

## Mapping malaria by sharing spatial information between incidence and prevalence datasets

Tim C.D. Lucas

*Big Data Institute, University of Oxford, Oxford, UK.*

E-mail: [timcdlucas@gmail.com](mailto:timcdlucas@gmail.com)

Anita K. Nandi<sup>1</sup>

Elisabeth G. Chestnutt<sup>1</sup>

Katherine A. Twohig<sup>1</sup>

Suzanne H. Keddie<sup>1</sup>

Emma L. Collins<sup>1</sup>

Rosalind E. Howes<sup>1</sup>

Michele Nguyen<sup>1</sup>

Susan F. Rumisha<sup>1</sup>

Andre Python<sup>1</sup>

Rohan Arambepola<sup>1</sup>

Amelia Bertozzi-Villa<sup>1,2</sup>

Penelope Hancock<sup>1</sup>

Punam Amratia<sup>1</sup>

Katherine E. Battle<sup>1</sup>

Ewan Cameron<sup>1</sup>

Peter W. Gething<sup>1, 3, 4</sup>

Daniel J. Weiss<sup>1</sup>

*1. Big Data Institute, University of Oxford, Oxford, UK.*

*2. Institute for Disease Modeling, Bellevue, WA, USA*

*3. Telethon Kids Institute, Perth Childrens Hospital, Perth, Australia*

*4. Curtin University, Perth, Australia*

**Summary.** 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 have great potential. Using case studies in Indonesia, Senegal and Madagascar, we compare two methods for incorporating point-level, spatial information into disaggregation regression models. The first simply fits a 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

## 2 Daniel J. Weiss<sup>1</sup>

and polygon incidence data. We find that the simple model generally performs better than a baseline disaggregation model while the joint model performance was mixed. More generally, our results demonstrate that combining these types of data improves estimates of malaria incidence.

### 0.1. Keywords

Disaggregation regression, disease mapping, geostatistics, joint modelling, spatial statistics.

## Introduction

Global malaria incidence has decreased dramatically over the last 20 years (Bhatt et al., 2015; Weiss et al., 2019; Battle et al., 2019). This decrease has been accompanied by a strategic shift aiming for elimination in low incidence countries (World Health Organization, 2016; Newby et al., 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 (Sturrock et al., 2016; Cohen et al., 2017). Mapping malaria in low burden countries presents new challenges as traditional mapping of prevalence (Gething et al., 2011; Bhatt et al., 2017; Gething et al., 2012; Bhatt et al., 2015) using cluster-level surveys and model-based geostatistics are not necessarily effective in these areas (Sturrock et al., 2016, 2014). 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. In most cases, these large sample sizes are neither logistically nor financially feasible. However, the availability and quality of routine surveillance data of malaria case counts, typically aggregated by administrative unit polygons, is improving, thus providing an alternative source for mapping malaria burden (Sturrock et al., 2016; Ohrt et al., 2015; Cibulskis et al., 2011). Advantageously, the 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).

Disaggregation regression methods have been proposed as a way to model malaria burden using polygon-level, routine surveillance records of incidence (Sturrock et al., 2014; Wilson and Wakefield, 2018; Law et al., 2018; Taylor et al., 2017; Li et al., 2012; Johnson et al., 2019). 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 and Wakefield, 2018). 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

### Mapping malaria by sharing spatial information 3

of times in the literature (Wang et al., 2018; Liu et al., 2011) but lacks the natural epidemiological interpretation provided by arithmetic summation in the response space.

The intent of this research is to assess the utility of extending disaggregation regression modelling approaches to map malaria incidence using both point- and polygon-level response data. In particular it focuses on two aspects where improvements in the disaggregation regression may be possible: (1) for modelling low-burden areas which have better coverage of prevalence surveys than polygon incidence data; and (2) for building the statistical relationships between the response (i.e. malaria incidence) and geospatial predictor data in heterogeneous landscapes. An ancillary benefit of a hybridized approach is that it will simultaneously produce estimates of both prevalence and incidence metrics, which may both be useful for policy makers (Cohen et al., 2017). Models are typically fitted to observations of one metric and then a secondary model is used to convert between prevalence and incidence *post hoc* (Battle et al., 2019; Bhatt et al., 2015), thus missing an opportunity to learn the relationship between prevalence and incidence at the same time as fitting the geographic model.

There are two broad ways that spatial information from prevalence surveys could be included in a disaggregation regression model of incidence. Firstly, 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 provide additional degrees of freedom in order to more accurately learn relationships between malaria risk and the environment, nor does it allow joint predictions of incidence and prevalence from a single model. 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 modeling 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 degrees of freedom 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  $[0, \infty]$  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

#### 4 Daniel J. Weiss<sup>1</sup>

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 (Wilson and Wakefield, 2018; Taylor et al., 2017), 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. Secondly, we formulate a joint model that combines polygon incidence data and prevalence point-surveys using separate likelihoods for both data types. 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 and Wakefield, 2018). 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.

