Received: 6 September 2019 | Accepted: 22 February 2021

DOI: 10.1111/rssc.12484

ORIGINAL ARTICLE

Mapping malaria by sharing spatial information between incidence and prevalence data sets

Tim C. D. Lucas[1] | Anita K. Nandi[1] | Elisabeth G. Chestnutt[1] | Katherine A. Twohig[1] | Suzanne H. Keddie[1] | Emma L. Collins[1] | Rosalind E. Howes[1] | Michele Nguyen[1] | Susan F. Rumisha[1] | Andre Python[1] | Rohan Arambepola[1] | Amelia Bertozzi-Villa[1,2] | Penelope Hancock[1] | Punam Amratia[1] | Katherine E. Battle[1] | Ewan Cameron[1] | Peter W. Gething[1,3,4] | Daniel J. Weiss[1]

[1]Big Data Institute, University of Oxford, Oxford, UK

[2]Institute for Disease Modeling, Bellevue, Washington, USA

[3]Telethon Kids Institute, Perth Children's Hospital, Perth, Australia

[4]Curtin University, Perth, Australia

Correspondence
Tim C. D. Lucas, Big Data Institute, University of Oxford, Oxford, UK.
Email: timcdlucas@gmail.com

Abstract

As malaria incidence decreases and more countries move towards elimination, maps of malaria risk in low-prevalence areas are increasingly needed. For low-burden areas, disaggregation regression models have been developed to estimate risk at high spatial resolution from routine surveillance reports aggregated by administrative unit polygons. However, in areas with both routine surveillance data and prevalence surveys, models that make use of the spatial information from prevalence point-surveys might make more accurate predictions. Using case studies in Indonesia, Senegal and Madagascar, we compare the out-of-sample mean absolute error for two methods for incorporating point-level, spatial information into disaggregation regression models. The first simply fits a binomial-likelihood, logit-link, Gaussian random field to prevalence point-surveys to create a new covariate. The second is a multi-likelihood model that is fitted jointly to prevalence point-surveys and polygon incidence data. We find that in most cases there is no difference in mean absolute error between models. In only one case, did the new models perform the best. More generally, our results demonstrate that combining these types of data has the potential to reduce absolute error in estimates of malaria incidence but that simpler baseline models should always be fitted as a benchmark.

KEYWORDS

disaggregation regression, disease mapping, geostatistics, joint modelling, spatial statistics

1 | INTRODUCTION

Global malaria incidence has decreased dramatically over the last 20 years (Battle et al., 2019; Bhatt et al., 2015; Weiss et al., 2019). This decrease has been accompanied by a strategic shift aiming for elimination in low-incidence countries (Newby et al., 2016; World Health Organization, 2016). Accurate, high-resolution maps of malaria risk are vital in countries in the elimination and pre-elimination phases as they highlight the areas with ongoing *Plasmodium* transmission most in need of interventions (Cohen et al., 2017; Sturrock et al., 2016). Two important data sources for malaria mapping are cluster-level surveys of prevalence (Bhatt et al., 2015; Bhatt et al., 2017; Gething et al., 2011; Gething et al., 2012) and routine surveillance data, typically aggregated by administrative unit polygons (Cibulskis et al., 2011; Ohrt et al., 2015; Sturrock et al., 2016). These data sources have different strengths and different spatial coverage. In low-burden areas, very large sample sizes are needed before a prevalence survey is informative because so few individuals have detectable parasitaemia that most sample points will have no cases. Routine surveillance data can be more sensitive than prevalence point-surveys in low-transmission areas because the entire public health system is being used to passively monitor disease occurrence continually over a period of time (Cibulskis et al., 2011). The availability and quality of routine surveillance data of malaria case counts can be poor, but is improving (Cibulskis et al., 2011; Ohrt et al., 2015). Therefore, when the study area contains both low- and medium/high-burden areas, or when prevalence surveys and routine surveillance provide complementary spatial coverage, models that use both of these data sources have the potential to improve estimates of malaria prevalence and incidence.

Disaggregation regression methods have been proposed as a way to model malaria burden using polygon-level, routine surveillance records of incidence (Arambepola et al., 2020; Johnson et al., 2019; Li et al., 2012; Sturrock et al., 2014; Taylor et al., 2018; Wilson & Wakefield, 2020). Disaggregation regression requires an aggregation step in which the high-resolution estimates of disease incidence are summed to match the level of the administrative unit at which the incidence data are observed. An important consideration is whether the aggregation step occurs in link function space or in the response space. In the case of the identity link function, the two cases are the same (Moraga et al., 2017; Roksvåg et al., 2019; Wilson & Wakefield, 2020). However, when using a non-linear link function, the two cases imply very different models. In the case of the Normal–Poisson pairing with a log-link function, performing the aggregation step in the link space before transformation back to the response space produces a ‘geometric sum’ operation. This formulation has been used for computational convenience a number of times in the literature (Liu et al., 2011; Wang et al.,

2018) but lacks the natural epidemiological interpretation provided by arithmetic summation in the response space.

There are two broad ways that spatial information from prevalence surveys could be included in a dissaggregation regression model of incidence. First, the information from prevalence surveys could be summarised using a separate model and then included as a covariate in the disaggregation model. If the model used to summarise the prevalence surveys was explicitly spatial, this approach would make the spatial information in the prevalence data available to the disaggregation model, thereby enhancing the ability to spatially disaggregate polygon-level cases within administrative units. However, this approach does not use any additional statistical power from the prevalence data in order to more accurately learn relationships between malaria risk and the environment. When the study area contains medium-burden areas, this is a missed opportunity. When the spatial coverage of data is different but the whole study area is low burden, this method may be appropriate. This broad approach of summarising the information in a different data set using a separate model has previously been used in a number of contexts, including information on animal hosts (Shearer et al., 2016) or summarising temperature suitability for malaria parasites (Weiss et al., 2014b), which were subsequently used as inputs for modelling malaria prevalence (Bhatt et al., 2015; Weiss et al., 2019).

Fully combining observations of incidence and prevalence in a joint model, with multiple likelihoods, addresses the limitations of a simple model using a prevalence map as a covariate. Advantageously, as the additional malariometric data are being used as response data, they provide more statistical power with which to learn relationships between malaria risk and the environment. Such a model can also learn the relationship between different types of malaria response metrics at the same time as making spatial estimates, thereby producing statistically and epidemiologically consistent outputs for both incidence and prevalence. While a joint model provides the opportunity to learn the relationship between prevalence and incidence, this is technically challenging as these two data types measure disease intensity on different scales. Point-surveys are a measurement of prevalence in the range [0, 1] that quantify parasite rate at a specific point in time. In contrast, routine surveillance measures incidence in the range [0, ∞] over a longer period of time (e.g. a year) during which individuals can have multiple malaria infections. The case of using areal and point data together with different likelihoods and different link functions has been examined previously (Wang et al., 2018) but has required that the aggregation step be performed in the link function space. Disaggregation regression models in which the aggregation step is performed in the natural response space have been examined (Taylor et al., 2018; Wilson & Wakefield, 2020), but without combining point data with areal data or using dual likelihoods for multi-metric data.

