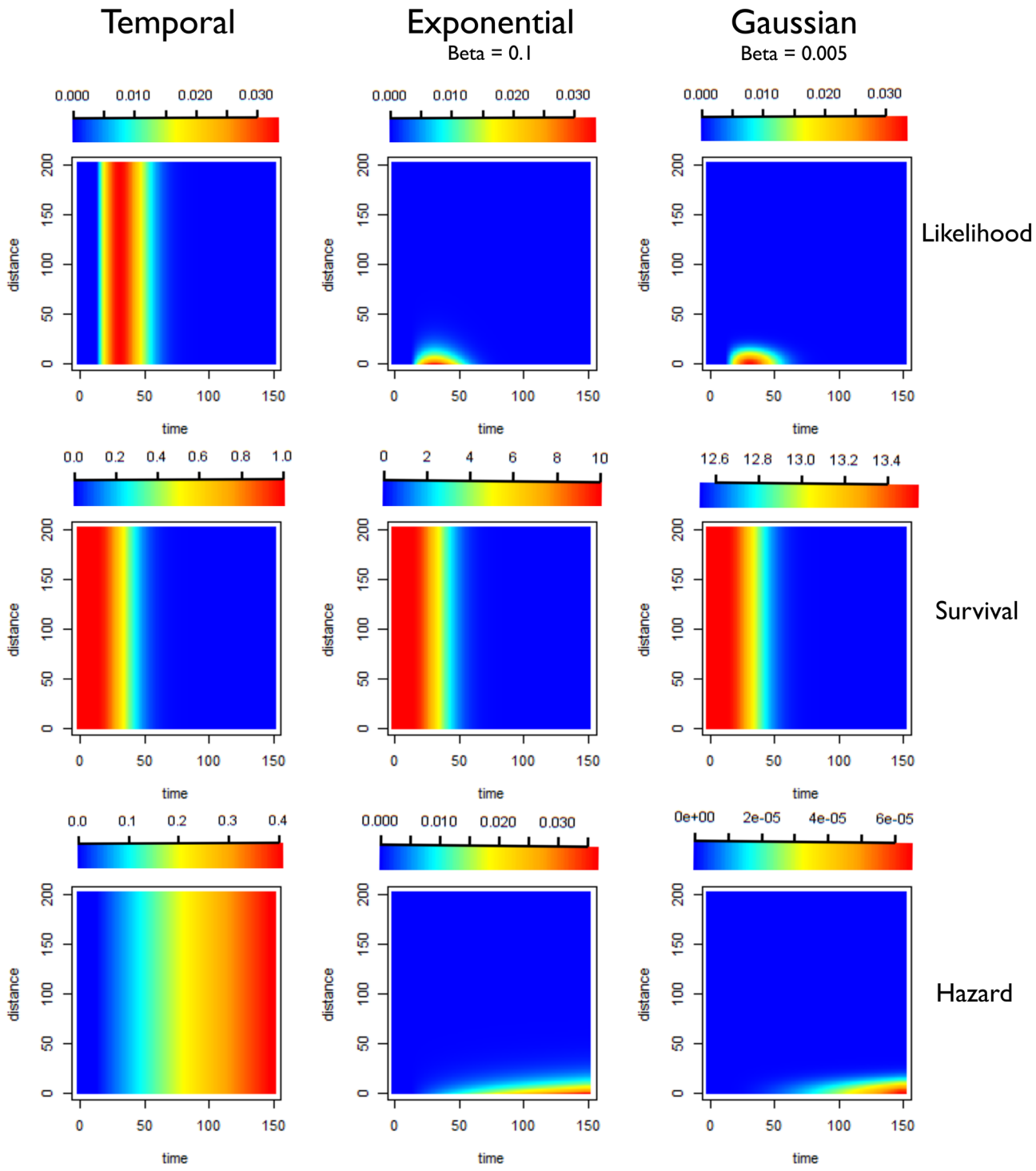

Figure 15. Illustration of likelihoods, hazards and survivals for highly restrictive kernels (Human movement unlikely, most movement under 10 km). Plots showing how the pairwise likelihoods, survivals and hazards vary with time and distance under different model structures. The first row of plots shows the pairwise likelihoods, the second row shows the pairwise survival and the third row shows the pairwise hazard values for different combinations of distance (in kilometres) and time between symptom onset (days). The first column shows the results for a time-only version of the algorithm. The second column shows results for an exponential kernel and the third column shows results for a Gaussian kernel. In this example more restrictive values for beta, the shaping parameter for the distance kernels have been chosen, representing a context where there is very little movement of parasites, with very little movement beyond 10–20 km expected.