## Materials and methods

### *Malaria data*

We used two data sources that quantify malaria burden: prevalence point-surveys and polygon incidence data. Prevalence point-surveys consist of geo-located survey clusters wherein all sampled individuals are tested for malaria and the positive cases as well as the total number of children tested is recorded. Polygon incidence data is aggregated to administrative units (e.g. districts or provinces) summarizing 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). As the prevalence point-surveys cover different age ranges they were standardised to the 2–10 year-range using a previously published model (Smith et al., 2007). 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; 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 one year of polygon incidence data and the surrounding five years of prevalence point-survey data for each country. Within this five 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 1,233 survey clusters (i.e. unique locations), representing 230,747 individuals. For Senegal we selected 2015 for polygon incidence data (41 administrative units) and 2013 to 2017 for prevalence data (804 clusters, 17,037 individuals). Finally, for Madagascar we selected 2013 for polygon incidence (110 administrative units) and 2011 to 2015 for prevalence data (1,049 clusters, 36,411 individuals).

### *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 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).

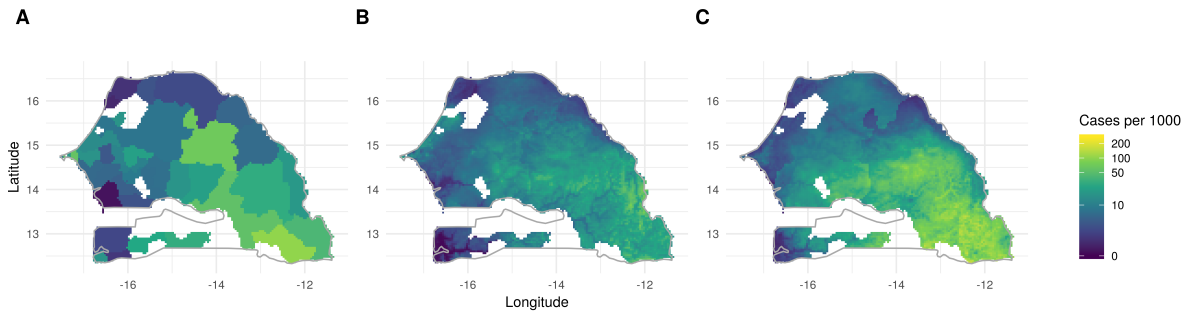
### *Covariate data*

We considered a suite of environmental and anthropological covariates, at a resolution of approximately  $5 \times 5$  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), tasseled cap brightness, tasseled 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 tasseled 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  $5 \times 5$  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.

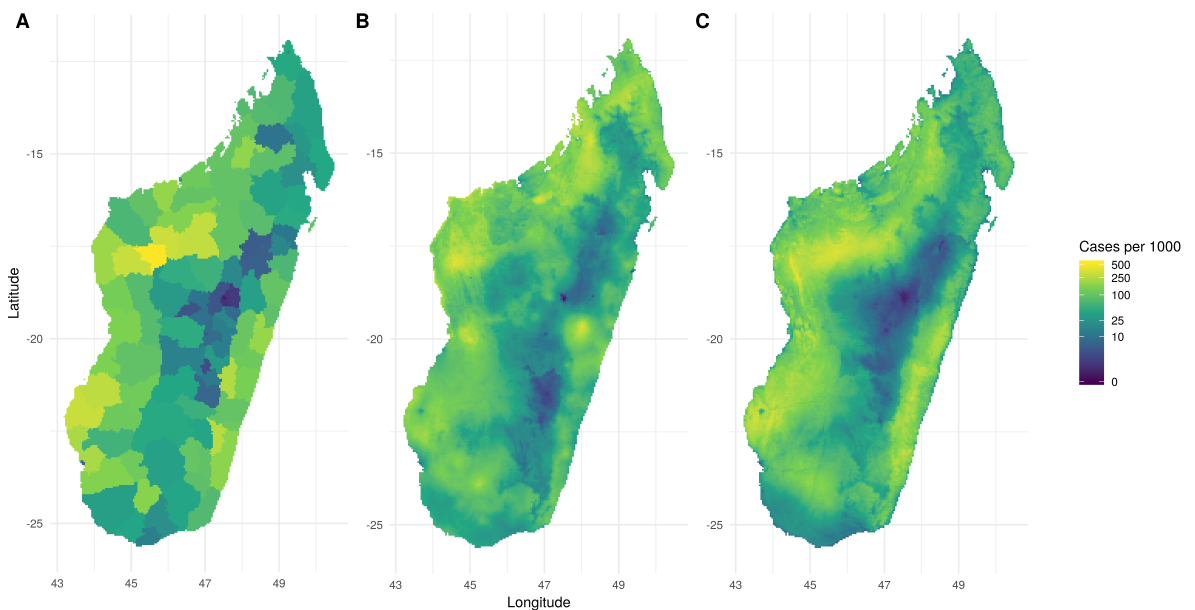
### *Baseline Disaggregation Model*

Values at the aggregated, polygon level are given the subscript  $a$  while pixel or point level variables are indexed with  $b$ . The polygon incidence case count data,  $y_a$  is given a Poisson likelihood