Here we compare two methods for using spatial information from prevalence surveys to inform a disaggregation model fitted to polygon incidence data of *Plasmodium falciparum* malaria. The first, simpler, model summarises the spatial information in the prevalence point-surveys by fitting a spatial Gaussian process model to the surveys. Predictions from this model are then used as a covariate in the disaggregation model. Second, we formulate a joint model that combines polygon incidence data and prevalence point-surveys using separate likelihoods for both data types. This model therefore has the potential to learn from the prevalence data in two ways. First, it has the potential to learn more accurate relationships between malaria incidence and the environment from any prevalence data that is in medium transmission areas. Second, it can learn from the spatial information in the prevalence data as well. We relate the differing malariometric measures by using a previously estimated relationship within the model (Cameron et al., 2015), which is then adjusted as part of the model fitting process. Unlike previous studies, this model combines areal and point level data, with different likelihoods, without performing the aggregation step in the link function space. We then compare results from the two models with those made using a polygon-only, disaggregation model similar to previous models (Sturrock et al., 2014; Wilson & Wakefield, 2020). All models are fitted to data from Indonesia, Senegal and Madagascar to provide a set of case studies from disparate geographic settings and with differing levels of malaria endemicity.

2 | MATERIALS AND METHODS

2.1 | Malaria data

We used two data sources that quantify malaria burden: prevalence point-surveys and polygon incidence data. Prevalence point-surveys consist of geolocated survey clusters wherein all sampled individuals are tested for malaria and the positive cases as well as the total number of people tested is recorded. Polygon incidence data are aggregated to administrative units (e.g. districts or provinces) summarising data reported from hospitals and health facilities. Unlike the point data, polygon-level reports only include numbers of cases and not the numbers of individuals in each administrative unit. As such, to determine an incidence rate we rely on gridded population surfaces, summarised to administrative unit boundaries, to provide the denominator. The prevalence point-survey data were extracted from the Malaria Atlas Project database (Bhatt et al., 2015; Guerra et al., 2007; Pfeffer et al., 2018). This data includes DHS and MIS data as well as prevalence surveys collated from the literature. As the prevalence point-surveys cover different age ranges they were standardised to the 2- to 10-year range using a previously published model (Smith et al., 2007). This standardisation generally makes small adjustments to the prevalence values. Carrying the uncertainty from this model through to the final model was deemed out of scope for this work. As described, the age standardisation model gives the surveys with zero positive cases a small positive prevalence. The polygon incidence data were collated from various government reports and adjusted for incompleteness using methods defined by Cibulskis and colleagues (Cibulskis et al., 2011). Full details of all the preprocessing performed on this data is given in the supplementary materials of Weiss et al. (2019). These adjustments account for underreporting of clinical cases due to lack of treatment seeking, missing case reports (from a health facility that reported for 11 months in a year for example) and cases that sought medical attention outside the public health systems (Battle et al., 2016). Where species-specific reports were given, these were used, and in reports that did not distinguish between species of *Plasmodium* the national estimate of the ratio between *P. falciparum* and *Plasmodium vivax* cases was used to estimate numbers of *P. falciparum* cases specifically. These adjustments were uniform across each country. The polygon incidence data can be seen in Panel A of Figures 1–3.

We selected Indonesia, Senegal and Madagascar as case examples as they all have abundant subnational surveillance data and country-wide surveys from approximately the same periods. To minimise temporal effects, we selected 1 year of polygon incidence data and the surrounding 5 years of prevalence point-survey data for each country. Within this 5-year period, we considered malaria unchanging and did not model time explicitly. For Indonesia, we selected polygon incidence data from 2012 that covers 379 administrative units, and prevalence data from 2010 to 2014 that consists of 1233 survey clusters (i.e. unique locations), representing 230,747 individuals. For Senegal, we selected 2015 for polygon incidence data (41 administrative units) and 2013–2017 for prevalence data (804 clusters, 17,037 individuals). Six hundred and ninety-eight of these surveys were from the Senegal continuous DHS survey with 201, 114, 161 and 222 surveys in the years 2008–2011. Finally, for Madagascar, we selected 2013 for polygon incidence (110 administrative units) and 2011–2015 for prevalence data (1049 clusters, 36,411 individuals). Of these, 898 surveys were

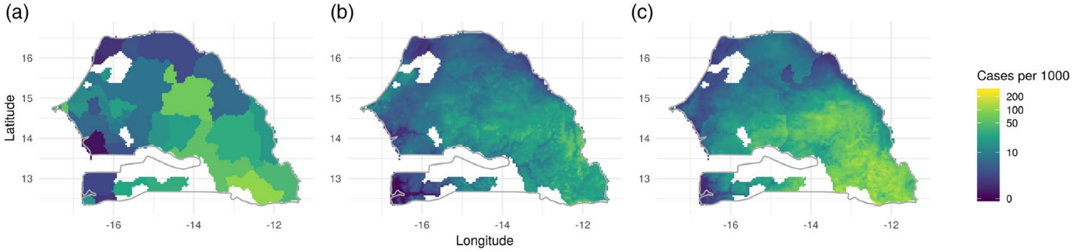

FIGURE 1 Reported incidence data and modelled incidence maps for Senegal. The national boundary of Senegal is shown in grey and missing data is left white. The adjusted input aggregated data is plotted in Panel (a), while Panel (b) maps the predictions of the prevalence Gaussian Process model for spatially cross-validated out-of-sample polygons and Panel (c) maps the predicted incidence from the joint model

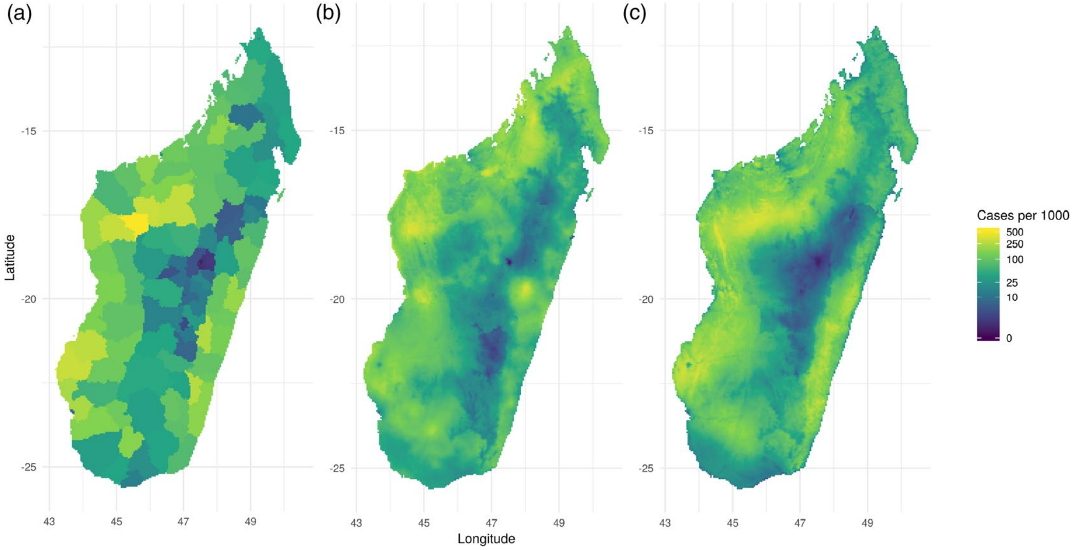

FIGURE 2 Reported incidence data and modelled incidence maps for Madagascar. The adjusted input aggregated data is plotted in Panel (a), while Panel (b) maps the predictions of the prevalence Gaussian Process model for spatially cross-validated out-of-sample polygons and Panel (c) maps the predicted incidence from the joint model

from the Malaria Indicator Surveys in 2011 (266 locations), 2013 (274 locations) and 2016 (358 locations). In DHS surveys, but not MIS surveys, urban areas are typically oversampled. However, this oversampling should be accounted for by the inclusion of an accessibility to cities covariate as described below.

