
There may be differences between this version and the published version. You are advised to consult the publisher’s version if you wish to cite from it.

[http://eprints.gla.ac.uk/188671/](http://eprints.gla.ac.uk/188671/)

Deposited on: 19 June 2019
Achieving explanatory depth and spatial breadth in infectious disease modelling: integrating active and passive case surveillance.

Abstract:
Ideally, the data used for robust spatial prediction of disease distribution should be both high-resolution and spatially expansive. However, such in-depth and geographically broad data are rarely available in practice. Instead, researchers usually acquire either detailed epidemiological data with high resolution at a small number of active sampling sites, or more broad-ranging but less precise data from passive case surveillance.

We propose a novel inferential framework, capable of simultaneously drawing insights from both passive and active data types. We developed a Bayesian latent point process approach, combining active data collection in a limited set of points, where in-depth covariates are measured, with passive case detection, where error-prone, large-scale disease data are accompanied only by coarse or remotely-sensed covariate layers.

Using the example of malaria, we tested our method’s efficiency under several hypothetical scenarios of reported incidence in different combinations of imperfect detection and spatial complexity of the environmental variables.

We provide a simple solution to a widespread problem in spatial epidemiology, combining latent process modelling and spatially autoregressive modelling. By using active sampling and passive case detection in a complementary way, we achieved the best-of-both-worlds, in effect, a formal calibration of spatially extensive, error-prone data by localised, high-quality data.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.
Achieving explanatory depth and spatial breadth in infectious disease modelling: integrating active and passive case surveillance.

Luca Nelli1*, Heather M. Ferguson1, Jason Matthiopoulos1

1Universe of Glasgow, Institute of Biodiversity Animal Health and Comparative Medicine, Glasgow G12 8QQ, UK

*corresponding author: luca.nelli@glasgow.ac.uk

Abstract

Ideally, the data used for robust spatial prediction of disease distribution should be both high-resolution and spatially expansive. However, such in-depth and geographically broad data are rarely available in practice. Instead, researchers usually acquire either detailed epidemiological data with high resolution at a small number of active sampling sites, or more broad-ranging but less precise data from passive case surveillance.

We propose a novel inferential framework, capable of simultaneously drawing insights from both passive and active data types. We developed a Bayesian latent point process approach, combining active data collection in a limited set of points, where in-depth covariates are measured, with passive case detection, where error-prone, large-scale disease data are accompanied only by coarse or remotely-sensed covariate layers.

Using the example of malaria, we tested our method’s efficiency under several hypothetical scenarios of reported incidence in different combinations of imperfect detection and spatial complexity of the environmental variables.

We provide a simple solution to a widespread problem in spatial epidemiology, combining latent process modelling and spatially autoregressive modelling. By using active sampling and passive case detection in a complementary way, we achieved the best-of-both-worlds, in effect, a formal calibration of spatially extensive, error-prone data by localised, high-quality data.
Keywords: Bayesian modelling, disease mapping, imperfect detection, latent point process, N- mixture models, spatial epidemiology.
INTRODUCTION

Predictive maps of disease risk, typically obtained by modelling the spatial heterogeneity in disease incidence as a function of underlying covariates, can be crucial for targeting effective control and surveillance. However, reliable prediction at the landscape scale is often hindered by lack of appropriate, high resolution spatial data. Traditionally, incidence data and potential explanatory covariates are collected either systematically – using active sampling by researchers – or opportunistically – from clinical records reported at health facilities. Each of these sampling strategies has its own limitations. For example, by collecting detailed data for both disease incidence and related covariates, data from active sampling allows models to achieve high explanatory power but not to make large-scale extrapolation and predictions in areas where fine scale covariates are not directly measurable. On the other hand, passive sampling yields data from a large number of geographically dispersed cases which are more amenable for large scale predictions, but these data often suffer from severe reporting biases and can be paired with only coarse environmental covariates that have limited explanatory power. As the drawbacks of one strategy are clearly the strengths of the other, modelling frameworks that consider these two types of data simultaneously and complementarily would strengthen our biological insights and predictive power.

Active sampling is typically conducted by research teams that focus on a small number of predetermined locations, with collection of detailed environmental or epidemiological variables including clinical samples, entomological indicators (for vector-borne disease), human demographic and socio-economic factors or fine-scale environmental conditions. Such data can provide high power for explaining variation in risk across focal sites, but lack predictive breadth across space because many of the crucial covariates are not available for un-sampled locations.
Clinical records from passive case detection offer the potential of expansive descriptions of spatial incidence patterns. However, since these incidence data often arise from self-reporting at health centres, they can be biased by their opportunistic nature. Reporting bias is well acknowledged for numerous infectious disease systems and can be expressed as a combined function of distance from health facilities, the likelihood of asymptomatic cases and sociodemographic factors or more complex measures of travel time. Despite this limitation, health centre surveys remain the primary source of information for disease monitoring. Another drawback of spatial models of incidence data gathered from passive case detection, relates to the availability of environmental predictor data. If the locality of the patient is recorded, incidence data can be spatially plotted but researchers and public health workers are unlikely to be able to directly measure some detailed explanatory variables at those localities. Therefore, when modelling the incidence data, only large-scale but coarse layers are customarily considered. While these bring more geographically expansive information than the highly localised survey data, they generally consist of remotely sensed covariates and summary records such as bioclimatic, geomorphological, vegetation indexes, human population density or road networks that typically contribute limited explanatory power.

Some studies make use of data from both active and passive case detections together, but focus on independent analysis and comparison of results from these separate data sources rather than integrating them. Analysing these two data sources jointly can be viewed as challenging because their limitations imply a trade-off between explanatory depth and predictive breadth. However, there is clearly an opportunity to achieve complementarity by analysing them on an integrated inferential framework. Here, for the first time, we develop a spatial statistical model combining these two sources of incidence data to harness the maximum amount of information for explanatory and predictive objectives.

Our framework takes a novel approach to both the response and the explanatory variables. The dual nature of the incidence data requires specification of a statistical model that considers two different
aspects of likelihood, one for the localised but precise survey data, and another for the spatially
extensive but imperfect clinical reporting data. We build this part of the approach on two
cornerstones of the statistical literature: the point process model and the methodology of point
transects. Point processes model events (e.g. infection cases) that occur continuously in space
according to an unknown intensity (a spatial surface to be estimated as a function of covariates). We
observe these events as arising from two different point transects, each having its own spatially
heterogeneous observation model. The first type of observation point is the active sampling
location, where cases are detected near-perfectly, but only for that particular set of geographical
coordinates. The second type of observation point is a clinic, where cases of the disease are reported
from a broad geographical region but with probabilities of detection that decay with distance from
the clinic. Regarding the explanatory variables, some environmental variables are easily collectable
for both passive case detection and active sampling points, but more important and powerful
variables may be available only at the latter. By importing ideas from latent process modelling
we use the spatially extensive clinical data together with the data-rich survey data to reconstruct
latent covariates that may be hidden from direct or remote observation.

To validate the ability of our model to retrieve correct parameter values we require these scenarios
to be accompanied by known intensity surfaces for both incidence and latent explanatory variables.
These requirements cannot be satisfied by real data sets, so here we have acquired our scenarios via
realistic simulation, motivating our examples from a real system of a vector-borne disease. To
illustrate the generality of our approach, we have hypothesized multiple contrasting scenarios of
reporting bias and spatial distribution of the latent process underlying disease incidence.

We chose malaria in West Africa as an ideal example of an important environmentally-dependent
infectious disease, for which human exposure and infection risk is highly spatially
heterogeneous and dependent on crucial environmental variables that influence interactions
between people, mosquitoes and parasites. Control measures such as long-lasting insecticide
treated nets (LLINs) have been crucial for impeding contact between mosquitoes and people, and
have led to substantial declines in malaria prevalence across Africa in the last decade 50, 51. However,
the success of such an approach may be undermined by development of insecticide resistance in
mosquitoes, particularly in West Africa where rates are among the highest in the world 53-55. Copious
and widespread data on reported cases are often available from clinics (see for example
www.malariasurveys.org or www.dhsprogram.com), but detailed information on mosquito vector
ecology and insecticide resistance is only available for a limited number of sites (e.g. 54, 56-58). These
challenges exist for many other vector-borne diseases whose transmission is dependent on an
ecological reservoir and rely on insecticide use for control, such as for example dengue, Zika and
chikunguya viruses 59, Lyme and other tick-borne disease 60, schistosomiasis 61, Rift Valley Fever 62,
human African trypanosomias 63 or West Nile Virus 64.