6 Daniel J. Weiss<sup>1</sup>

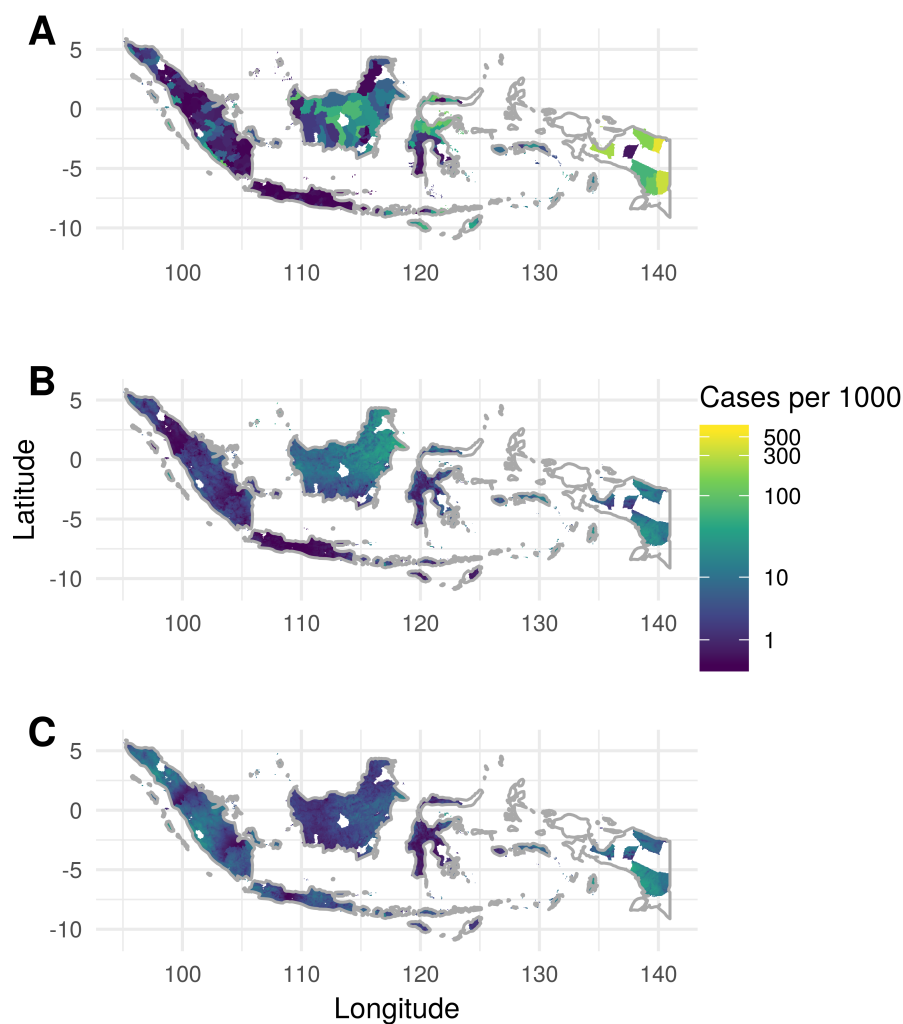


**Fig. 1.** Reported incidence data and modelled incidence maps for Senegal. The national boundary of Senegal is shown in grey and missing data is left white. The adjusted input aggregated data is plotted in Panel A, while Panel B maps the predictions of the prevalence Gaussian Process model for for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



**Fig. 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 for spatially cross-validated out-of-sample polygons and Panel C maps the predicted incidence from the joint model.





**Fig. 3.** Reported incidence data and modelled incidence maps for Indonesia. The national boundary of Indonesia 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 for spatially cross-validated out-of-sample polygons and Panel C maps the predicted incidence from the joint model.

$$y_a \sim \text{Poisson}(i_a \text{pop}_a)$$

where  $i_a$  is the estimated polygon incidence rate and  $\text{pop}_a$  is the population at risk within that admin unit polygon (as apposed to the true health centre catchment area).

Incidence rate is linked to latent pixel-level incidence ( $i_b$ ), prevalence ( $p_b$ ) and predictor variables by the following system of equations.

$$i_a = \frac{\sum_{b \in a} i_b \text{POP}_b}{\sum_{b \in a} \text{POP}_b}$$

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

$$i_b = \text{PrevInc}(p_b).$$

Here  $\text{PrevInc}$  is a function from a previously fitted model (Cameron et al., 2015)

$$\text{PrevInc} : f(p_b) = 2.616p_b - 3.596p_b^2 + 1.594p_b^3.$$

The linear predictor of the model,  $\eta_b$ , is related to the latent prevalence scale by a typical logit link function.

$$p_b = \text{logit}^{-1}(\eta_b)$$

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.