2.2 | Population data

Raster surfaces of population for the years 2005, 2010 and 2015 were created using a hybrid mosaic of data from the Gridded Population of the World v4 (NASA, 2018) and WorldPop (Tatem, 2017), with the latter taking priority for those pixels where both sources had population data. For each year, the

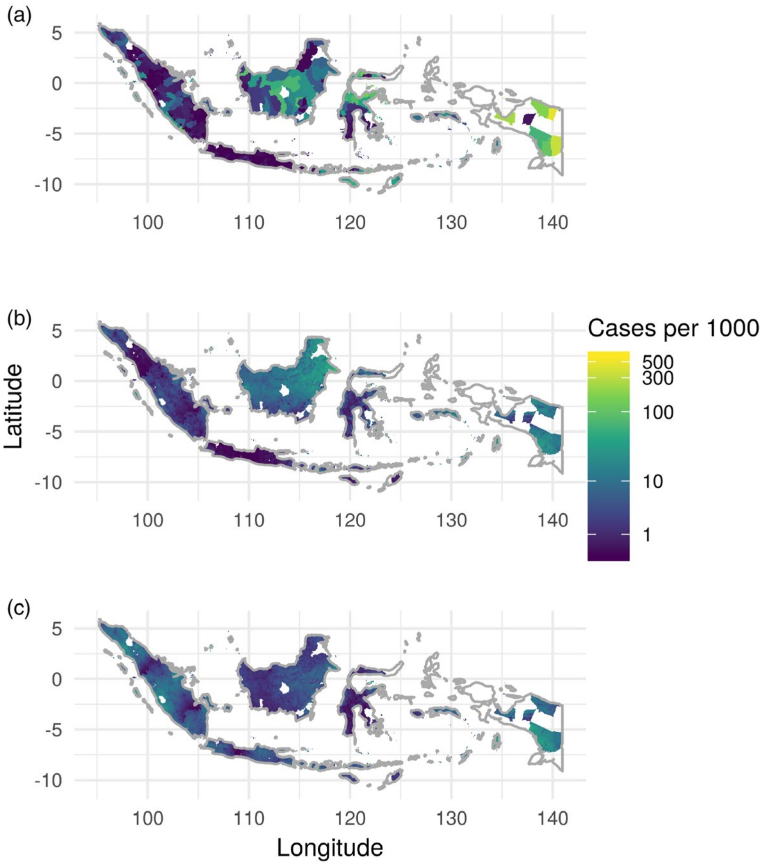

FIGURE 3 Reported incidence data and modelled incidence maps for Indonesia. The national boundary of Indonesia is shown in grey and missing data are left white. The adjusted input aggregated data is plotted in Panel (a), while Panel (b) maps the predictions of the prevalence Gaussian Process model for spatially cross-validated out-of-sample polygons and Panel (c) maps the predicted incidence from the joint model

interpolated population surfaces were adjusted to match national population estimates from the UN. Finally, the population surfaces were masked by environmental suitability so that only populations at risk were included (Weiss et al., 2019).

2.3 | Covariate data

We considered a suite of environmental and anthropological covariates, at a resolution of approximately 5 × 5 kilometres at the equator that included land surface temperature annual mean and standard deviation, enhanced vegetation index (EVI), *P. falciparum* temperature suitability index (Weiss et al., 2014b), elevation (NASA LP DAAC, 2013), tassel cap brightness, tassel cap wetness, accessibility to cities (Weiss et al., 2018), night lights (Elvidge et al., 2017) and proportion of urban land cover (Esch et al., 2018). The land surface temperature, EVI and tasselled cap indices were derived from satellite imagery and gap-filled to remove missing data caused by factors like cloud-cover (Weiss et al., 2014a) and rescaled to a spatial resolution of approximately 5 × 5 km (Weiss et al., 2015) that defined the output of the final prevalence and incidence maps. Some covariates were log-transformed to remove skewness or removed due to multicollinearity with other predictor variables using the threshold of 0.8. The covariates were standardised to have a mean of zero and a standard deviation of one.

2.4 | Baseline disaggregation model

Values at the aggregated, polygon (or areal) level are given the subscript A while pixel or point level variables are indexed with P. The polygon incidence case count data, Count_A is given a Poisson likelihood

$$\text{Count}_A \sim \text{Poisson}(\text{Inc}_A\ \text{pop}_A)$$

where Inc_A is the estimated polygon incidence rate and pop_A is the population at risk within that admin unit polygon (as opposed to the true health-centre catchment area).

The incidence rate is linked to latent pixel-level incidence (Inc_P), prevalence (Prev_P) and predictor variables by the following system of equations.

$$\text{Inc}_A = \frac{\sum_{P \in A} \text{Inc}_P\ \text{pop}_P}{\sum_{P \in A} \text{pop}_P}.$$

Here, $P \in A$ denotes that the summation is over the pixels in polygon A. Incidence is related to prevalence by

$$\text{Inc}_P = \text{PrevInc}(\text{Prev}_P).$$

Here PrevInc is a function from a previously fitted model (Cameron et al., 2015)

$$\text{PrevInc:} f\left(\text{Prev}_P\right) = 2.6 \cdot \text{Prev}_P - 3.6 \cdot (\text{Prev}_P)^2 + 1.6 \cdot (\text{Prev}_P)^3.$$

The linear predictor of the model, η_P, is related to the latent prevalence scale by a typical logit link function

$$\text{Prev}_P = \text{logit}^{-1}(\eta_P).$$

The baseline model is described in terms of prevalence as well as incidence, despite including no prevalence data, so that any differences in predictive ability compared to the full joint model are not simply due to the changed link function. Furthermore, the form of this set of link functions means we calculated predictions of prevalence and incidence simultaneously whether both data types or just one were used which can be useful in applied settings.