METHODOLOGY

Modelling approaches

For a given area of interest subdivided into a regular grid, we consider as our sampling unit the grid
cell $i \in \{1, \ldots, K\}$. We first assume an underlying stochastic process $f$ that generates numbers of cases
$N_i$ according to an underlying, spatially heterogeneous rate $\lambda_i$. We also assume an observation
process $g$ that allows a subset of the $N_i$ cases to be reported at different sampling stations. We
distinguish between two types of sampling stations: $S$ is the number of active sampling points (about
which we are assuming a perfect and exclusive detection but at a small distance, i.e. within the cell
that contains them). We denote by $J$ the number of clinics (about which we are assuming an
imperfect but long-ranging detection). The observation process $g$ is therefore generating the vector
of incidence data reported in each $i^{th}$ cell at different stations $I_i =$

$$\{I_{1,i}, \ldots, I_{S,i}, I_{(S+1),i}, \ldots, I_{(S+J),i}, U_i\}$$

given the vector of probabilities $P_i =$

$$\{P_{1,i}, \ldots, P_{S,i}, P_{(S+1),i}, \ldots, P_{(S+J),i}, Q_i\}$$

$U_i$ represents the number of completely unreported cases in
each $i^{th}$ cell (which is a missing value in the data), given the probability $Q_i$ of not reporting.
The general likelihood function of our models can be expressed as follows:

\[
L = \prod_{i=1}^{K} f(N_i | \lambda_i) \, g(I_i | P_i, N_i) \tag{1}
\]

We built our approach incrementally, developing three distinct modelling approaches with an increasing level of complexity to allow comparison between the routes that might have traditionally been followed to analyse data arising from active sampling (model 1) and passive case detection (model 2) with our new proposed route (model 3), which reconstructs the latent processes and estimates the emergent patterns of disease incidence with increased precision and accuracy.

**Model 1 – active sampling data only**

Here, we consider data that would be collected from active sampling at just a limited number \( S \) of active survey sites. To analyse the relationship between disease incidence and detailed measures of covariates at a set of predetermined survey points, model 1 takes the form of a Poisson Generalised Linear Model without any spatially explicit component.

Although this is a straightforward model to fit using likelihood-based libraries in all statistical platforms, we fitted it using Bayesian methods for consistency in the comparison with models 2 and 3 that follow. The response variable is the number of observed diseases cases \( N_i \) at the location of the \( i^{th} \) survey. We assume here (for simplicity, but with no loss of generality) that all the cases at the survey location are recorded (hence, a local detection probability of 1 for each case), although we acknowledge that with conventional diagnostic tests some percentage of cases can be missed.

If data are available on diagnostic sensitivity and specificity, our method can be readily extended by incorporating false negatives or positives.

The model takes the form

\[
N_i \sim \text{Poisson}(\lambda) \tag{2}
\]
where the rate \( \lambda_i \) of disease incidence is

\[
\ln(\lambda_i) = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ik}
\]  \[3\]

The linear predictor on the right-hand-side of this expression comprises a set of \( n \) coefficients \( \beta \) and \( n \) explanatory variables \( X \) measured at the \( i^{th} \) survey location.

Equations [2] and [3] can be generalised to take better account of specific features of the data. For example, it may be relevant to use overdispersed forms of the likelihood (relaxing the Poisson assumption) or more complicated functional forms of the linear predictor, involving polynomials, interactions or splines.

**Model 2 – passive case detection only**

Here, we considered only data coming from passive case detection. This model maintained the basic structure of model 1, i.e. it is a Bayesian Poisson regression, with reported disease cases at human dwellings or communities surrounding the health centres as the response variable and the set of environmental variables as predictors. Under our scenarios, we assumed that one of the key predictor variables (insecticide resistance \( IR, \text{ see model validation} \)) could only be measured experimentally in active sampling sites, therefore we couldn’t include it in eq. [3].

We introduced the estimation of bias in reporting disease cases given by the distance from the health centres, borrowing concepts from distance sampling theory \(^{47}\), a group of methods, widely used to estimate the absolute abundance or spatial density of animal or plant populations. The key underlying concept is the estimation of a detection function \( P(d) \), which represents the decay in the probability of detecting an object with increasing distance \( d \) from the observer. Given the detection function and encounter rate, the absolute density of a population can be modelled at a given point, assuming perfect detection at the location of the observer \( P(0) = 1 \). In our application, this has the interpretation that if a case arises in the immediate vicinity of the clinic \( d \approx 0 \), then it is
certain to be reported. A plausible, but flexible decay function is fitted to paired data of detections and distances. For example, detection of a malaria case from the $i^{th}$ location at the $j^{th}$ clinic, and can be modelled as a half-normal function of distance from the health centre $d_{ij}$, by the following $^{47}$:

$$p(d_{ij}) = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)$$

[4] where $\sigma$ is the shape parameter of the half-normal function (regulating how quickly the detection probability drops with distance). The distance $d$ can be Euclidean, or a more complicated function of accessibility (e.g. affected by proximity between points along a given road network).

Any given case may be reported to any one of the available clinics, but clinics nearby are more likely to receive the report. The probability of any one case being reported to any one clinic (accounting for other clinics) can be modelled in terms of the distances of all the clinics from the point of occurrence of the case, as follows

$$P_{ij} = \frac{p(d_{ij})}{\sum_{j=1}^{J} p(d_{ij}) + Q_i}$$

[5] The denominator here represents all possible outcomes, i.e. the probabilities that the case is reported to any one of $J$ centres, and the probability $P_{ij}$ that the case goes completely unreported:

$$Q_i = \prod_{j=1}^{J} (1 - p(d_{ij}))$$

[6] Note that $P_{ij}$ is the standardised form of $p(d_{ij})$. In fact, $p(d_{ij})$ is the probability of a case being reported at a given clinic (considered in isolation), purely as a function of distance, whereas $P_{ij}$ is the probability of reporting at a clinic, accounting for the effects of other clinics that are “competing” for the same reports and including $Q_i$, that is the probability of a case not being reported at all.
The likelihood of a data set comprising clinic reports may then be written as a multinomial process. In particular, for a given number of actual cases \( N_i \) (see eq. [2]), the likelihood of reported disease cases \( I_i \) in the \( i^{th} \) cell for the \( J \) clinics in the dataset is determined by the detection probabilities \( P_i \) that are function of distances between the \( i^{th} \) location and the clinics, by

\[
I_i \sim \text{Multinomial}(N_i, P_i)
\]  

where \( P_i = \{P_{1,i}, \ldots, P_{J,i}, Q_i\} \).

Fitting model 2 to the data yielded estimates of the shape parameter of the detection function (eq. [4]) and parameters of eq. [3]. Although it had no spatially explicit component, we used model 2 to generate a reconstruction of the patterns of incidence across space, based on the coarse-level environmental covariates. Hence this model did not benefit from the fine-resolution covariates that could only be measured by detailed experimental methods at survey points.

Model 3 – active and passive data combined

The process and observation model for this joint approach to data took the form of eqs. [2] and [7] respectively. However, just like in model 1, eq. [3] used the full set of predictors, including the partly-latent variable (i.e. insecticide resistance, available only for active sampling points but not for regions of passive case detection data collection and the rest of space). Our model for the latent variable \( IR \) postulated a spatial autocorrelation structure \( ^{67} \), implying that even though we may not know the values of the latent variable at two points in space, we can express a relationship about their expected degree of similarity. Any pair of \( K \) cells in our grid, say \( i \in \{1, \ldots, K\} \) and \( k \in \{1, \ldots, K\} \), were assumed to have a covariance, specified as a decreasing function of their distance

\[
cov_{i,k} = \exp(-\rho d_{i,k})
\]  

where \( \rho \) is the spatial autocorrelation parameter.
With $\rho \geq 0$. Again, this is one of many possible structures and our overall approach is not constrained to this functional form. The distribution of the latent variable $\mathbf{IR} = \{IR_1, \ldots, IR_K\}$ in all the $K$ cells, was therefore modelled as a Gaussian field from an $m$-dimensional multivariate normal distribution, where each of the dimensions represented the probability density of a cell in space.

$$IR_i \sim MVN(\mu, \Sigma) \quad [9]$$

Here, the mean vector $\mu$ has length $K$ (the total number of cells in geographical space), and $\Sigma$ is a $K \times K$ spatial covariance matrix with values of 1 on the diagonal and values $\text{cov}_{ik}$ for the $i$ row and $k$ column from eq. [8].

Model 3, hence, is fitted exactly as model 2 according to eq. [7], but the linear predictor function (eq. [3]), included all the covariates, unlike model 2, which was missing the covariate of IR. In particular, IR observations were used where available (at active sampling points), assuming that they were realisations from eq. [9].

Model validation

We used simulated data on malaria incidence and insecticide resistance within the primary mosquito vectors to validate our models. Our specific validation aims were to 1) evaluate the match between the posterior distribution of the coefficients and the simulation process that generated the data; 2) estimate bias in reporting the clinical data as a function of distance between the location of a clinic and the village where the patient resides; 3) recreate the missing covariate of insecticide resistance $\mathbf{IR}$ and to reconstruct the true incidence $\hat{N}$.