The linear predictor is composed of an intercept,  $b_0$ , covariates,  $X$ , and a vector of regression coefficients  $\beta$ . We also include a spatial, Gaussian random field,  $u_s(\rho, \sigma_u)$  and a polygon-level iid random effect,  $v_a(\sigma_v)$ .

$$\eta_b = \beta_0 + \beta X + u_s(\rho, \sigma_u) + v_a(\sigma_v)$$

The Gaussian spatial effect  $u(s, \rho, \sigma_u)$  has a Matérn covariance function and two hyper parameters:  $\rho$ , the nominal range on the longitude-latitude scale (beyond which correlation is  $< 0.1$ ) and  $\sigma_u$ , the marginal standard deviation. The iid random effect,  $v_a \sim \text{Normal}(0, \sigma_v)$ , was grouped by polygon, with all pixels within polygon  $j$  being grouped together. 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  $\beta_0, \beta, \rho, \sigma_u$  and  $\sigma_v$ . The intercept was given a wide prior,  $b_0 \sim \text{Normal}(-2, 4)$ , 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 prevalences between 0.004 and 0.27. This prior encodes our belief that the full range of malaria transmission can not 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  $\rho$  and  $\sigma_u$  a joint penalised complexity prior (Fuglstad et al., 2018) such that  $P(\rho < \zeta) = 0.00001$  and  $P(\sigma_u > \xi) = 0.00001$ . We used different  $\zeta$  and  $\xi$  values for each country: Indonesia  $\zeta = 3, \xi = 1$ , Senegal  $\zeta = 1, \xi = 0.5$  and Madagascar  $\zeta = 1, \xi = 1$ . We believe that a large proportion of the variance of malaria prevalence and incidence cannot be explained by a linear combination of the covariates selected 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  $\xi$  to a smaller value for this country.

We assigned  $\sigma_v$  a penalised complexity prior (Simpson et al., 2017) such that  $P(\sigma_v > 0.05) = 0.0000001$ . This was based on a comparison of the variance of Poisson random variables, with rates given by the number of cases observed, and an 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 and Rue, 2015) can be used to fit disaggregation models but only when a linear link function is being used (Wilson and Wakefield, 2018). 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.

### *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 binomial likelihood, hierarchical Gaussian random field with the same hyperpriors for  $\rho$  and  $\sigma_u$  as above. These models were fitted using R-INLA (Lindgren and Rue, 2015). To be in the correct scale for the disaggregation model, the inverse logit of the predicted Gaussian field (i.e. the linear predictor of the model) was used as the additional covariate.

### *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

$$z_b \sim \text{Binomial}(p_b, n_b)$$

where  $p_b$  is the estimated prevalence and  $n_b$  is the observed survey sample size. As this model has both prevalence and incidence data we add a parameter  $\alpha$  that modifies the relationship between the two.

$$i_b = \exp(\alpha) \text{PrevInc}(p_b).$$

The only further additions to the baseline model are in the linear predictor which becomes

$$\eta_b = \beta_0 + \mathbf{1}_p \beta_p + \beta X + u_s(\rho, \sigma_u) + v_a(\sigma_v) + w_b(\sigma_w).$$

10 *Daniel J. Weiss<sup>†</sup>*

**Table 1.** Summary of out-of-sample accuracy for all cross-validation experiments. Mean absolute error of predicted incidence rate against out-of-sample observed data for three countries.

Cross-validation	Country	Baseline	Prev GP	Joint
Random	Indonesia	13.95	14.09	<b>13.79</b>
	Senegal	12.41	<b>12.37</b>	13.07
	Madagascar	39.06	<b>35.82</b>	36.36
Spatial	Indonesia	<b>14.77</b>	<b>14.77</b>	16.46
	Senegal	13.09	<b>12.21</b>	15.15
	Madagascar	67.73	50.38	<b>44.05</b>

As well as the global intercept,  $\beta_0$ , this model has a prevalence survey specific intercept  $\beta_p$  where the indicator function,  $\mathbf{1}_p$  denotes that this term is zero except when a prevalence point-survey is being considered. The iid random effect,  $v_a \sim \text{Norm}(0, \sigma_v)$ , 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,  $w_b \sim \text{Normal}(0, \sigma_w)$ , 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  $\sigma_w$  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  $\phi = 0.3$ .

Given that the PrevInc relationship is fitted to the best available data, we have fairly strong *a priori* confidence in it. Therefore, our prior belief is that  $\exp(\alpha)$  is close to one (i.e. the relationship remains unchanged) and therefore that  $\alpha$  is close to zero. We set our prior as  $\alpha \sim \text{Normal}(0, 0.001)$ .

## Experiments

To compare the three models we used two cross-validation schemes. In the first (random), the incidence data was split into ten 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 due to their differing sizes and epidemiological settings. This scheme tests specifically whether the joint model can improve predictions by increasing geographic data coverage.