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Confirmed	260	11,830	252	822	6631	87
Probable	0	176	0	0	693	311

Table 4. Cases by diagnosis type (probable and confirmed) and species across China.

	Mixed infection	*P. falciparum*	*P. malariae*	*P. ovale*	*P. vivax*	Untyped
Local	5	92	4	1	1711	95
Imported	255	11,914	248	821	5613	303

Table 5. Cases by imported/local status and species across China.

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

Table 6. Equations for f1, f2, hazard and survival for time-only, Gaussian and Exponential spatial kernels.

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

To illustrate this approach by applying to several malaria line-lists, we used a shifted Rayleigh distribution to model serial interval distributions, $f_1(t_i|t_j;\ \alpha_{i,j})$. For the second part of the likelihood which model the relationship between space and the likelihood of transmission $f_2(x_i|x_j;\ \beta)$, Gaussian and Exponential diffusion kernels were used (Table 6).

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

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

The survival term simplifies to:

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

And the hazard simplifies to:

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

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

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

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

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

Integrating over time returns

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

Integrating over all distances gives

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

Following Eq. (10), the hazard is equivalent to

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

Which simplifies to

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

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

Adding epsilon, ε, as a competing hazard and survival returns:

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

The objective function is then:

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

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

The algorithm was written in TensorFlow, implemented in R via the *rTensorflow* package[33]. A prior probability was defined for the parameter shaping the serial interval of malaria, informed by previous characterisations of the serial interval of malaria[34]. Because data about how likely the cases were to have moved long distances or the likelihood of a case has been infected by an unobserved source of infection were not available for the contexts explored here, several different parameterisations of the model were used to represent different scenarios (Table 1) and a detailed sensitivity analysis was carried out (Table 3). The versions of the model which are described in Table 1 and Figs. 13, 14 and 15 represent different patterns of human/parasite movement, ranging from a context where there may be small amounts of movement (almost all under 10 km) to moderate amounts of movement/travel (almost all under 50 km) to a less restrictive parameterisation, where near cases were more likely but far away cases were not completely excluded. We applied different versions of the algorithm, as well as temporal-only algorithm to these datasets to explore the impact of different assumptions about the impact of space on estimated R_c values and their variation over time and space. We also evaluated the performance of each approach by comparing differences in the second order AIC (ΔAICc), and the corresponding Akaike Weights.

Twelve scenarios (Table 1) were considered when defining parameters for each dataset. These scenarios consider three different levels of likelihood of transmission in relationship to Euclidian distance (due to the limited range of mosquito travel, this is considered in the context of human mobility), which was defined for both exponential and Gaussian kernels. These are illustrated in Figs. 13, 14 and 15. Then the values for epsilon were set at 0.001 and 0.1, representing different levels of missing cases likely. This can be interpreted as the chance of a case having an unobserved source of infection. For example, 0.1 would represent P(unobserved source of infection) = 0.1.

The timeseries of R_c and its spatial patterns were illustrated for each dataset and parameter combination and compared to the results of the time-only version of the algorithm. The results were also mapped to compare how spatial patterns in R_c were affected by assumptions about space and unobserved infections. The maps shown in Figs. 5, 6, 7 and 8 were created in Tableau version 2019.1 using base maps from OpenStreetMap and OpenStreetMap Foundation (www.openstreetmap.org).

In order to compare models quantifiably, the second order Akaike Information Criterion (AIC_c) was calculated using the equation $AIC_c = -2\log f(x) + 2K\left(\frac{n}{n-k-1}\right)$, where $f(x)$ is the model likelihood, K is the number of parameters estimated and n is the sample size of the data used to fit the parameters. The AIC[35] is used in model comparison, by creating a comparison of negative log likelihood that penalises increases in model parameters, to prevent overfitting. AICc is recommended for use with smaller datasets with larger numbers of parameters, and as the sample size n increases AICc converges to AIC[22]. The differences in AICc for each model, known as

ΔAICc, were calculated to compare models. Typically, a ΔAICc of greater than 10 is considered strong evidence that that model performs worse than the model it is being compared to.

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

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

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

Code availability

Code for algorithm to generate reproduction numbers, estimate unobserved cases and calculate metrics of model performance are available at https://github.com/IzzyRou/spatial_rcs.