Our simulation borrowed its setting from a study currently ongoing in Southwest Burkina Faso (MiRA – Malaria in Insecticide Resistant Africa, Wellcome Trust 200222/Z/15/Z). The study covers an area of approx. 6000 km$^2$ in the health district of Banfora in south-western Burkina Faso, comprising primarily West Sudanian savannah which experiences a rainy season from May to October with little
228 rain in other months. Malaria transmission is stable throughout the year but peaks from May to
229 November. The major vectors are *Anopheles gambiae* and *An. funestus*. Like many other areas of
230 Africa, the primary malaria control strategy is long lasting insecticidal nets (LLINs) that are
231 distributed at high coverage across the country (Burkina Faso National Malaria Control Program,
232 unpublished data). In contrast to some areas of Africa, recent LLIN distribution campaigns have had
233 little impact on malaria prevalence and it is hypothesized that this may be due to high levels of
234 insecticide resistance in local vector populations, which are amongst the highest on record.
235 Resistance to pyrethroid insecticides is widespread. Mortality after exposure (defined by the World
236 Health Organization (WHO) as the response to the stipulated discriminating dose of permethrin)
237 ranges from 5-20% \(^{20}\). For the purposes of data simulation we assume that active sampling of malaria
238 infections and insecticide resistance levels is carried out in 12 villages, and that passive case data is
239 available from patients reporting to from 8 health centres distributed throughout the study area.
240 This number and distribution of passive and active sampling site was selected to represent the
241 distribution of health facilities and likely maximum amount of active survey data available.
242 For the simulation, we considered a square grid with a 1km\(^2\) resolution covering the study area. We
243 generated a dataset with reported incidence in each cell of the grid under a binomial *N*-mixture
244 model \(^{70, 71}\) by combining two different processes: a state model, i.e. the biological process that
245 generates malaria infection cases, and an observation model, i.e. the process that affects the
246 probability that infection cases are reported to a health centre.
247 To simulate the biological process, we considered the average altitude in the cell, average yearly
248 temperature (TEMP), annual rainfall (RAIN), human density (HUM), normalised difference vegetation
249 index (NDVI) and insecticide resistance (IR) in mosquitoes as potential predictors \(^{9, 38, 72-77}\).
250 Temperature and rainfall were derived from the WorldClim database (www.worldclim.org). NDVI
251 values were obtained using the package *MODIS*\(\text{Stsp}\) for R \(^{78}\). To create the layer of human density, we
252 used a kernel density estimation \(^{79}\) using GPS points of the villages (307) in the study area and the
population census in each village (1755 ± 1804 mean ± dev. std., Institut national de la statistique et
de la démographie, unpublished data) as weight field. Kernel bandwidth was chosen so as to
minimize the least-squares cross validation score ($h_{lscv}$)

Insecticide resistance reporting has improved over time, and global maps of insecticide resistance at
crude resolutions are now becoming available. However, little is known about its spatial
distribution at local scale. Therefore, to explore out model’s ability to retrieve latent variables of
differing spatial complexity, insecticide resistance was simulated by hypothesizing 3 different
scenarios of increasing spatial autocorrelation, with parameter $\rho$ of eq. [8] set respectively to $\rho_1$
$= 3.0$, $\rho_2 = 0.7$ and $\rho_3 = 0.3$ (Fig. 1, IR1, IR2, IR3).

The number of malaria cases, or true incidence, in each cell ($N_i$) was assumed to have a positive
relationship with temperature, rainfall, human density, NDVI and insecticide
resistance, and was simulated from eq. [2] using the linear predictor

$$\log (\lambda_i) = \beta_0 + \beta_{HUM} HUM_i + \beta_{NDVI} NDVI_i + \beta_{RAIN} RAIN_i + \beta_{TEMP} TEMP_i + \beta_{IR} IR_i$$

We set the equation’s coefficients to the values $\beta_0 = 2.90$, $\beta_{HUM} = 0.50$, $\beta_{NDVI} = 0.30$, $\beta_{RAIN} = 0.20,$ $\beta_{TEMP} = 0.25, \beta_{IR} = 0.50$. Having 3 distinct scenarios of insecticide resistance $IR1$, $IR2$ and $IR3$ we
obtained 3 scenarios of malaria infection cases $N1_i$, $N2_i$ and $N3_i$.

For the observation process, we accounted for simulated bias in reporting cases in each cell of the
grid, by considering a probability of reporting as a function of the distance between a given cell and
each health centre. We set the detection probabilities in each cell $P(i,j)$ in accordance with eq. [4]
with $p(d_{ij})$ being the Euclidean distance between the centroid of the $i^{th}$ cell of the grid and each $j^{th}$
health centre. We employed 3 different shapes of the detection function, using different values of
the shape parameter $\sigma_A = 10$, $\sigma_B = 15$, $\sigma_C = 20$ (Fig. 1, $P_A$, $P_B$ and $P_C$). Probability of reporting at
active sampling stations was deliberately set at 1, to ensure that all the infection cases occurring at
the sampling stations were recorded.
By combining the three scenarios of disease incidence given by the biological process with the three scenarios of detection function, we generated nine different scenarios of reported incidence for each cell ($I_i$), under a multinomial process given by [7]. For each combination scenario the response data comprised the number of reported cases per cell (Fig. 1, $I_{1A}$ to $I_{3C}$).

Preliminary manipulation of environmental layers was done using the software QGIS, the simulations were conducted in the statistical environment R, and Bayesian model fitting to the simulated data was carried out using the program JAGS, interfaced with R via the package rjags.

We analysed the simulated incidence data, using each of the three models described above. We used Markov Chain Monte Carlo (MCMC) algorithms (code provided in Appendix S1) to fit each of the models to the combination of environmental and incidence data. Relatively non-informative priors where chosen for all process and observation parameters and for the cells of the map relating to the latent variable. To make this a conservative test of the methodology, we employed priors wide variances. For the coefficients of the environmental covariates we chose diffuse normal priors centred at zero, corresponding to a null hypothesis of no-effect for each covariate. For the distance decay parameter $\sigma$ of the detection function, we adopted a uniform prior with limits 0-1000. For parameter $\rho$ of the covariance matrix describing spatial autocorrelation in the latent covariate, we used a gamma prior (shape = 0.1, rate = 0.1). To achieve convergence, model 1 and 2 were run for $3 \times 10^4$, whereas model 3 was run for $1.2 \times 10^6$ iterations.

Means of posterior distributions with corresponding credible intervals were obtained for each model coefficient $\hat{\beta}_k$ as well as the shape parameters of the detection function $\hat{\sigma}$, (only relevant for models 2 and 3). For each model and each simulated scenario, we generated spatial predictions of the expected true incidence $\hat{N}$ and the latent covariate of insecticide resistance $\hat{IR}$. The accuracy of each parameter in the complete set $\theta = (k, \sigma)$ was examined by calculating its relative bias from the true underlying value, as
RESULTS

The full results with posterior summaries for all model parameters are reported in the supplementary material (S2). Plots showing the relationship between the simulated and reconstructed malaria incidence and between the simulated and reconstructed insecticide resistance are also presented in supplementary material (S3). Here, we present an overview of these detailed results, by reporting on the values of relative bias $|RB|$ for each explanatory variable, in each model, under the nine different scenarios of reported malaria incidence (Fig. 2).

Model 1 considered only the active sampling points, hence the single column under model 1 in Fig. 2 does not include extended results pertaining to the clinic detection function (see supplementary material S2.1 for full results). Under model 1, the simulated malaria incidence was affected only by the environmental covariates (that were common to all scenarios) including insecticide resistance.

Overall, the results from model 1 showed an average $|RB| = 0.11$ (std. dev. = 0.08). This was a persistent finding across all three simulated patterns for the latent variable (IR), with low values of relative bias arising regardless of the degree of spatial autocorrelation of the simulated insecticide resistance layer.

Model 2, which considered only data from passive case detection, was less able to capture the underlying effects of predictors on the reported malaria incidence (see supplementary material S2.2 for full results). The posterior means of all coefficients showed an overall average $|RB| = 0.89$ (std. dev. = 1.52). A pattern of increasing bias emerged in particular when considering scenarios of increasing spatial autocorrelation in the latent variable of insecticide resistance (Fig. 2). Since model 2 only included the passive detection cases, the latent variable was completely missing from the list.
of covariates. In scenarios $I1A$, $I1B$ and $I1C$, given by the same $IR1$, (low spatial autocorrelation),
the average $|RB|$ was 0.87 (std. dev. = 1.53). Scenarios that assumed an intermediate level of spatial
autocorrelation in insecticide resistance (latent variable $IR2$) generated an average $|RB|$ of 0.89
(std. dev. = 1.54) whereas models assuming the most spatially autocorrelated distribution of
insecticide resistance ($IR3$) generated an average $|RB|$ of 0.91 (std. dev. = 1.50). Contrary to the
coefficients of the process model, posteriors pertaining to the observation model were not sensitive
to the different shapes of the detection function (cases $PA$, $PB$ or $PC$). Posteriors for the parameter $\delta$
of the detection function were highly accurate, with absolute values of relative biases ranging from
0.06 to 0.09 (Fig. 2). This model was able to partly reconstruct disease incidence, but not in areas
with relatively higher levels of insecticide resistance (Fig. 3a).

Model 3 gave the best results in terms of estimating coefficients with low relative biases (see
supplementary material S2.3 for full results). Of particular note is the fact that the parameter for the
latent insecticide resistance variable $RB_{IR}$ showed a low $|RB|$ varying between 0.02 and 0.08.
Overall, the average $|RB|$ across all variables was 0.07 (std. dev. = 0.07). As with model 1, but in
contrast to model 2, the magnitude of bias in estimated parameters was unrelated to the degree of
spatial autocorrelation assumed in the latent variable. Similar to model 2, the parameter associated
with the case detection function ($\delta$) was estimated with good accuracy, but model 3 was more
accurate in mapping case distribution (Fig. 3a, see also comparison of plots in supplementary
materials S3.1 vs S3.2). Additionally, the latent distribution of the layer of insecticide resistance was
accurately reconstructed using model 3 (Fig. 3b, and supplementary material S3.3).