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. To assess how well the models were calibrated we considered coverage of the 80% predictive credible intervals on the hold-out data.

## Results

Under the random cross-validation scheme, the prevalence GP model performed best in Senegal and Madagascar while the joint model performed best in Indonesia (Table 1). The differences were relatively small in all three countries. 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 4).

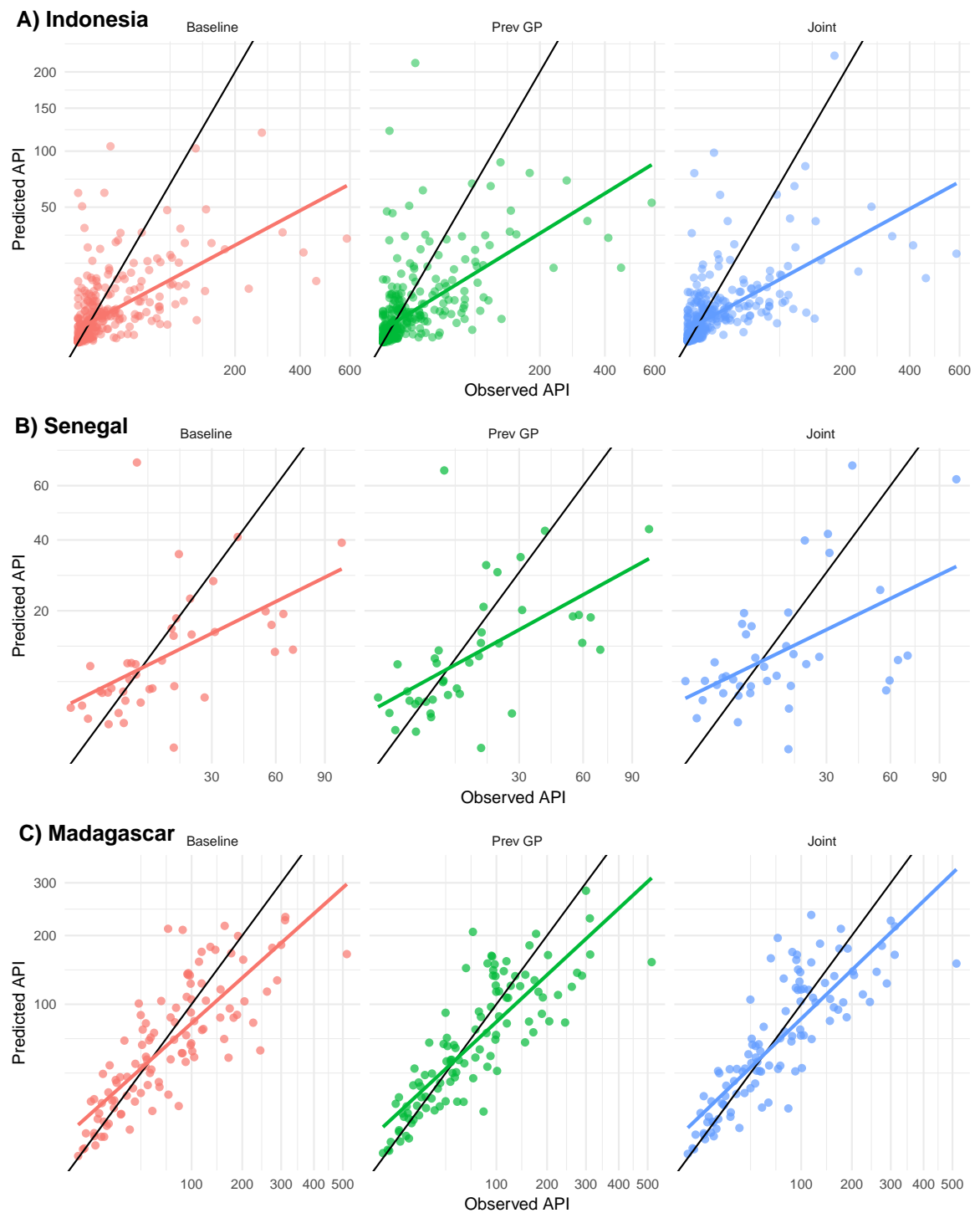
Under the spatial cross-validation scheme, the baseline model and prevalence GP models performed best in Indonesia, the prevalence GP performed best in Senegal while the joint model performed best in Madagascar (Table 1). In contrast to the random cross-validation results, the differences between models was quite strong. Furthermore, notable differences can be seen in the scatter plots of observed and predicted values (Figure 5). In Indonesia it can be seen that the joint model is more strongly biased at low incidence values with many data points being overpredicted. However, the joint model clearly performs better in Madagascar with the polygon-only model unable to predict high incidence observations accurately. Out-of-sample predictions, under spatial cross validation, from the prevalence GP model and full joint model can be seen in Figures 1 – 3.

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.

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). Firstly we can 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).