The linear predictor is composed of an intercept, β_0, covariates, X_P, and a vector of regression coefficients $\boldsymbol{\beta}$. We also include a spatial, Gaussian random field, $u_P(\rho, \sigma_u)$ and a polygon-level iid random effect, $v_{A_p}(\sigma_v)$.

$$\eta_P = \beta_0 + \boldsymbol{\beta} X_P + u_P(\rho, \sigma_u) + v_{A_p}(\sigma_v).$$

The Gaussian spatial effect $u_P(\rho, \sigma_u)$ has a Matérn covariance function and two hyper parameters: ρ, the nominal range on the longitude–latitude scale (beyond which correlation is < 0.1) and σ_u, the marginal standard deviation. The iid random effect, $v_{A_p} \sim \text{Normal}(\mu = 0, \sigma = \sigma_v)$, was grouped by polygon, with all pixels within polygon A being grouped together (all subsequent normal distributions are also parameterised in terms of the standard deviation). Internally, this effect is parameterised as the log of the precision, $\omega_v = \log(\tau_v) = \log(\frac{1}{\sigma_v^2})$ to improve numeric stability. This random effect modelled both missing covariates and extra-Poisson sampling error.

Finally, we complete the model by setting priors on the parameters $\beta_0, \boldsymbol{\beta}, \rho, \sigma_u$ and σ_v. The intercept was given a wide prior, $\beta_0 \sim \text{Normal}(-2, 4)$, with a mean relating to a prevalence of 0.12 as we know a priori that these countries have low or medium levels of malaria transmission. We set independent, regularising priors on the regression coefficients $\beta_i \sim \text{Normal}(0, 0.04)$. Given the standardised covariates, an intercept of -3 and a regression coefficient from the 95% interquartile range of this distribution, each covariate would be able to predict prevalence between 0.004 and 0.27. This prior encodes our belief that the full range of malaria transmission cannot be explained by a single covariate and our desire to regularise the model. This regularisation is particularly important given the small number of administrative units in Senegal ($n = 46$) and Madagascar ($n = 110$).

We assigned ρ and σ_u a joint penalised complexity prior (Fuglstad et al., 2019) such that $P(\rho < \zeta) = 0.00001$ and $P(\sigma_u > \xi) = 0.00001$. We used different ζ and ξ values for each country: Indonesia $\zeta = 3$, $\xi = 1$, Senegal $\zeta = 1$, $\xi = 0.5$ and Madagascar $\zeta = 1$, $\xi = 1$. This gives a 95% prior credible interval for σ_u of (0.0020, 0.18) and for ρ of (2.2, 7.9) in Madagascar. We believe that a large proportion of the variance of malaria prevalence and incidence cannot be explained by a linear combination of the selected covariates at the scale of individual countries (Bhatt et al., 2017), so we set this prior such that the random field could explain most of the range of the data. As Senegal has a lower range of incidences in the data we set ξ to a smaller value for this country. Plots of this prior are shown in Figures S4 and S5.

We assigned σ_v a penalised complexity prior (Simpson et al., 2017) such that $P(\sigma_v > 0.05) = 0.0000001$. This gives a 95% prior credible interval of $(7.8 \times 10^{-5}, 1.1 \times 10^{-2})$. This was based on a comparison of the variance of Poisson random variables, with rates given by the number of cases observed, and a separately derived upper and lower bound for the case counts using the approach defined by Cibulskis and colleagues (Cibulskis et al., 2011). We found that an iid effect with a standard deviation of 0.05 was able to account for the discrepancy between the assumed Poisson error and the separately derived measurement error.

The models were implemented and fitted in R (R Core Team, 2018) using Template Model Builder (Kristensen et al., 2016) which allows a Laplace approximation of the posterior to be calculated. We note that R-INLA (Lindgren & Rue, 2015) can be used to fit disaggregation models but only when a linear link function is being used (Wilson & Wakefield, 2020). The hyperparameters are fitted using empirical Bayes whereby the hyperparameters are learned from the data but are treated as point estimates rather than using the full posterior of the hyperparameters. We further note that MCMC, even using efficient samplers, is prohibitively slow for these sorts of models (Nandi et al., 2020).

2.5 | Prevalence Gaussian process covariate model

The prevalence Gaussian process model (henceforth the prevalence GP model) is the same as the baseline disaggregation model except that it has one extra covariate. This covariate is created by fitting a Gaussian random field to the prevalence survey data. For each country, we fitted a model with a binomial likelihood

$$z_P \sim \text{Binomial}(\text{Prev}_P, n_P).$$

Here Prev_P is the estimated prevalence and n_P is the observed survey sample size. The model only depended on a Gaussian random field having no covariates

$$\text{Prev}_P = \text{logit}^{-1}(\beta_{\text{GP}} + u_P(\rho, \sigma_u)).$$

We used the same hyperpriors for ρ and σ_u as above. These models were fitted using R-INLA (Lindgren & Rue, 2015). To be in the correct scale for the dissagregation model, the inverse logit of the predicted Gaussian field (i.e. the linear predictor of the model) was used as the additional covariate.

2.6 | Full joint model

The final model is a joint-likelihood model with separate likelihoods for prevalence point-surveys and polygon incidence data. The polygon data are assigned a Poisson likelihood as before. Additionally, the point-survey data, with positive cases z_b, are given a binomial likelihood

$$z_P \sim \text{Binomial}(\text{Prev}_P, n_P)$$

where Prev_P is the estimated prevalence and n_P is the observed survey sample size. As this model has both the prevalence and incidence data, we add a parameter α that modifies the relationship between the two

$$\text{Inc}_P = \exp(\alpha)\text{PrevInc}(\text{Prev}_P).$$

The only further change to the baseline model are in the linear predictor. For incidence data, the linear predictor remains unchanged. The linear predictor for the prevalence data is

$$\eta_P = \beta_0 + \beta_{\text{Prev}} + \boldsymbol{\beta} X_P + u_P(\rho, \sigma_u) + v_{A_p}(\sigma_v) + w_P(\sigma_w).$$

As well as the global intercept, β_0, this model has a prevalence survey specific intercept β_{Prev}. The iid random effect, $v_{A_p} \sim \text{Norm}(0, \sigma_v)$, was again grouped by polygon, with all pixels and point-surveys within polygon A being in the same group as polygon A. The second iid random effect, w_P~Normal$(0, \sigma_w)$, was applied to each point-survey. To improve numeric stability, this effect is also parameterised internally as the log of the precision, $\omega_w = \log(\tau_w) = \log(\frac{1}{\sigma_w^2})$. This effect modelled extra-binomial sampling noise. As such, this random effect is not included in the predicted uncertainty in the incidence or prevalence layers.

We assigned σ_w a penalised complexity prior such that $P(\sigma_w > \phi) = 0.0000001$. This was chosen by finding the maximum difference in prevalence between point-surveys (with a sample size greater than 500 individuals) within the same raster pixel. The differences between points within the same pixel can only be accounted for by the binomial error and this iid effect. Given that the error on a prevalence estimate with sample size greater than 500 is quite small, the iid effect needs to be able to explain this difference. In Senegal and Madagascar, this value was relatively small so we set $\phi = 0.05$. In Indonesia, however, there was a high density of prevalence surveys and heterogeneity in estimated prevalence within single pixels. Therefore we set $\phi = 0.3$.