**DISCUSSION**

By analysing a wide range of plausible, simulated data sets of disease incidence and environmental
variables arising from active sampling and passive case detection, we uncovered some of the
disadvantages of analysing these two data types in isolation. Additionally, we propose a novel
modelling framework aimed at achieving complementarity between the two. We found that such an
integrated, spatially-explicit model, which acknowledges both active sampling and passive case
detection, leads to great improvements in precision and accuracy but also enables the
reconstruction of maps for the hidden variable across unsurveyed space.

As expected, when modelling data arising only from active sampling, we achieved high explanatory
power and relatively low bias, because the model had access to measurements of all the covariates
underlying disease incidence. The model considering only data coming from passive case detection
allowed us to estimate the map of malaria incidence with high accuracy. However, posterior
distributions for most parameters were biased which was likely due to missing data for the
important variable of insecticide resistance. This condition reflects a common situation in
epidemiological studies, where passive case detection at health centres can provide a large amount
of long-term data with relatively moderate effort. Our simultaneous estimation of detection
functions as part of model inferences shows how to take account of imperfect reporting which is an
integral characteristic of such opportunistic data.\textsuperscript{12, 27-29, 35}

With our proposed 3\textsuperscript{rd} model, we achieved a good synergy between depth and breadth in inference
by combining the strengths of the first two models, and allowing them to compensate for each
other’s limitations. In contrast to model using only passive case detection, our hybrid modelling
framework allowed us to investigate the effect of all the variables (including the latent one), and to
produce accurate predictive maps of the disease incidence and latent variable which were not
possible with the model considering only active sampling. An important achievement of our
proposed model was the capability to deal with a latent variable, regardless of its level of spatial
autocorrelation. Thus, even in the absence of assumptions or any preliminary information on the
spatial structure of the latent variable (e.g. whether it is akin to uncorrelated “background noise” or
has a highly geography-dependent distribution) this model framework has potential to reconstruct
it.
Our incremental approach showed that the gains in the accuracy of the results, moving from model 1 to model 3, were a direct result of increases in the spatial complexity used by the analytical approaches. Model 1 had no explicit spatial component. Model 2 was used to generate predictions in space but it didn’t explicitly consider spatial structure in its formulation. Model 3, by including the spatial autocorrelation structure in the partly latent variable, led to the best results.

Our approach to latent variables, readily generalises to processes other than insecticide resistance. We chose this particular example of a latent variable, because IR has potential to impact the transmission and control of a wide range of vector-borne diseases, including malaria, but is typically labour-intensive, time-consuming and expensive to measure. Although WHO guidelines classify insecticide resistance in a binary way, the raw data from Tube test bioassays measure the % survival of cohorts of similarly aged females after a given time period of exposure to insecticide treated surfaces. Therefore, to greatly increase the inferential value acquired from such data, we treated insecticide resistance as a continuous variable ranging from 0 to 1. Our approach can be easily extended to more specific measures of insecticide resistance, such as metabolic, cuticular and behavioural resistance, or to other types of predictor data that can be collected in the field through active sampling but are not easily obtainable via passive case detection, such as vector abundance and density.

When simulating and modelling the latent variable, we made an assumption of stationarity (the autocorrelation function didn’t change in space or in time) and monotonicity (the autocorrelation always decreased with distance). These two assumptions can be plausibly relaxed extending our autocorrelation function. For example, non-stationary formulations could be achieved by expressing the rate of autocorrelation decay (ro) as a function of latitude and longitude or time. Alternatively, ro could be expressed as log-linear combination of environmental covariates. Non-monotonic formulations of the autocorrelation function could be produced for cases where periodic patterns
exist in space, but we currently see very little justification for such formulations based on biological first principles.

The ability to account for reporting bias of our response variable, makes our approach easily applicable to other scenarios where an imperfect detection needs to be considered, such as citizen science data or mobile phone surveillance tools. When modelling the detection function, we made similar assumptions (stationarity and monotonicity) to those of the autocorrelation function for the latent variable and we hypothesized the observation process was only affected by distance from health centres. In several real-world scenarios, additional covariates of reporting probability may be involved, such as age and sex of the patient and socioeconomic factors. Borrowing fundamental concepts from Distance sampling, we assumed that at zero distance the probability of reporting the disease was 100%, however asymptomatic disease in apparently healthy people is common, and would not be observed in clinical data. Thus, incomplete detection at zero distance (based on additional calibration data on the frequency of asymptomatic cases) must be considered. Finally, human mobility is unlikely to be strictly related to Euclidean distance, so it may be preferable to use the distance according to road network, when applying this model to real data. Global digital layers quantifying underlying landscape resistance, describe the travel time between any two points on the globe, based on data such as road density, terrain morphology and an political borders, could be easily included in a spatially explicit epidemiological model such as ours. For all of these reasons, we suggest that preliminary analysis using pilot data and focussing only on modelling the detection probability should be carried out before integrating it into the final model.

Our likelihood could be deployed using either a Bayesian or a frequentist setting. It is likely that in real life, most epidemiological data sets will be accompanied by sufficient expert opinion to lead to influential priors, hence we have illustrated using a Bayesian approach. However, we did not assume the existence of expert opinion here, because we were seeking to construct a conservative test of
our methods. The models presented here (in particular our model 3, using both data types) require a high computational effort (see supplementary material for details). Notwithstanding their theoretical simplicity, the need to take spatial structure into account with a large dataset slows down the Bayesian MCMC inference. Other model fitting approaches such as the Integrated Nested Laplace approximation (INLA)\(^93\), may prove capable of providing similarly accurate results but with faster processing \(^94\).

In quantitative ecology, data simulation, by generating random realisations from stochastic processes described by a series of distributional statements, is exceedingly useful \(^71\). Although simulated studies are not guaranteed to be the same as a real epidemiological system, they allow objective validation of proposed frameworks on a wide range of plausible scenarios, easily adaptable to other epidemiological studies. Although our simulation was borrowing its settings from a study specifically looking at malaria, we demonstrated its applicability on a broad range of contrasting scenarios. Therefore we believe that such a framework can successfully work under different epidemiological systems, where a combination of large-scale but opportunistic data are collected at the same time as conducting a small number of localised scientific surveys.

The strength of our proposed analytical approach lies in its ability to use distinct solutions, such as latent process modelling and spatially autoregressive modelling, in a fully integrated framework. In particular, we demonstrated how active sampling and passive case detection, that have so far been considered independently in the context of spatial epidemiology, can be used simultaneously and complimentarily in a package where the strength of one compensates for the drawback of the other. Our method shows promise for complex spatial epidemiology studies, by allowing different parts of the model to glean information from different types of data. Such egalitarian and complementary use of two, or more data types, can be extended to make use of digital or hard copy primary care records, irrespective of the sophistication of the health provision systems, the density of the human population, or the nature of the disease.
Acknowledgements

We would like to thank Laurie Baker, Julie Barker, Fergus Chadwick, Heather McDevitt and Hilary Ranson for their useful comments on an earlier draft of the manuscript. We also thank the two anonymous reviewers who provided essential suggestions to improve the presentation of the statistical elements.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This project has received funding from the Wellcome Trust under grant agreement no [200222/Z/15/Z] MiRA.
REFERENCES


598 59. Mayer SV, Tesh RB and Vasilakis N. The emergence of arthropod-borne viral diseases: A
600 58. Guelbeogo WM, Sagnon NF, Liu F, Besansky NJ and Costantini C. Behavioural divergence of
602 57. Govella NJ, Chaki PP and Killeen GF. Entomological surveillance of behavioural resilience and
604 56. Parham PE, Pople D, Christiansen-Jucht C, Lindsay S, Hinsley W and Michael E. Modeling the
605 role of environmental variables on the population dynamics of the malaria vector Anopheles
607 55. Ranson H and Lissenden N. Insecticide Resistance in African Anopheles Mosquitoes: A
608 role of environmental variables on the population dynamics of the malaria vector Anopheles
610 54. Parham PE, Pople D, Christiansen-Jucht C, Lindsay S, Hinsley W and Michael E. Modeling the
611 role of environmental variables on the population dynamics of the malaria vector Anopheles
613 53. Ranson H, N’Guessan R, Lines J, Moiroux N, Nkuni Z and Corbel V. Pyrethroid resistance in
621 49. Ramsey DS, Caley PA and Robley A. Estimating population density from presence–absence
623 48. Chandler RB and Royle JA. Spatially explicit models for inference about density in unmarked
625 47. Buckland ST. Introduction to Distance Sampling: Estimating Abundance of Biological
627 46. Wiegand T and Moloney KA. Handbook of spatial point-pattern analysis in ecology. CRC
629 45. Illian J, Penttinen A, Stoyan H and Stoyan D. Statistical analysis and modelling of spatial point
631 44. Zhou G, Afrane YA, Malla S, Githeko AK and Yan G. Active case surveillance, passive case
632 surveillance and asymptomatic malaria parasite screening illustrate different age distribution, spatial
635 Finding in the Visceral Leishmaniasis Elimination Initiative in India, Bangladesh, and Nepal.
637 42. Kuznetsov VN, Grjibovski AM, Mariandyshev AO, Johansson E and Bjune GA. A comparison
639 41. Kuznetsov VN, Grjibovski AM, Mariandyshev AO, Johansson E and Bjune GA. A comparison
641 40. Parham PE, Pople D, Christiansen-Jucht C, Lindsay S, Hinsley W and Michael E. Modeling the
642 role of environmental variables on the population dynamics of the malaria vector Anopheles


ecurrents.outbreaks.90e80717c4e67e1a830f17f6e9aaf85de.