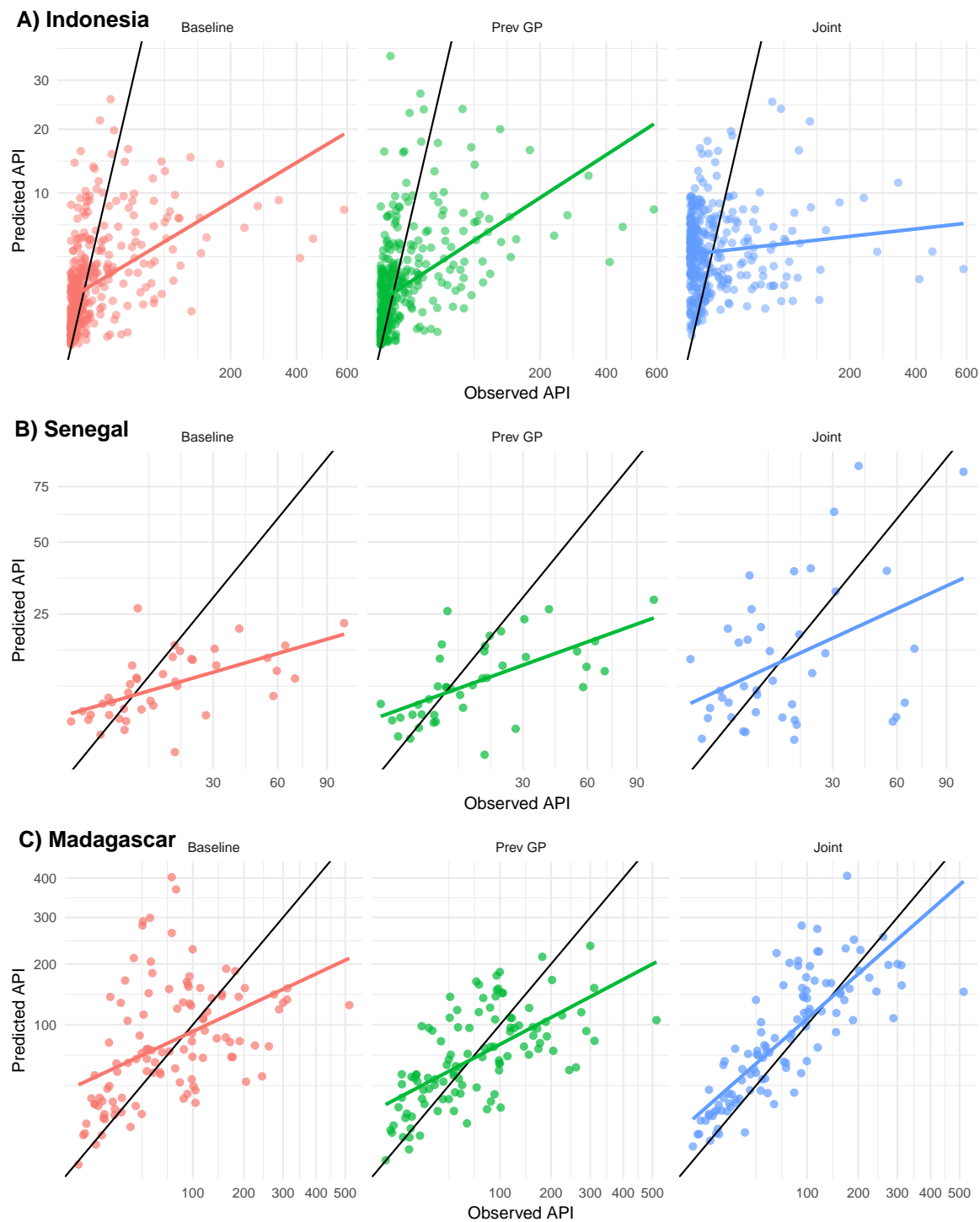
In Madagascar the joint model performed the best in the spatial cross-validation scheme and better than the baseline in the random cross-validation scheme. Comparing the estimated parameters of the joint model between Madagascar and the other two countries therefore is useful. The prevalence intercept,  $\beta_p$ , 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

12 Daniel J. Weiss<sup>1</sup>



**Fig. 4.** Observed-predicted plots (square root scale) of modelled annual malaria incidence (cases per 1000) by country from the random 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.

Mapping malaria by sharing spatial information 13



**Fig. 5.** Observed-predicted plots (square root scale) 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.

14 *Daniel J. Weiss<sup>†</sup>*

**Table 2.** Summary of coverage of 80% credible intervals. 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.

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	<b>0.51</b>
	Senegal	0.78	0.88	0.71
	Madagascar	<b>0.67</b>	0.70	0.72

larger in Indonesia (mean  $\omega_w = -2.6$ , sd  $\omega_w = 0.09$  which corresponds to a mean of  $\sigma_w$  of 13.46) than in Senegal (mean  $\omega_w = -1.03$ , sd  $\omega_w = 0.13$  which corresponds to a mean of  $\sigma_w$  of 2.80) or Madagascar (mean  $\omega_w = -0.77$ , sd  $\omega_w = 0.11$  which corresponds to a mean of  $\sigma_w$  of 2.16). This implies there is a lot of noise in the prevalence data in Indonesia.