Received: 11 November 2020; Accepted: 11 June 2021
Published online: 14 July 2021

I've now edited this however in the process accidentally created a duplicate reference (36) which I cannot seem to delete. Please delete reference 36.References

1. Lourenço, C. *et al.* Strengthening surveillance systems for malaria elimination: A global landscaping of system performance, 2015–2017. *Malar. J.* **18**, 1–11 (2019).
2. Sturrock, H. J. W. *et al.* Mapping malaria risk in low transmission settings: Challenges and opportunities. *Trends Parasitol.* **32**, 635–645 (2016).
3. Wesolowski, A. *et al.* Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med.* **16**, 190 (2018).
4. Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542–548 (2001).
5. Ghani, A. *et al.* The early transmission dynamics of H1N1pdm influenza in the United Kingdom. *PLoS Curr.* **1**, 66 (2010).
6. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
7. Haydon, D. T. *et al.* The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. R. Soc. B Biol. Sci.* **270**, 121–127 (2003).
8. Jombart, T. *et al.* Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLOS Comput. Biol.* **10**, e1003457 (2014).
9. Reiner, R. C. Mapping residual transmission for malaria elimination. *eLife* **4**, 09520 (2015).
10. Routledge, I. *et al.* Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting. *Nat. Commun.* **9**, 2476 (2018).
11. Routledge, I. *et al.* Tracking progress towards malaria elimination in China: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. *PLOS Comput. Biol.* **16**, 1007707 (2020).
12. Prothero, R. M. Disease and mobility: A neglected factor in epidemiology. *Int. J. Epidemiol.* **6**, 259–267 (1977).
13. Pindolia, D. K. *et al.* Human movement data for malaria control and elimination strategic planning. *Malar. J.* **11**, 205 (2012).
14. Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E. & Snow, R. W. Mobile phones and malaria: Modeling human and parasite travel. *Travel Med. Infect. Dis.* **11**, 15–22 (2013).
15. Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012).
16. Cotter, C. *et al.* The changing epidemiology of malaria elimination: New strategies for new challenges. *The Lancet* **382**, 900–911 (2013).
17. Wang, L., Ermon, S. & Hopcroft, J. E. Feature-Enhanced Probabilistic Models for Diffusion Network Inference. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 7524 LNAI 499–514 (2012).
18. Gomez-Rodriguez, M., Leskovec, J. & Krause, A. Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data TKDD* **5**, 1–37 (2012).
19. Rodriguez, M. G., Balduzzi, D. & Schölkopf, B. Uncovering the temporal dynamics of diffusion networks. ArXiv Prepr. arXiv: 11050697 (2011).
20. Unwin, H. J. T. *et al.* Using Hawkes Processes to model imported and local malaria cases in near-elimination settings. *medRxiv* https://doi.org/10.1101/2020.07.17.20156174 (2020).
21. Gomez-Rodriguez, M., Leskovec, J., Balduzzi, D. & Schölkopf, B. Uncovering the structure and temporal dynamics of information propagation. *Netw. Sci.* **2**, 26–65 (2014).
22. Hurvich, C. M. & Tsai, C. L. Regression and time series model selection in small samples. *Biometrika* **76**, 297–307 (1989).
23. Bateman, A. J. Is gene dispersion normal?. *Heredity* **4**, 353–363 (1950).
24. Huber, J. H. *et al.* Inferring person-to-person networks of pathogen transmission: Is routine surveillance data up to the task? *medRxiv* https://doi.org/10.1101/2020.07.17.20156174 (2020).
25. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
26. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).

27. Kraemer, M. U. G. *et al.* Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 5151 (2019).
28. Prosper, O., Ruktanonchai, N. & Martcheva, M. Assessing the role of spatial heterogeneity and human movement in malaria dynamics and control. *J. Theor. Biol.* **303**, 1–14 (2012).
29. Marshall, J. M. *et al.* Mathematical models of human mobility of relevance to malaria transmission in Africa. *Sci. Rep.* **8**, 7713 (2018).
30. Bousema, T. & Drakeley, C. Epidemiology and infectivity of *Plasmodium falciparum* and *Plasmodium vivax* gametocytes in relation to malaria control and elimination. *Clin. Microbiol. Rev.* **24**, 377–410 (2011).
31. Reiner, R. C. *et al.* Mapping residual transmission for malaria elimination. *eLife* **4**, e09520 (2015).
32. Lai, S. *et al.* Changing epidemiology and challenges of malaria in China towards elimination. *Malar. J.* **18**, 107 (2019).
33. Abadi, M, *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint* arXiv:1603.04467 (2016).
34. Huber, J. H., Johnston, G. L., Greenhouse, B., Smith, D. L. & Perkins, T. A. Quantitative, model-based estimates of variability in the generation and serial intervals of *Plasmodium falciparum* malaria. *Malar. J.* **15**, 490 (2016).
35. Akaike, H. A new look at the statistical model identification. *Autom. Control IEEE Trans.* **19**, 716–723 (1974).

Author contributions

IR devised the project, carried out analysis, generated figures, curated data, wrote paper. HJTU advised on simulation work and edited paper draft. SB supervised the project and edited paper draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-93238-0.

Correspondence and requests for materials should be addressed to I.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.