FIGURE CAPTIONS

Figure 1 – Location of active sampling sites, simulation of reported malaria reported incidence ($I_{1A}$, ..., $I_{3C}$) under 3 scenarios of insecticide resistance ($IR_1$, $IR_2$, $IR_3$) and 3 scenarios of reporting probability as a function of distance from health centres ($P_A$, $P_B$, $P_C$).

Figure 2 – Visual summary of results of the three Bayesian models of reported malaria incidence ($I$) under different simulated scenarios of insecticide resistance spatial patterns ($IR$) and probability of reporting at health centres ($P$). Model 1 used only active sampling data from some localised surveys, model 2 only passive case detections at health centres, model 3 combined both data sources together. The colour scale refers to the absolute values of the relative bias between the simulated coefficients of the variables involved in the biological process (1 to 6), or the shape parameter of the detection function (7), and the estimate of the same coefficient obtained by the mean of Markov Chain Monte Carlo (MCMC) posterior distributions. (•) indicates that the simulated coefficient is within the corresponding 95% posterior credible interval, (×) indicates that it falls outside.

Figure 3 – Reconstructions of a simulated scenario of a) malaria incidence and b) insecticide resistance using Bayesian models. Figure refers to scenario 3B (see Fig. 1), with a high level of insecticide resistance spatial autocorrelation and an intermediate shape of the detection function. Model 1 used only active sampling data from a small set of localised surveys, model 2 only passive case detections at health centres, model 3 combined both data sources together.
Active sampling sites
Health centres

Probability of case reporting

Insecticide resistance
Reported malaria incidence
For Peer Review
Supplementary material S1

Nelli L., Ferguson H.M., Matthiopoulos J.

Achieving depth and breadth in spatial models of vector-borne diseases: an integrated framework for active survey and passive surveillance data.

The following R scripts are used to generate the simulated dataset and to run the three models presented in the paper. Note that in the paper we presented 9 different scenarios, whereas here we are simulating only 1 scenario with an intermediate level of spatial autocorrelation of insecticide resistance (ro=0.7, see L 57) and an intermediate scenario of detectability (sigma=15, see L 126). Also, note that the dataset provided as supplementary material is a subset of the entire data that we analysed in the paper, therefore some difference in the final results might be expected.

For constant updates on this scripts, please visit https://lucanelli.wordpress.com/r-codes.

R code for data simulation and models

```r
library(rgdal)
library(mvtnorm)
library(raster)
library(rgeos)
library(R2jags)
library(jagstools)

#read the grid in shapefile format. Please find it as paper's supplementary material
"grid.shp"
grid<-readOGR("your.folder.path","grid", GDAL1_integer64_policy=T)

# set the distance from clinics=0 in active sampling points (to set probability of reporting
as 1)
grid$dist_1[grid$sampling==1] <- 0
grid$dist_2[grid$sampling==1] <- 0
grid$dist_3[grid$sampling==1] <- 0
grid$dist_4[grid$sampling==1] <- 0
grid$dist_5[grid$sampling==1] <- 0
grid$dist_6[grid$sampling==1] <- 0
grid$dist_7[grid$sampling==1] <- 0
grid$dist_8[grid$sampling==1] <- 0

# ############################################
# #### - simulate Insecticide Resistance - ###
# ############################################

# function to make a distance matrix (side*side) 2D array
dist.matrix <- function(side)
{
  row.coords <- rep(1:side, times=side)
col.coords <- rep(1:side, each=side)
row.col <- data.frame(row.coords, col.coords)
D <- dist(row.col, method="euclidean", diag=TRUE, upper=TRUE)
D <- as.matrix(D)
return(D)
}

# function to simulate an autocorrelated surface, with exponential decay given by ro
cor.surface <- function(side, global.mu, ro)
{
  D <- dist.matrix(side)
  # scaling the distance matrix by the exponential decay
  SIGMA <- exp(-ro*D)
  mu <- rep(global.mu, times=side*side)
  # sampling from the multivariate normal distribution
  M <- matrix(nrow=side, ncol=side)
}
```
```r
M[] <- rmvnorm(1, mu, SIGMA)
return(M)
}

# parameters
dimension = max(c(grid@bbox[1,2]-grid@bbox[1,1],grid@bbox[2,2]-grid@bbox[2,1]))/1000  
ro <- 0.7  
# Arena  
global.mu <- 0

# simulating the autocorrelated raster
set.seed(1)
ir <- cor.surface(side = side, ro = ro, global.mu = global.mu)  
image(ir)  

# now transform it into a raster and assign it to the cells
ir.raster <- raster(ir)
extent(ir.raster) <- (grid@bbox)
summary(ir.raster)

grid$IR<-(extract(ir.raster, gCentroid(grid,byid=TRUE), na.rm = T, small = T, df = T))$layer
summary(grid$IR)

#scaling environmentalvariables
scale2 <- function(x) {
  sdx <- sqrt(var(x))
  meanx <- mean(x)
  return((x - meanx)/sdx)
}

grid$NDVI<-scale2(grid$NDVI)
grid$RAIN<-scale2(as.integer(grid$RAIN))
grid$TEMP<-scale2(as.integer(grid$TEMP))
grid$HUM<-scale2(grid$HUM)

# set the simulated coefficients and create the matrix of coefficients 'a'
sim.intercept     <- 2.90
sim.beta.NDVI     <- 0.30
sim.beta.RAIN     <- 0.20
sim.beta.TEMP     <- 0.25
sim.beta.HUM      <- 0.20
sim.beta.IR       <- 0.50
a<-rbind(sim.intercept, sim.beta.NDVI, sim.beta.RAIN, sim.beta.TEMP, sim.beta.HUM, sim.beta.IR)

#Create a the covariate matrix 'x'
x<- matrix(c(rep(1,length(grid)), grid$NDVI, grid$RAIN, grid$TEMP, grid$HUM, grid$IR), nrow=1)

#generate a poisson rates L
grid$L<- as.vector(x %*% a)

#generate true incidence N (number of malaria cases under the biological process)
grid$lambda<-exp(grid$L)
set.seed(1)
grid$N<-rpois(length(grid),grid$lambda)

# set probability of reporting (Pd) at each clinic, according to a half normal function with
sigma=15
HN<-function (x,si) {exp(-(x^2)/(2*sigma^2))}
```

---

**Notes:**

- The code block is a simulation for simulating malaria cases in a grid environment.
- It involves various steps including generating a normal distribution for the initial cases, creating an autocorrelated raster, transforming it into a raster, and assigning it to the grid cells.
- The code then scales environmental variables and simulates coefficients for a Poisson regression model, generating Poisson rates and true incidence of malaria cases.
- Probability of reporting is calculated using a half normal function.

---

**Context:**

This code is likely part of a research project or a statistical analysis related to malaria cases, using spatial simulation techniques. The simulations are aimed at understanding the distribution and reporting of malaria cases in a defined area. The code is written in R, a programming language widely used for statistical computing and graphics.
sigma<-15
grid$Pd1<-HN(grid$dist_1,si)
grid$Pd2<-HN(grid$dist_2,si)
grid$Pd3<-HN(grid$dist_3,si)
grid$Pd4<-HN(grid$dist_4,si)
grid$Pd5<-HN(grid$dist_5,si)
grid$Pd6<-HN(grid$dist_6,si)
grid$Pd7<-HN(grid$dist_7,si)
grid$Pd8<-HN(grid$dist_8,si)

# set the overall probability of not reporting any case at all (Q)
grid$Q<-(1-grid$Pd1)*
   (1-grid$Pd2)*
   (1-grid$Pd3)*
   (1-grid$Pd4)*
   (1-grid$Pd5)*
   (1-grid$Pd6)*
   (1-grid$Pd7)*
   (1-grid$Pd8)

# generate the reported incidence (I) under the observation process
n.clinics<-8
grid$I<-matrix(nrow = length(grid), ncol = n.clinics+1)
for (i in 1:length(grid)) {
  set.seed(i)
  grid$I[i,]<-rmultinom(1, grid$N[i], cbind(grid$Pd1[i],
                grid$Pd2[i],
                grid$Pd3[i],
                grid$Pd4[i],
                grid$Pd5[i],
                grid$Pd6[i],
                grid$Pd7[i],
                grid$Pd8[i],
                grid$Q[i]))
}

# total reported cases at each cell
grid$total.reported.cases<-rowSums(grid$I[,1:n.clinics])

# reported cases from clinics only (set NA in active sampling cells)
grid$incidence.clinics<-grid$I
grid$incidence.clinics[grid$sampling==1]<-NA