The PrevInc relationship was fitted to the best available data (matched prevalence and incidence studies both at point level) and this data set. Therefore, we have some a priori confidence in it and from a regularisation standpoint would not wish it to be easily overridden by the data used in these models which are mismatched in spatial scale. Therefore, our prior belief is that $\exp(\alpha)$ is close to one (i.e. the relationship remains unchanged) and therefore that α is close to zero. We set our prior as $\alpha \sim$ Normal(0, 0.001).

2.7 | Experiments

To compare the three models, we used two cross-validation schemes. In the first (random), the incidence data were split into 10 cross-validation folds while all the prevalence data was used in each case (Figure S1). In the second validation scheme, the incidence data was split into spatial cross-validation folds, using k means clustering on polygon centroids, while again all prevalence points were used in all folds (Figure S2). The number of folds was seven for Indonesia, five for Senegal and three for Madagascar. The number of folds was chosen based on the geographical sizes of the countries. This scheme tests specifically whether the joint model can improve predictions by increasing geographic data coverage. In each case, we tested whether the differences between the new models and the baseline model was greater than expected by chance using a paired Wilcox test.

We considered the ability of the model to predict polygon incidence to be our main objective and our performance metric for this was mean absolute error (MAE). As the models were fitted on data on different scales, we found that observations and predictions were sometimes correlated but shifted from the one-one line (i.e. were biased) and therefore correlation metrics were misleading. We also calculated mean errors to assess the bias of the models. To assess how well the models were calibrated we considered coverage of the 80% predictive credible intervals on the hold-out data.

2.8 | Sensitivity analysis

We ran sensitivity analyses for particular aspects of the model and prior specification. We limited our sensitivity analysis to Madagascar and ran random cross-validation on the full joint model with a number of alterations. We tested a weaker prior on α, using $\alpha \sim$ Normal(0, 1) instead of $\alpha \sim$ Normal(0, 0.001). We also examined the sensitivity of the model to the coefficient values used in the PrevInc function. We reran the cross-validation using five draws from a normal distribution, centred on the original parameter values, and with a standard deviation of 0.1. Finally, we reran the cross-validation using weaker priors on the random effects but keeping the relative strengths of the iid and spatial random effects the same. We used $P(\sigma_u > 1) = 0.01$ and $P(\sigma_v > 0.05) = 0.001$.

3 | RESULTS

Under the random cross-validation scheme, there was no significant differences in model performance between models (Table 1). This lack of strong differences is highlighted by there being no clear differences in scatter plots of observed and predicted data across the three methods (Figure S3). Under the spatial cross-validation scheme, the prevalence GP model performed significantly better than baseline in Madagascar and was not significantly different from baseline in Indonesia or Senegal (Table 1, Figure 4). The joint model performed significantly worse than baseline in Indonesia and in Senegal but was the best performing model in Madagascar. Furthermore, notable differences can be seen in the scatter plots of observed and predicted values (Figure 4). In Indonesia, it can be seen that the joint model is more strongly biased at low incidence values with many data points being overpredicted.

TABLE 1 Summary of out-of-sample accuracy for all cross-validation experiments

Cross-validation	Country	Baseline	Prev GP	Joint
Random	Indonesia	13.95	14.09	13.79
	Senegal	12.41	12.37	13.07
	Madagascar	39.06	35.82	39.01
Spatial	Indonesia	14.77	14.77	16.46*
	Senegal	13.09	12.21	15.15*
	Madagascar	67.73	50.38†	44.05†

Mean absolute error of predicted incidence rate against out-of-sample observed data for three countries. Results that are significantly better or worse than the baseline (at the 95% significance level) are indicated with a dagger or asterisk respectively.

However, the joint model clearly performs better in Madagascar with the polygon-only model unable to predict high incidence observations accurately. Maps of out-of-sample predictions, under spatial cross-validation, can be seen in Figures 1–3. In the Indonesian plots, one clear difference is that the joint model overpredicts on the island of Java.

All models seem to be fairly well calibrated (Table 2). The proportion of out-of-sample incidence datapoints being within their 80% credible intervals ranged between 0.51 and 0.88. However, in most cases, coverage was between 0.7 and 0.8 implying that the models were a little overconfident in their predictions. There was no clear difference in calibration between the different models. The bias in the models was generally fairly small (less than 15) with most models having a tendency to underpredict. While the differences in bias were not large, the joint model had the lowest bias in four out six cases (Tables S4–S5). Plots of model error against observed API, under spatial cross are shown in Figures S35–37.

We can further investigate why the models performed as they did by examining the parameters estimated in the models fitted to all data (Tables S1–S3). We can see that for all covariate regression parameters in Madagascar and Senegal (Tables S1–S2) the standard deviation of the posterior is smaller in the joint model than in the prevalence GP model. For Indonesia, the standard deviation of the posteriors is smaller in the prevalence GP model 6 our of 8 times. Similarly, the standard deviation of the posteriors of the hyperparameters of the random field is smaller for the joint model in Senegal and Madagascar but larger for Indonesia. Given that these models are using the additional prevalence data as response data, this is to be expected. While noting that the coverage is similarly acceptable in all models (Table 2), this implies that more is being learned by the joint model and that these models might be useful for analyses that are more focused on estimating specific parameters.

We can also compare the regression parameter for the prevalence GP covariate in the three countries noting that in Indonesia the prevalence GP model performed worse than baseline under random cross-validation and had equal performance to baseline under spatial cross-validation. We see that the regression parameter for this covariate was small in Indonesia (mean = 0.06, sd = 0.12) but relatively large and positive in both Senegal (mean = 0.30, sd = 0.17) and Madagascar (mean = 0.36, sd = 0.07).

In Madagascar, the joint model performed the best in the spatial cross-validation scheme. Comparing the estimated parameters of the joint model between Madagascar and the other two countries therefore is useful. The prevalence intercept, β_p, is large in the Senegal fit (mean = 1.36, sd = 0.13) but small in Indonesia (mean = 0.03, sd = 0.20) and Madagascar (mean = 0.07, sd = 0.10). This implies there is a strong discrepancy (given the prevalence to incidence model) between the prevalence and incidence data in Senegal. Furthermore, the standard deviation of the prevalence point iid effect, $w_b(\sigma_w)$, is much larger in Indonesia (mean $\omega_w = -2.6$, sd $\omega_w = 0.09$ which corresponds to a mean of σ_w of 13.46) than in Senegal (mean $\omega_w = -1.03$, sd $\omega_w = 0.13$ which corresponds to a mean of σ_w of 2.80)