We set a strong prior on  $\alpha$  being close to one, encoding our belief that the incidence prevalence relationship should be close to the previously fitted model. The estimated value for  $\alpha$  in all three countries is very close to one (Tables S1–S3). While this might be driven by the prior, we can conclude that there is no strong evidence from the data that this relationship should be scaled differently by country.

Overall, inclusion of the spatial information from prevalence surveys yielded predictions that were as good or better than the baseline model in all six experiments (three countries and two cross-validation schemes). The prevalence GP model was as good or better than baseline in five out of six experiments. In contrast, the joint model was only better than baseline in three out of six experiments.

## Discussion

We have compared the predictive performance 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 Gaussian process (GP) model; and a model that jointly learns from polygon incidence data and prevalence point-surveys. Overall the prevalence GP model appeared to perform best. While the joint model sometimes performed best it also performed worse than baseline in half of the experiments. 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. However, given that this comparison was conducted on datasets from only three countries, it is challenging to draw firm conclusions.

A full joint model using both prevalence surveys and incidence data gains a large number of additional degrees of freedom compared to the baseline or prevalence GP models. Therefore, it is worth considering why the performance of this model was generally less good than the simpler prevalence GP model that did not benefit from the additional degrees of freedom. One potential reason is that the malariometric data are on different scales. Here we have used a previously fitted model (Cameron et al., 2015) to



inform the joint model. However, this model was calibrated using relatively few matched prevalence and incidence surveys as few of these have been conducted and published. Although we added the parameter  $\alpha$  that scales this relationship, it is a very simple scaling. Furthermore the true relationship between prevalence and incidence is likely to vary spatially as aspects such as immunity, seasonality, and population age-structure are not constant (Cameron et al., 2015; Battle et al., 2015; Reiner et al., 2015). In using a joint model we are accepting these limitations in the hope that the benefits of including additional data outweigh the costs of using mismatched data.

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 original model could be used to regularise this joint fit both to prevent improbable inferences but also because if the relationship were too flexible, 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  $\beta_p$ . 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. Firstly, 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). 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 (Sissoko et al., 2017; Sewe et al., 2017; Hundessa et al., 2018), 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. Given the small number of case studies it is hard to determine when these methods are likely to be most effective. However, the greatest benefits here were seen in Madagascar, a country with

more intermediate transmission intensities. In the future, two groups of countries might particularly benefit from the methods presented here. Firstly countries who have had large prevalence surveys in the past and whose reporting systems are improving, such as Ethiopia, might benefit from these methods. Secondly, 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.

## Conclusion

Overall, we have shown that including spatial information from prevalence surveys generally improves the predictive performance of disaggregation regression of aggregated incidence data. However, we found that the more complex joint model was unreliable in its predictive performance. In contrast, summarising the spatial information from the prevalence surveys by fitting a spatial Gaussian process model and using predictions from this model as a new covariate nearly always improved predictive performance. As more countries produce reliable routine surveillance data, and as more countries reduce their malaria prevalences 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 utilized and further refined to improve disaggregation regression results where and when the requisite data are available.

## Supporting information

### Acknowledgments

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

### References

- Battle, K. E., Bisanzio, D., Gibson, H. S., Bhatt, S., Cameron, E., Weiss, D. J., Mappin, B., Dalrymple, U., Howes, R. E., Hay, S. I. et al. (2016) Treatment-seeking rates in malaria endemic countries. *Malaria Journal*, **15**, 20.
- Battle, K. E., Cameron, E., Guerra, C. A., Golding, N., Duda, K. A., Howes, R. E., Elyazar, I. R., Price, R. N., Baird, J. K., Reiner, R. C. et al. (2015) Defining the relationship between *Plasmodium vivax* parasite rate and clinical disease. *Malaria Journal*, **14**, 191.
- Battle, K. E., Lucas, T. C., Nguyen, M., Howes, R. E., Nandi, A. K., Twohig, K. A., Pfeffer, D. A., Cameron, E., Rao, P. C., Casey, D. et al. (2019) Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. *The Lancet*, **394**.