# Jags model
model.1<-function() {
  # Priors
  alpha          ~ dnorm(0,0.001)
  beta.NDVI      ~ dnorm(0,0.001)
  beta.RAIN      ~ dnorm(0,0.001)
  beta.TEMP      ~ dnorm(0,0.001)
  beta.HUM       ~ dnorm(0,0.001)
  beta.IR        ~ dnorm(0,0.001)
  si             ~ dgamma(0.01, 0.01)

  for (i in 1:n) {
    count[i]~dbin(1,raw.count[i])
    raw.count[i]~dpois(lambda[i])
    log(lambda[i]) <- alpha +
                        beta.NDVI*NDVI[i] +
                        beta.RAIN*RAIN[i] +
                        beta.TEMP*TEMP[i] +
                        beta.HUM*HUM[i] +
                        beta.IR*IR[i] +
                        eps[i]
params <- c("alpha", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.HUM", "beta.IR")
n.iterations <- 2000

jags.data.survey <- list(count = grid.sub$total.reported.cases,
NDVI = grid.sub$NDVI,
RAIN = grid.sub$RAIN,
TEMP = grid.sub$TEMP,
HUM = grid.sub$HUM,
IR = grid.sub$IR,
n = length(grid.sub))

jags.out.survey <- jags(data = jags.data.survey,
model.file = model.1,
n.chains = 3,
n.iter = n.iterations,
parameters.to.save = params)

##################################################
# Model 2 - passive case detection  ##################
学说 M1

model.2 <- function() {

  # Priors
  alpha        ~ dnorm(0,0.0001)
  beta.NDVI    ~ dnorm(0,0.0001)
  beta.RAIN    ~ dnorm(0,0.0001)
  beta.TEMP    ~ dnorm(0,0.0001)
  beta.HUM     ~ dnorm(0,0.0001)
  sigma        ~ dunif(0,100)
  si           ~ dgamma(0.01, 0.01)

  # Likelihood
  for (i in 1:n) {
    counts[i,1:8] ~ dmulti(probs[i,1:8], reported.cases[i])
    probs[i,1] <- exp(-(dist_1[i]*dist_1[i])/(2*sigma*sigma))
    probs[i,2] <- exp(-(dist_2[i]*dist_2[i])/(2*sigma*sigma))
    probs[i,3] <- exp(-(dist_3[i]*dist_3[i])/(2*sigma*sigma))
    probs[i,4] <- exp(-(dist_4[i]*dist_4[i])/(2*sigma*sigma))
    probs[i,5] <- exp(-(dist_5[i]*dist_5[i])/(2*sigma*sigma))
    probs[i,6] <- exp(-(dist_6[i]*dist_6[i])/(2*sigma*sigma))
    probs[i,7] <- exp(-(dist_7[i]*dist_7[i])/(2*sigma*sigma))
    probs[i,8] <- exp(-(dist_8[i]*dist_8[i])/(2*sigma*sigma))

    reported.cases[i] ~ dbinom(overall_prob[i], true.incidence[i])
    overall_prob[i] <- 1-Q[i]
    Q[i] <- q1[i]*q2[i]*q3[i]*q4[i]*q5[i]*q6[i]*q7[i]*q8[i]
    q1[i] <- 1-probs[i,1]
    q2[i] <- 1-probs[i,2]
    q3[i] <- 1-probs[i,3]
    q4[i] <- 1-probs[i,4]
    q5[i] <- 1-probs[i,5]
    q6[i] <- 1-probs[i,6]
    q7[i] <- 1-probs[i,7]
    q8[i] <- 1-probs[i,8]

    true.incidence[i] ~ dpois(lambda[i])
    log(lambda[i]) <- alpha + beta.NDVI*NDVI[i] +
                      beta.RAIN*RAIN[i] +
                      beta.TEMP*TEMP[i] +
                      beta.HUM*HUM[i] +
                      ep[i]
\begin{verbatim}
  eps[i] ~ dnorm(0, si)
}

params<-c("alpha", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.HUM", "sigma", "true.incidence")

jags.data.clinics <- list(counts=grid$incidence.clinics[,1:8],
                          NDVI=grid$NDVI,
                          RAIN=grid$RAIN,
                          TEMP=grid$TEMP,
                          HUM=grid$HUM,
                          dist_1=grid$dist_1,
                          dist_2=grid$dist_2,
                          dist_3=grid$dist_3,
                          dist_4=grid$dist_4,
                          dist_5=grid$dist_5,
                          dist_6=grid$dist_6,
                          dist_7=grid$dist_7,
                          dist_8=grid$dist_8,
                          n=length(grid))

jags.out.clinics <- jags(data=jags.data.clinics,
                          model.file=model.2,
                          inits=inits,
                          n.chains=3,
                          n.iter=n.iterations,
                          parameters.to.save=params)

############################################################
####### Model 3 - active and passive case together ########
############################################################

DM<-gDistance(gCentroid(grid,byid=TRUE),byid=T)/1000 # calculate distances for covariance matrix

model.3<-function() {
  # Priors
  alpha        ~ dnorm(0,0.0001)
  beta.NDVI    ~ dnorm(0,0.0001)
  beta.RAIN    ~ dnorm(0,0.0001)
  beta.TEMP    ~ dnorm(0,0.0001)
  beta.HUM     ~ dnorm(0,0.0001)
  beta.IR      ~ dnorm(0,0.0001)
  si           ~ dgamma(0.01, 0.01)
  sigma        ~ dunif(0, 100)
  ro           ~ dgamma(0.1,0.1)
  global.mu    ~ dnorm(0,0.0001)

  # Likelihood
  for (i in 1:n) {
    counts[i,1:8]~dmulti(probs[i,1:8], reported.cases[i])
    probs[i,1]<-exp(-(dist_1[i]*dist_1[i])/(2*sigma*sigma))
    probs[i,2]<-exp(-(dist_2[i]*dist_2[i])/(2*sigma*sigma))
    probs[i,3]<-exp(-(dist_3[i]*dist_3[i])/(2*sigma*sigma))
    probs[i,4]<-exp(-(dist_4[i]*dist_4[i])/(2*sigma*sigma))
    probs[i,5]<-exp(-(dist_5[i]*dist_5[i])/(2*sigma*sigma))
    probs[i,6]<-exp(-(dist_6[i]*dist_6[i])/(2*sigma*sigma))
    probs[i,7]<-exp(-(dist_7[i]*dist_7[i])/(2*sigma*sigma))
    probs[i,8]<-exp(-(dist_8[i]*dist_8[i])/(2*sigma*sigma))
    reported.cases[i]~dbinom(overall_prob[i], true.incidence[i])
    overall_prob[i]<-1-Q[i]
    Q[i] <- q1[i]*q2[i]*q3[i]*q4[i]*q5[i]*q6[i]*q7[i]*q8[i]
    q1[i] <- 1-probs[i,1]
  }
}
\end{verbatim}
true.incidence[i] ~ dpois(lambda[i])
log(lambda[i]) <- alpha + beta.NDVI*NDVI[i] + 
beta.RAIN*RAIN[i] +
beta.TEMP*TEMP[i] +
beta.HUM*HUM[i] + eps[i]
eps[i] ~ dnorm(0, sigma)

for(j in 1:n)
{
    # turning the distance matrix to covariance matrix
    C.w[i,j] <- exp(-ro*D[i,j])
}

# turning covariances into precisions
P.w[i,i] <- inverse(C.w[i,i])
mu[i] <- global.mu
IR[i] ~ dmnorm(mu[i], P.w[i,i])

n.iterations<-20000
params<-c("alpha", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma", "ro",
"global.mu", "true.incidence", "IR")

grid$IR.NA<-grid$IR
grid$IR.NA[grid$sampling==0]<-NA #set NA in all the cells but those of sampling points
counts=grid[I[,1:8]]

index<-c(1:nrow(grid))[-which(grid$sampling==1)]     # create an index for NAs cells (that is, all but the 12 sampling stations)
jags.data.both <- list (counts=grid$I[,1:8],
NDVI=grid$NDVI,
RAIN=grid$RAIN,
TEMP=grid$TEMP,
HUM=grid$HUM,
dist_1=grid$dist_1,
dist_2=grid$dist_2,
dist_3=grid$dist_3,
dist_4=grid$dist_4,
dist_5=grid$dist_5,
dist_6=grid$dist_6,
dist_7=grid$dist_7,
dist_8=grid$dist_8,
IR=grid$IR.NA,
n=length(grid),
nn=length(index),
index=index,
D=DM)
jags.out.both<-jags(data=jags.data.both,           #this will take a lot.
model.file=model.3,
n.chains=3,
init=init,
n.iter=n.iterations,
# n.thin=1,
parameters.to.save=params)
# RESULTS - MODEL 1

mu.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,1]

ds.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,2]

LCI.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,3]

UCI.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,7]

Rhat.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,8]

n.eff.model1 <- jagsresults(x=jags.out.survey, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[,9]

res.model1 <- data.frame(mu.vect=mu.model1, sd.vect=sd.model1, LCI=LCI.model1, UCI=UCI.model1, Rhat=Rhat.model1, n.eff=n.eff.model1)