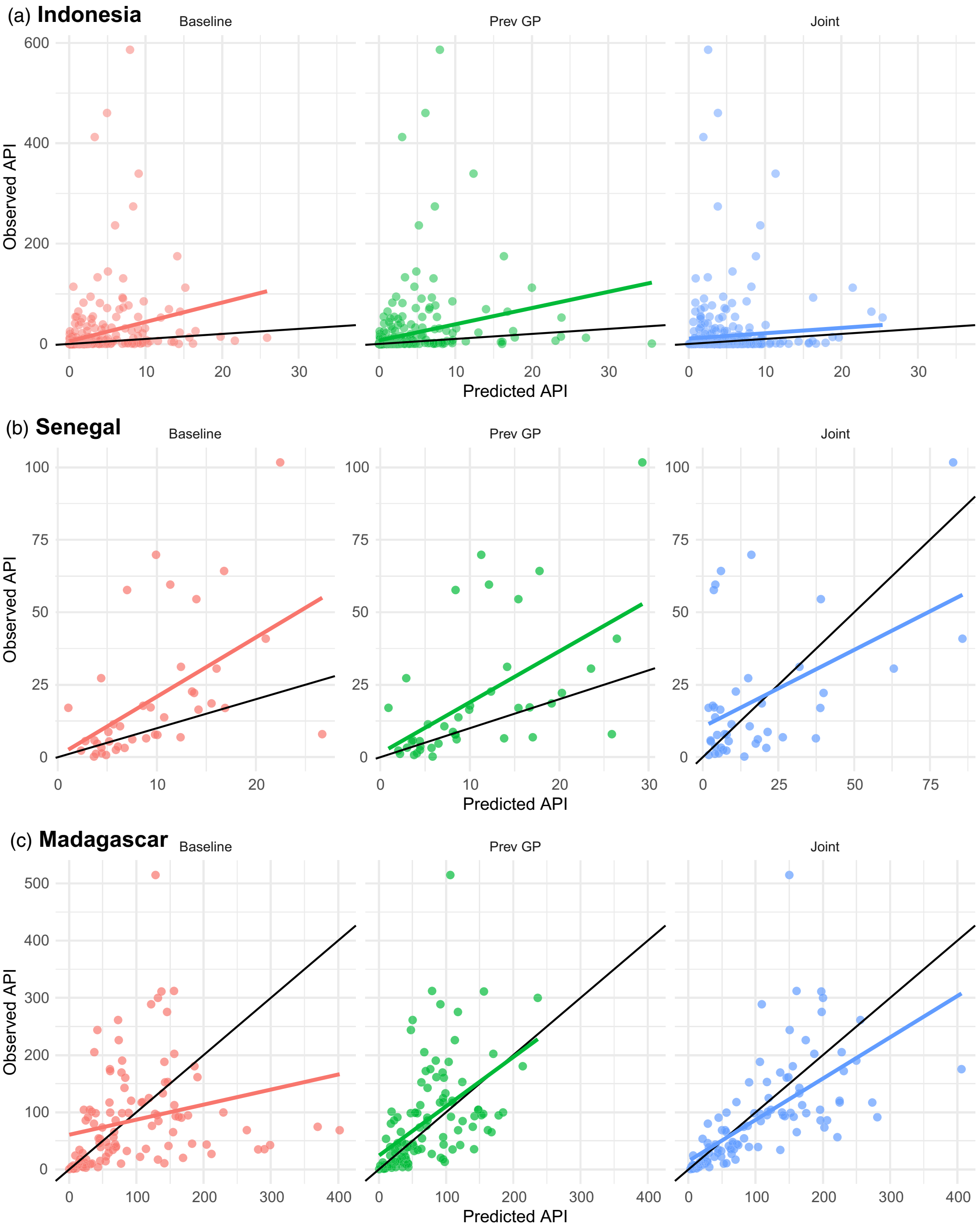

FIGURE 4 Observed-predicted plots of modelled annual malaria incidence (cases per 1000) by country from the spatial cross-validation experiments for Indonesia (Panel a), Senegal (Panel b) and Madagascar (Panel c). Results from the baseline disaggregation model are shown in red, the prevalence GP model is shown in green while the joint model is shown in blue. The one-one line is shown with a black line and a simple linear regression through the points is shown by a coloured line [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Summary of coverage of 80% credible intervals

Cross-validation	Country	Baseline	Prev GP	Joint
Random	Indonesia	0.73	0.72	0.72
	Senegal	0.76	0.78	0.80
	Madagascar	0.78	0.79	0.78
Spatial	Indonesia	0.71	0.72	**0.51**
	Senegal	0.78	0.88	0.71
	Madagascar	**0.67**	0.70	0.72

The proportion of held out data points that fall within their 80% credible intervals. Cases where this is below 0.7 are highlighted in bold.

or Madagascar (mean $\omega_w = -0.77$, sd $\omega_w = 0.11$ which corresponds to a mean of σ_w of 2.16). This implies there is a lot of noise in the prevalence data in Indonesia.

We set a strong prior on α being close to one, encoding our belief that the incidence–prevalence relationship should be close to the previously fitted model. The estimated value for α in all three countries is very close to one (Tables S1–S3). While this might be driven by the prior, this indicates that there is no strong evidence from the data that this relationship should be scaled differently by country. The MAE for Madagascar does not change when run with a much weaker prior on this parameter. However, the MAE for Madagascar varies quite a lot when adding random noise to the parameter values (MAE increases by between 0 and 5.7). Changing the priors on the iid effect and the marginal standard deviation of the Gaussian random field did not alter predictive performance.

4 | DISCUSSION

We have compared the predictive performance (MAE) of three models: a baseline polygon-only model; a disaggregation model with spatial information from prevalence surveys included as an additional covariate from a separate GP model; and a model that jointly learns from polygon incidence data and prevalence point-surveys. The prevalence GP model never performed worse than baseline and performed better than baseline in one case. The joint model performed best in one case but also performed worse than baseline in two cases. Therefore, fitting a spatial Gaussian process to prevalence points and including these predictions seems to be a more reliable way of using spatial information from prevalence points while the full joint model seems to have potential for being the best-performing model in a given situation. The significant differences between models all occurred under spatial cross-validation. This implies that the models are not using the additional data to learn more accurate relationships between the environment and malaria incidence. Instead, it suggests that the models are using the spatial information in the data to improve predictions.

A full joint model using both prevalence surveys and incidence data gains additional statistical power compared to the baseline or prevalence GP models. When the prevalence survey random effect ($w_P(\sigma_w)$) becomes very flexible (when σ_w goes to infinity), this joint model collapses and becomes the baseline model. Therefore, it is worth considering why the performance of this model was sometimes less good than the baseline model that did not benefit from the additional statistical power. The two comparisons in which the joint model did worse than baseline were the cross-validation scheme for Senegal and Indonesia. In Figure S36, we can see that in Indonesia, the joint model overpredicts a number of very low-incidence areas, and by looking at Figure 3, we can see that the island of Java is overpredicted. In Figure S37, we can see that a number of low-incidence and medium-incidence areas in Senegal are overpredicted and Figure S39 shows maps of the polygon level errors. We can see that the baseline model clearly underpredicts many areas while the joint model overpredicts some areas in the centre of the country and in the south-east.