*Mapping malaria by sharing spatial information* 17

- Bhatt, S., Cameron, E., Flaxman, S. R., Weiss, D. J., Smith, D. L. and Gething, P. W. (2017) Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of The Royal Society Interface*, **14**, 20170520.
- Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C., Henry, A., Eckhoff, P. et al. (2015) The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, **526**, 207.
- Cameron, E., Battle, K. E., Bhatt, S., Weiss, D. J., Bisanzio, D., Mappin, B., Dalrymple, U., Hay, S. I., Smith, D. L., Griffin, J. T. et al. (2015) Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nature Communications*, **6**.
- Cibulskis, R. E., Aregawi, M., Williams, R., Otten, M. and 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., Moonen, B., Schapira, A. and Smith, D. L. (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. and 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., Roth, A., Üreyen, S., Zeidler, J., Dech, S. et al. (2018) Where we live—a summary of the achievements and planned evolution of the Global Urban Footprint. *Remote Sensing*, **10**.
- Fuglstad, G.-A., Simpson, D., Lindgren, F. and Rue, H. (2018) Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 1–8.
- Gething, P. W., Elyazar, I. R., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., Patil, A. P., Tatem, A. J., Howes, R. E., Myers, M. F. et al. (2012) A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Neglected Tropical Diseases*, **6**, e1814.
- Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L., Tatem, A. J. and Hay, S. I. (2011) A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal*, **10**, 378.
- Guerra, C. A., Hay, S. I., Lucioparedes, L. S., Gikandi, P. W., Tatem, A. J., Noor, A. M. and Snow, R. W. (2007) Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malaria Journal*, **6**, 17.
- Hundessa, S., Williams, G., Li, S., Li Liu, D., Cao, W., Ren, H., Guo, J., Gasparrini, A., Ebi, K., Zhang, W. 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.

18 Daniel J. Weiss<sup>†</sup>

- Johnson, O., Diggle, P. and 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. and 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. and Fukumizu, K. (2018) Variational learning on aggregate outputs with gaussian processes. In *Advances in Neural Information Processing Systems*, 6081–6091.
- Li, Y., Brown, P., Gesink, D. C. and Rue, H. (2012) Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, **21**, 479–507.
- Lindgren, F. and Rue, H. (2015) Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, **63**, 1–25. URL: <http://www.jstatsoft.org/v63/i19/>.
- Liu, Z., Le, N. D. and Zidek, J. V. (2011) An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, **22**, 340–353.
- Moraga, P., Cramb, S. M., Mengersen, K. L. and 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.
- NASA (2018) Gridded Population of the World (GPW), v4. URL: <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). URL: <https://lpdaac.usgs.gov>.
- Newby, G., Bennett, A., Larson, E., Cotter, C., Shretta, R., Phillips, A. A. and Feachem, R. G. (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. and 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., Dalrymple, U., Guerra, C. A., Moyes, C. L., Thorn, M. 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. URL: <https://www.R-project.org/>.

- Reiner, R. C., Geary, M., Atkinson, P. M., Smith, D. L. and Gething, P. W. (2015) Seasonality of *Plasmodium falciparum* transmission: a systematic review. *Malaria Journal*, **14**, 343.
- Roksvåg, T., Steinsland, I. and 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. and 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., Pigott, D. M., Brady, O. J., Putaporntip, C., Jongwutiwes, S. 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., Yena, M., Dessay, N., Piarroux, R., Doumbo, O. K. 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. and Hay, S. I. (2007) Standardizing estimates of the *Plasmodium falciparum* parasite rate. *Malaria Journal*, **6**, 131.
- Sturrock, H. J., Bennett, A. F., Midekisa, A., Gosling, R. D., Gething, P. W. and Greenhouse, B. (2016) Mapping malaria risk in low transmission settings: challenges and opportunities. *Trends in Parasitology*, **32**, 635–645.
- Sturrock, H. J., Cohen, J. M., Keil, P., Tatem, A. J., Le Menach, A., Ntshalintshali, N. E., Hsiang, M. S. and Gosling, R. D. (2014) Fine-scale malaria risk mapping from routine aggregated case data. *Malaria Journal*, **13**, 421.
- Tatem, A. J. (2017) Worldpop, open data for spatial demography. *Scientific Data*, **4**, 170004–170004.
- Taylor, B. M., Andrade-Pacheco, R. and Sturrock, H. J. (2017) Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A*, **181**.
- Ton, J.-F., Flaxman, S., Sejdinovic, D. and Bhatt, S. (2018) Spatial mapping with Gaussian processes and nonstationary Fourier features. *Spatial Statistics*, **28**, 59–78.
- Wang, C., Puhani, M. A., Furrer, R., Group, S. S. et al. (2018) Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, **23**, 72–90.

20 Daniel J. Weiss<sup>†</sup>

- Weiss, D., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N. 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., Atkinson, P. M., Bhatt, S., Mappin, B., Hay, S. I. and 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. and Gething, P. W. (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., Lucas, T. C., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., Cameron, E., Twohig, K. A., Pfeffer, D. A., Rozier, J. A. et al. (2019) Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet*, **394**.
- Weiss, D. J., Mappin, B., Dalrymple, U., Bhatt, S., Cameron, E., Hay, S. I. and Gething, P. W. (2015) Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach. *Malaria Journal*, **14**, 68.
- Wilson, K. and Wakefield, J. (2018) Pointless spatial modeling. *Biostatistics*.
- World Health Organization (2016) *World malaria report 2016*. World Health Organization.