# reconstruction of true incidence from model 1

grid$rec.incid.model1 <- jags.out.survey$BUGSoutput$mean$true.incidence

# transform into a raster (for nicer plotting)

r <- raster()
extent(r) <- extent(grid)

rec.incid.model1.rast <- rasterize(grid, r, 'rec.incid.model1')

N.rast <- rasterize(grid, r, 'N')

par(mfrow=c(1, 3))
plot(N.rast, main="True Incidence")
plot(rec.incid.model1.rast, main="Reconstructed Incidence")
plot(grid$rec.incid.model1-grid$N, main="True vs Reconstructed Incidence")

par(mfrow=c(1, 1))

# RESULTS - MODEL 2

mu.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,1]

ds.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,2]

LCI.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,3]

UCI.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,7]

Rhat.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,8]

n.eff.model2 <- jagsresults(x=jags.out.clinics, params=c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "sigma"))[,9]

res.model2 <- data.frame(mu.vect=mu.model2, sd.vect=sd.model2, LCI=LCI.model2, UCI=UCI.model2, Rhat=Rhat.model2, n.eff=n.eff.model2)

# reconstruction of true incidence from model 2

grid$rec.incid.model2 <- jags.out.clinics$BUGSoutput$mean$true.incidence

# transform into a raster (for nicer plotting)

r <- raster()
extent(r) <- extent(grid)

rec.incid.model2.rast <- rasterize(grid, r, 'rec.incid.model2')

N.rast <- rasterize(grid, r, 'N')

par(mfrow=c(1, 3))
plot(N.rast, main="True Incidence")
plot(rec.incid.model2.rast, main="Reconstructed Incidence")
plot(grid$rec.incid.model2-grid$N, main="True vs Reconstructed Incidence")

par(mfrow=c(1, 1))
### RESULTS - MODEL 3 ###

# results of model 3

```r
mu.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 1]
sd.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 2]
LCI.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 3]
UCI.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 7]
Rhat.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 8]
n.eff.model3 <- jagsresults(x = jags.out.both, params = c("alpha", "beta.HUM", "beta.NDVI", "beta.RAIN", "beta.TEMP", "beta.IR", "sigma"))[, 9]
```

```r
res.model3 <- data.frame(mu.vect = mu.model3,
                         sd.vect = sd.model3,
                         LCI = LCI.model3,
                         UCI = UCI.model3,
                         Rhat = Rhat.model3,
                         n.eff = n.eff.model3)
```

```r
res.model3
```

# reconstruction of true incidence from model 3

```r
grid$rec.incid.model3 <- jags.out.both$BUGSoutput$mean$true.incidence
```

```r
# transform into a raster (for nicer plotting)
r <- raster()
extent(r) <- extent(grid)
rec.incid.model3.rast <- rasterize(grid, r, "rec.incid.model3")
```

```r
par(mfrow = c(1, 3))
plot(N.rast, main = "True Incidence")
plot(rec.incid.model3.rast, col = rev(heat.colors(255)), main = "Reconstructed Incidence")
plot(grid$rec.incid.model3~grid$N, main = "True vs Reconstructed Incidence")
par(mfrow = c(1, 1))
```

# reconstruction of insecticide resistance from model 3

```r
grid$rec.InsRes.model3 <- jags.out.both$BUGSoutput$mean$IR
```

```r
# transform into a raster (for nicer plotting)
r <- raster()
extent(r) <- extent(grid)
rec.InsRes.model3.rast <- rasterize(grid, r, "rec.InsRes.model3")
```

```r
par(mfrow = c(1, 3))
plot(ir.raster, col = rev(heat.colors(255)), main = "True insecticide resistance")
plot(rec.InsRes.model3.rast, col = rev(heat.colors(255)), main = "Reconstructed insecticide resistance")
plot(grid$rec.InsRes.model3~grid$IR, main = "True vs Reconstructed insecticide resistance")
par(mfrow = c(1, 1))
```
Supplementary material S2

Nelli L., Ferguson H.M., Matthiopoulos J.

Achieving depth and breadth in spatial models of vector-borne diseases: an integrated framework for active survey and passive surveillance data.

S2.1 – Result of Bayesian models of reported malaria incidence under different scenarios of insecticide resistance patterns. The table shows results of model 1, which considered only active sampling data from some localised surveys. \( \theta \): simulated coefficient, \( \hat{\theta} \): mean of posterior distribution, CI: credible interval, RB: relative bias.

<table>
<thead>
<tr>
<th>Scenario ( I_A=I_B=I_C )</th>
<th>Variable</th>
<th>( \theta )</th>
<th>( \hat{\theta} ) (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>2.90</td>
<td>2.89 (2.82, 2.95)</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>0.24 (0.14, 0.34)</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.27 (0.20, 0.33)</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.22 (0.14, 0.31)</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.21 (0.08, 0.34)</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>IR</td>
<td>0.50</td>
<td>0.49 (0.44, 0.55)</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>2.90</td>
<td>2.90 (2.83, 2.96)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>0.24 (0.14, 0.34)</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.29 (0.21, 0.36)</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.18 (0.09, 0.27)</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.19 (0.05, 0.32)</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>IR</td>
<td>0.50</td>
<td>0.44 (0.37, 0.50)</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>2.90</td>
<td>2.86 (2.79, 2.92)</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>0.15 (0.04, 0.26)</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.32 (0.24, 0.39)</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.21 (0.11, 0.32)</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.30 (0.15, 0.47)</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>IR</td>
<td>0.50</td>
<td>0.52 (0.44, 0.60)</td>
<td>0.04</td>
</tr>
</tbody>
</table>