In Indonesia, it seems that the prevalence–incidence relationship is the main factor driving the poor predictions. The relationship increases too quickly at very low incidence values due to the limited flexibility in the model structure and the lack of low incidence points used in model fitting (Cameron et al., 2015). In Figure S38, we show a zoomed in plot of Java with both the incidence data and the prevalence data having been converted to incidence space using the relationship from (Cameron et al., 2015). The prevalence data are sensibly distributed across the island and mostly from a systematic survey. However, the prevalence surveys give a considerably higher incidence than the incidence data. Fitting joint models to just Java, with a more flexible relationship between prevalence and incidence will be considered for future work but is beyond the scope of this paper.

In contrast, in Senegal, there is no clear driver of the poor performance. There is very little incidence data in Senegal and it is probably the least reliable data of the three countries. Figure S39 demonstrates the strong bias in the baseline model compared to the less biased, but higher variance predictions of the joint model. Furthermore, we can see in Figure 1 that the baseline better captures the expected increase in incidence in the south east of the country driven by the transition from desert to savannah. Overall, while the MAE is lower for the baseline model, we think that a subjective argument could be made that the joint model is in fact better as it is less biased and captures expected large scale trends. These predictions would possibly be more useful for risk stratification for example.

In our sensitivity analysis, we found that changing the coefficients in the relationship considerably reduced predictive ability (though weakening the prior on α did not). We might interpret this as suggesting that the prior was too strong but that as the data did not contradict the prior relationship, this did not matter. Future models could potentially be improved by using a more flexible approach for addressing the shortcomings of the prevalence–incidence relationship (Cameron et al., 2015) being used in this context. This could be by estimating the parameters of the polynomial jointly with the rest of the model. Informative priors based on the full posterior of the model fitted by Cameron et al. could be used to regularise this joint fit both to prevent improbable inferences. In contrast, vague priors would allow too much flexibility in this component of the model and the information from the prevalence data might not contribute to informing the regression parameters and spatial random field. This is particularly true for model forms such as a spline or a Gaussian process on the relationship between prevalence and incidence. For the model to handle noisy or biased prevalence point-surveys, the modeller can control the iid random effect on the point-surveys, w_b and the prevalence intercept β_p. Here we have tried to maximise the influence of the prevalence data by setting the prior based on the belief that the random effect should only explain extra-binomial variation that is impossible to derive from the covariates (e.g. based on the differences in prevalence surveys within the same pixel). Weakening this prior will allow the iid effect to explain more of the prevalence point-survey variation which both reduces the potential statistical power gained by adding the point-surveys but also reduces the effects of biased or noisy estimates.

In this research, we have used only linear covariates but previous work has demonstrated that simple linear combinations of environmental covariates cannot fully explain malaria risk (Bhatt et al., 2017). A number of methods could be used to include non-linear effects of covariates and interactions into the model. First, machine learning models could be fitted to the prevalence data and then predictions from these models could be used as covariates in the full model (Bhatt et al., 2017; Lucas et al., 2020). This approach is feasible but would not allow any information from the polygons to inform non-linear relationships. Directly modelling non-linear effects in the full model could be achieved by including simple non-linear functions such as splines (Hundessa et al., 2018; Sewe et al., 2017; Sissoko et al., 2017), though the increased model complexity would require more data than was used in Senegal and Madagascar in this study. Finally, Gaussian process regression, with smoothly varying effects in environmental and geographic space could be used (Law et al., 2018). Unfortunately, each of these options is computationally expensive without variational Bayes or other approximations (Law et al., 2018; Ton et al., 2018), which can be difficult to derive. Additionally these models require a large volume of response data and careful regularisation for good predictive performance.

We used three case studies, limited by the number of countries with good aggregated incidence data as well as good prevalence survey data. Two types of study region seemed likely to benefit from the methods presented here. First, countries like Madagascar with intermediate transmission intensities and good surveillance data and good prevalence data. Second, countries that have lots of prevalence surveys and are adjacent to countries with good reporting systems (e.g. Papua New Guinea and neighbouring Indonesia) might also benefit from models that share information between countries. Given the small number of case studies, it is hard to determine in which countries these methods are likely to be most effective. However, the benefits here were only seen in spatial cross-validation schemes suggesting that the models were only effectively using spatial information.

5 | CONCLUSION

We have presented two methods for combining spatial information from prevalence surveys with disaggregation regression models. We found that while the predicted maps were quite different, the polygon-level predictive performance was usually not significantly different. The prevalence GP model never performs worse than baseline and performs better than baseline in one case. The full joint model performs worse than baseline in two cases but is the best performing model in one case. As more countries produce reliable routine surveillance data, and as more countries reduce their malaria prevalence to the point where prevalence surveys are no longer sensitive, disaggregation regression will become more commonly used. Methods such as those presented here should be explored further and refined to improve disaggregation regression results where and when the requisite data are available.

ACKNOWLEDGEMENTS

The authors acknowledge the National Malaria Control Programme of Madagascar for sharing their routine case data for this analysis.

ORCID

Tim C. D. Lucas https://orcid.org/0000-0003-4694-8107
Andre Python https://orcid.org/0000-0001-8094-7226

REFERENCES

Arambepola, R., Lucas, T.C., Nandi, A.K., Gething, P.W. & Cameron, E. (2020) A simulation study of disaggregation regression for spatial disease mapping. *arXiv preprint arXiv:2005.03604*.

Battle, K.E., Bisanzio, D., Gibson, H.S., Bhatt, S., Cameron, E., Weiss, D.J. et al. (2016) Treatment-seeking rates in malaria endemic countries. *Malaria Journal*, 15, 20.

Battle, K.E., Lucas, T.C., Nguyen, M., Howes, R.E., Nandi, A.K., Twohig, K.A. et al. (2019) Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: A spatial and temporal modelling study. *The Lancet*, 394, 332–343.

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U. et al. (2015) The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526, 207.

Bhatt, S., Cameron, E., Flaxman, S.R., Weiss, D.J., Smith, D.L. & Gething, P.W. (2017) Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of the Royal Society Interface*, 14, 20170520.

Cameron, E., Battle, K.E., Bhatt, S., Weiss, D.J., Bisanzio, D., Mappin, B. et al. (2015) Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nature Communications*, 6, 1–10.

Cibulskis, R.E., Aregawi, M., Williams, R., Otten, M. & Dye, C. (2011) Worldwide incidence of malaria in 2009: Estimates, time trends, and a critique of methods. *PLoS Medicine*, 8, e1001142.

Cohen, J.M., Menach, A., Pothin, E., Eisele, T.P., Gething, P.W., Eckhoff, P.A. et al. (2017) Mapping multiple components of malaria risk for improved targeting of elimination interventions. *Malaria Journal*, 16, 459.

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C. & Ghosh, T. (2017) VIIRS night-time lights. *International Journal of Remote Sensing*, 38, 5860–5879.

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D. et al. (2018) Where we live—a summary of the achievements and planned evolution of the Global urban footprint. *Remote Sensing*, 10, 895.

Fuglstad, G.-A., Simpson, D., Lindgren, F. & Rue, H. (2019) Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114, 445–452.

Gething, P.W., Patil, A.P., Smith, D.L., Guerra, C.A., Elyazar, I.R., Johnston, G.L. et al. (2011) A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal*, 10, 378.