https://mc.manuscriptcentral.com/smmr
**S2.2** – Result of Bayesian models of reported malaria incidence under different scenarios of insecticide resistance patterns and detectability at health centres. The table shows results of model 2, which considered only passive case detections at health centres. $\sigma$: shape parameter of half-normal detection function, $\theta$: simulated coefficient, $\hat{\theta}$: mean of posterior distribution, CI: credible interval, RB: relative bias.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Variable</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{1A}$</td>
<td>Intercept</td>
<td>2.90</td>
<td>2.53 (2.51, 2.55)</td>
<td>-0.13</td>
<td>2.90</td>
<td>2.51 (2.49, 2.53)</td>
<td>-0.13</td>
<td>2.90</td>
<td>2.48 (2.46, 2.50)</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>1.01 (0.99, 1.04)</td>
<td>4.07</td>
<td>0.20</td>
<td>1.07 (1.04, 1.09)</td>
<td>4.33</td>
<td>0.20</td>
<td>1.05 (1.05, 1.10)</td>
<td>4.38</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.26 (0.24, 0.28)</td>
<td>-0.13</td>
<td>0.30</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.20</td>
<td>0.30</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.30 (0.28, 0.33)</td>
<td>0.51</td>
<td>0.20</td>
<td>0.30 (0.27, 0.32)</td>
<td>0.48</td>
<td>0.20</td>
<td>0.30 (0.30, 0.35)</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.25 (0.22, 0.27)</td>
<td>-0.02</td>
<td>0.25</td>
<td>0.23 (0.20, 0.25)</td>
<td>-0.10</td>
<td>0.25</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>10.00</td>
<td>10.35 (9.99, 10.39)</td>
<td>0.03</td>
<td>15.00</td>
<td>14.14 (14.00, 14.19)</td>
<td>-0.06</td>
<td>20.00</td>
<td>21.74 (19.64, 21.83)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Variable</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{1B}$</td>
<td>Intercept</td>
<td>2.90</td>
<td>2.50 (2.48, 2.52)</td>
<td>-0.14</td>
<td>2.90</td>
<td>2.50 (2.48, 2.52)</td>
<td>-0.14</td>
<td>2.90</td>
<td>2.50 (2.48, 2.52)</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>1.00 (1.07, 1.13)</td>
<td>4.51</td>
<td>0.20</td>
<td>1.04 (1.02, 1.07)</td>
<td>4.21</td>
<td>0.20</td>
<td>1.04 (1.02, 1.07)</td>
<td>4.20</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.19</td>
<td>0.30</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.19</td>
<td>0.30</td>
<td>0.22 (0.21, 0.24)</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.33 (0.30, 0.35)</td>
<td>0.63</td>
<td>0.20</td>
<td>0.31 (0.29, 0.34)</td>
<td>0.57</td>
<td>0.20</td>
<td>0.28 (0.26, 0.30)</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.20 (0.17, 0.22)</td>
<td>-0.21</td>
<td>0.25</td>
<td>0.23 (0.21, 0.26)</td>
<td>-0.07</td>
<td>0.25</td>
<td>0.23 (0.21, 0.26)</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>10.00</td>
<td>9.46 (9.42, 9.50)</td>
<td>-0.05</td>
<td>15.00</td>
<td>14.93 (14.87, 15.02)</td>
<td>0.00</td>
<td>20.00</td>
<td>20.36 (19.27, 20.45)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Variable</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\hat{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{1C}$</td>
<td>Intercept</td>
<td>2.90</td>
<td>2.42 (2.40, 2.45)</td>
<td>-0.16</td>
<td>2.90</td>
<td>2.42 (2.40, 2.43)</td>
<td>-0.17</td>
<td>2.90</td>
<td>2.39 (2.37, 2.41)</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>HUM</td>
<td>0.20</td>
<td>1.05 (1.02, 1.08)</td>
<td>4.25</td>
<td>0.20</td>
<td>1.05 (1.02, 1.07)</td>
<td>4.25</td>
<td>0.20</td>
<td>1.04 (1.02, 1.07)</td>
<td>4.22</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>0.30</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.19</td>
<td>0.30</td>
<td>0.23 (0.21, 0.25)</td>
<td>-0.23</td>
<td>0.30</td>
<td>0.22 (0.20, 0.24)</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td>RAIN</td>
<td>0.20</td>
<td>0.31 (0.29, 0.34)</td>
<td>0.56</td>
<td>0.20</td>
<td>0.26 (0.24, 0.29)</td>
<td>0.31</td>
<td>0.20</td>
<td>0.26 (0.23, 0.28)</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>TEMP</td>
<td>0.25</td>
<td>0.14 (0.11, 0.16)</td>
<td>-0.45</td>
<td>0.25</td>
<td>0.16 (0.14, 0.19)</td>
<td>-0.34</td>
<td>0.25</td>
<td>0.16 (0.14, 0.18)</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>10.00</td>
<td>9.84 (9.80, 10.01)</td>
<td>-0.02</td>
<td>15.00</td>
<td>14.59 (14.53, 15.65)</td>
<td>-0.03</td>
<td>20.00</td>
<td>19.32 (19.23, 20.41)</td>
<td>-0.03</td>
</tr>
</tbody>
</table>
S2.3 – Result of Bayesian models of reported malaria incidence under different scenarios of insecticide resistance patterns and detectability at health centres. The table shows results of model 3, which considered both active surveys and passive case detections at health centres. $\sigma$: shape parameter of half-normal detection function, $\theta$: simulated coefficient, $\tilde{\theta}$: mean of posterior distribution, CI: credible interval, RB: relative bias.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.90</td>
<td>2.91 (2.87, 2.96)</td>
<td>0.00</td>
<td>2.90</td>
<td>2.92 (2.87, 2.96)</td>
<td>0.01</td>
<td>2.90</td>
<td>2.90 (2.85, 2.94)</td>
<td>0.00</td>
</tr>
<tr>
<td>HUM</td>
<td>0.20</td>
<td>0.17 (0.15, 0.19)</td>
<td>-0.16</td>
<td>0.20</td>
<td>0.17 (0.16, 0.19)</td>
<td>-0.13</td>
<td>0.20</td>
<td>0.20 (0.18, 0.21)</td>
<td>-0.02</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.30</td>
<td>0.30 (0.29, 0.32)</td>
<td>0.01</td>
<td>0.30</td>
<td>0.31 (0.29, 0.32)</td>
<td>0.02</td>
<td>0.30</td>
<td>0.30 (0.29, 0.32)</td>
<td>0.01</td>
</tr>
<tr>
<td>RAIN</td>
<td>0.20</td>
<td>0.18 (0.16, 0.20)</td>
<td>-0.12</td>
<td>0.20</td>
<td>0.17 (0.15, 0.19)</td>
<td>-0.14</td>
<td>0.20</td>
<td>0.21 (0.19, 0.23)</td>
<td>0.03</td>
</tr>
<tr>
<td>TEMP</td>
<td>0.25</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.04</td>
<td>0.25</td>
<td>0.22 (0.20, 0.24)</td>
<td>-0.11</td>
<td>0.25</td>
<td>0.24 (0.22, 0.26)</td>
<td>-0.05</td>
</tr>
<tr>
<td>IR</td>
<td>0.50</td>
<td>0.47 (0.45, 0.48)</td>
<td>-0.07</td>
<td>0.50</td>
<td>0.48 (0.46, 0.49)</td>
<td>-0.04</td>
<td>0.50</td>
<td>0.48 (0.46, 0.49)</td>
<td>-0.04</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>10.00</td>
<td>10.37 (10.00, 10.41)</td>
<td>0.04</td>
<td>15.00</td>
<td>14.80 (14.74, 15.85)</td>
<td>-0.05</td>
<td>20.00</td>
<td>22.58 (22.48, 22.67)</td>
<td>0.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.90</td>
<td>2.92 (2.87, 2.96)</td>
<td>0.01</td>
<td>2.90</td>
<td>2.94 (2.89, 2.99)</td>
<td>0.01</td>
<td>2.90</td>
<td>2.92 (2.87, 2.97)</td>
<td>0.01</td>
</tr>
<tr>
<td>HUM</td>
<td>0.20</td>
<td>0.23 (0.21, 0.25)</td>
<td>0.16</td>
<td>0.20</td>
<td>0.16 (0.14, 0.18)</td>
<td>-0.22</td>
<td>0.20</td>
<td>0.16 (0.14, 0.17)</td>
<td>-0.22</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.30</td>
<td>0.27 (0.25, 0.28)</td>
<td>-0.11</td>
<td>0.30</td>
<td>0.30 (0.28, 0.31)</td>
<td>0.00</td>
<td>0.30</td>
<td>0.29 (0.27, 0.30)</td>
<td>-0.04</td>
</tr>
<tr>
<td>RAIN</td>
<td>0.20</td>
<td>0.16 (0.14, 0.18)</td>
<td>-0.19</td>
<td>0.20</td>
<td>0.18 (0.16, 0.20)</td>
<td>-0.09</td>
<td>0.20</td>
<td>0.15 (0.13, 0.17)</td>
<td>-0.23</td>
</tr>
<tr>
<td>TEMP</td>
<td>0.25</td>
<td>0.18 (0.16, 0.20)</td>
<td>-0.30</td>
<td>0.25</td>
<td>0.22 (0.20, 0.24)</td>
<td>-0.13</td>
<td>0.25</td>
<td>0.21 (0.19, 0.23)</td>
<td>-0.15</td>
</tr>
<tr>
<td>IR</td>
<td>0.50</td>
<td>0.49 (0.47, 0.51)</td>
<td>-0.02</td>
<td>0.50</td>
<td>0.46 (0.42, 0.48)</td>
<td>-0.08</td>
<td>0.50</td>
<td>0.47 (0.45, 0.49)</td>
<td>-0.06</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>10.00</td>
<td>9.47 (9.23, 10.10)</td>
<td>0.04</td>
<td>15.00</td>
<td>14.85 (14.79, 14.9)</td>
<td>-0.01</td>
<td>20.00</td>
<td>20.26 (19.98, 20.34)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
<th>$\theta$</th>
<th>$\tilde{\theta}$ (95% CI)</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.90</td>
<td>2.94 (2.89, 2.99)</td>
<td>0.01</td>
<td>2.90</td>
<td>2.94 (2.88, 2.98)</td>
<td>0.01</td>
<td>2.90</td>
<td>2.92 (2.87, 2.97)</td>
<td>0.01</td>
</tr>
<tr>
<td>HUM</td>
<td>0.20</td>
<td>0.2 (0.18, 0.22)</td>
<td>-0.01</td>
<td>0.20</td>
<td>0.19 (0.18, 0.21)</td>
<td>-0.03</td>
<td>0.20</td>
<td>0.20 (0.18, 0.21)</td>
<td>-0.02</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.30</td>
<td>0.3 (0.28, 0.31)</td>
<td>0.00</td>
<td>0.30</td>
<td>0.3 (0.28, 0.31)</td>
<td>0.00</td>
<td>0.30</td>
<td>0.3 (0.28, 0.31)</td>
<td>0.00</td>
</tr>
<tr>
<td>RAIN</td>
<td>0.20</td>
<td>0.18 (0.16, 0.20)</td>
<td>-0.10</td>
<td>0.20</td>
<td>0.18 (0.16, 0.20)</td>
<td>-0.12</td>
<td>0.20</td>
<td>0.17 (0.15, 0.19)</td>
<td>-0.15</td>
</tr>
<tr>
<td>TEMP</td>
<td>0.25</td>
<td>0.20 (0.18, 0.22)</td>
<td>-0.18</td>
<td>0.25</td>
<td>0.23 (0.21, 0.24)</td>
<td>-0.09</td>
<td>0.25</td>
<td>0.23 (0.21, 0.25)</td>
<td>-0.10</td>
</tr>
<tr>
<td>IR</td>
<td>0.50</td>
<td>0.46 (0.43, 0.48)</td>
<td>-0.08</td>
<td>0.50</td>
<td>0.48 (0.45, 0.49)</td>
<td>-0.05</td>
<td>0.50</td>
<td>0.48 (0.46, 0.50)</td>
<td>-0.03</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>10.00</td>
<td>9.76 (9.72, 10.81)</td>
<td>0.02</td>
<td>15.00</td>
<td>14.48 (14.42, 15.00)</td>
<td>-0.03</td>
<td>20.00</td>
<td>19.55 (19.27, 20.03)</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

https://mc.manuscriptcentral.com/smmr
Supplementary material S3

Nelli L., Ferguson H.M., Matthiopoulos J.

Achieving depth and breadth in spatial models of vector-borne diseases: an integrated framework for active survey and passive surveillance data.

S3.1 – Plots as showing the relationship between the simulated and reconstructed malaria incidence by model 2.
S3.2 – Plots as showing the relationship between the simulated and reconstructed malaria incidence by model 3.
S3.3 – Plots as showing the relationship between the simulated and reconstructed *insecticide resistance* by model 3.