Gething, P.W., Elyazar, I.R., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A. et al. (2012) A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Neglected Tropical Diseases*, 6, e1814.

Guerra, C.A., Hay, S.I., Lucioparedes, L.S., Gikandi, P.W., Tatem, A.J., Noor, A.M. et al. (2007) Assembling a global database of malaria parasite prevalence for the Malaria atlas project. *Malaria Journal*, 6, 17.

Hundessa, S., Williams, G., Li, S., Liu, D.L., Cao, W., Ren, H. et al. (2018) Projecting potential spatial and temporal changes in the distribution of *Plasmodium vivax* and *Plasmodium falciparum* malaria in China with climate change. *Science of the Total Environment*, 627, 1285–1293.

Johnson, O., Diggle, P. & Giorgi, E. (2019) A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. *Statistics in Medicine*, 38, 4871–4887.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B.M. (2016) TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70, 1–21.

Law, H.C., Sejdinovic, D., Cameron, E., Lucas, T., Flaxman, S., Battle, K. et al. (2018) Variational learning on aggregate outputs with Gaussian processes. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. (Eds.) *Advances in neural information processing systems*, Vol. 32. Cambridge: MIT Press, pp. 6084–6094.

Li, Y., Brown, P., Gesink, D.C. & Rue, H. (2012) Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21, 479–507.

Lindgren, F. & Rue, H. (2015) Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63, 1–25. Available from http://www.jstatsoft.org/v63/i19/.

Liu, Z., Le, N.D. & Zidek, J.V. (2011) An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, 22, 340–353.

Lucas, T.C.D., Nandi, A.K., Keddie, S.H., Chestnutt, E.G., Howes, R.E., Rumisha, S.F. et al. (2020) Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence. *Spatial and Spatio-temporal Epidemiology*, 100357. https://doi.org/10.1016/j.sste.2020.100357.

Moraga, P., Cramb, S.M., Mengersen, K.L. & Pagano, M. (2017) A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21, 27–41.

Nandi, A.K., Lucas, T.C., Arambepola, R., Gething, P. & Weiss, D.J. (2020) Disaggregation: An R package for Bayesian spatial disaggregation modelling. *arXiv preprint arXiv:2001.04847*.

NASA. (2018) Gridded Population of the World (GPW), v4. Available from http://sedac.ciesin.columbia.edu/data/collection/gpw-v4.

NASA LP DAAC. (2013) SRTMGL3S: NASA shuttle radar topography mission global 3 arc second sub-sampled. Version 003. (accessed 12 September 2017). Available from https://lpdaac.usgs.gov.

Newby, G., Bennett, A., Larson, E., Cotter, C., Shretta, R., Phillips, A.A. et al. (2016) The path to eradication: A progress report on the malaria-eliminating countries. *The Lancet*, 387, 1775–1784.

Ohrt, C., Roberts, K.W., Sturrock, H.J., Wegbreit, J., Lee, B.Y. & Gosling, R.D. (2015) Information systems to support surveillance for malaria elimination. *The American Journal of Tropical Medicine and Hygiene*, 93, 145–152.

Pfeffer, D.A., Lucas, T.C., May, D., Harris, J., Rozier, J., Twohig, K.A. et al. (2018) malariaAtlas: An R interface to global malariometric data hosted by the Malaria atlas project. *Malaria Journal*, 17, 352.

R Core Team. (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Roksvåg, T., Steinsland, I. & Engeland, K. (2019) A knowledge based spatial model for utilizing point and nested areal observations: A case study of annual runoff predictions in the Voss area. *arXiv preprint arXiv:1904.02519*.

Sewe, M.O., Tozan, Y., Ahlm, C. & Rocklöv, J. (2017) Using remote sensing environmental data to forecast malaria incidence at a rural district hospital in Western Kenya. *Scientific Reports*, 7, 2589.

Shearer, F.M., Huang, Z., Weiss, D.J., Wiebe, A., Gibson, H.S., Battle, K.E. et al. (2016) Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria. *PLoS Neglected Tropical Diseases*, 10, e0004915.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H. et al. (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32, 1–28.

Sissoko, M.S., Sissoko, K., Kamate, B., Samake, Y., Goita, S., Dabo, A. et al. (2017) Temporal dynamic of malaria in a suburban area along the Niger River. *Malaria Journal*, 16, 420.

Smith, D.L., Guerra, C.A., Snow, R.W. & Hay, S.I. (2007) Standardizing estimates of the *Plasmodium falciparum* parasite rate. *Malaria Journal*, 6, 131.

Sturrock, H.J., Cohen, J.M., Keil, P., Tatem, A.J., Le Menach, A., Ntshalintshali, N.E. et al. (2014) Fine-scale malaria risk mapping from routine aggregated case data. *Malaria Journal*, 13, 421.

Sturrock, H.J., Bennett, A.F., Midekisa, A., Gosling, R.D., Gething, P.W. & Greenhouse, B. (2016) Mapping malaria risk in low transmission settings: Challenges and opportunities. *Trends in Parasitology*, 32, 635–645.

Tatem, A.J. (2017) Worldpop, open data for spatial demography. *Scientific Data*, 4, 170004.

Taylor, B.M., Andrade-Pacheco, R. & Sturrock, H.J. (2018) Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society, Series A*, 181, 1125–1150.

Ton, J.-F., Flaxman, S., Sejdinovic, D. & Bhatt, S. (2018) Spatial mapping with Gaussian processes and nonstationary Fourier features. *Spatial Statistics*, 28, 59–78.

Wang, C., Puhan, M.A., Furrer, R., & SNC Study Group. (2018) Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, 23, 72–90.

Weiss, D.J., Atkinson, P.M., Bhatt, S., Mappin, B., Hay, S.I. & Gething, P.W. (2014a) An effective approach for gap-filling continental scale remotely sensed time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98, 106–118.

Weiss, D.J., Bhatt, S., Mappin, B., Van Boeckel, T.P., Smith, D.L., Hay, S.I. et al. (2014b) Air temperature suitability for *Plasmodium falciparum* malaria transmission in Africa 2000–2012: A high-resolution spatiotemporal prediction. *Malaria Journal*, 13, 171.

Weiss, D.J., Mappin, B., Dalrymple, U., Bhatt, S., Cameron, E., Hay, S.I. et al. (2015) Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: A data-intensive variable selection approach. *Malaria Journal*, 14, 68.

Weiss, D., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A. et al. (2018) *A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature*, 553, 333–336.

Weiss, D.J., Lucas, T.C., Nguyen, M., Nandi, A.K., Bisanzio, D., Battle, K.E. et al. (2019) Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: A spatial and temporal modelling study. *The Lancet*, 394, 322–331.

Wilson, K. & Wakefield, J. (2020) Pointless spatial modeling. *Biostatistics*, 21(2), e17–e32.

World Health Organization. (2016) World malaria report 2016. World Health Organization.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Lucas TC, Nandi AK, Chestnutt EG, et al. Mapping malaria by sharing spatial information between incidence and prevalence data sets. *J R Stat Soc Series C*. 2021;70:733–749. https://doi.org/10.1111/rssc.12484