Model ensembles with different response variables for base and meta models: malaria disaggregation regression combining prevalence and incidence data

Tim C. D. Lucas*[1]; Anita Nandi[1]; Michele Nguyen[1]; Susan Rumisha[1]; Katherine E. Battle[1]; Rosalind E. Howes[1]; Chantal Hendriks[1]; Andre Python[1]; Penny Hancock[1]; Ewan Cameron[1]; Pete Gething[1]; Daniel J. Weiss[1].

1. Malaria Atlas Project, Big Data Institute, University of Oxford, Oxford, UK - timcdlucas@gmail.com

Abstract

Maps of infection risk are a vital tool for the elimination of malaria. Routine surveillance data of malaria case counts, often aggregated over administrative regions, is becoming more widely available and can better measure low malaria risk than prevalence surveys. However, aggregation of case counts over large, heterogeneous areas means that these data are often underpowered for learning relationships between the environment and malaria risk. A model that combines point surveys and aggregated surveillance data could have the benefits of both but must be able to account for the fact that these two data types are different malariometric units. Here, we train multiple machine learning models on point surveys and then combine the predictions from these with a geostatistical disaggregation model that uses routine surveillance data. We find that, in tests using data from Colombia and Madagascar, using a disaggregation regression model to combine predictions from machine learning models trained on point surveys improves model accuracy relative to using the environmental covariates directly.

Keywords: Spatial statistics; Ensemble; Stacking; Epidemiology.

1. Introduction

High-resolution maps of malaria risk are vital for elimination but mapping malaria in low burden countries presents new challenges as traditional mapping of prevalence from cluster-level surveys (Gething et al., 2011; Bhatt et al., 2017; Gething et al., 2012; Bhatt et al., 2015) is often not effective because, firstly, so few individuals are infected that most surveys will detect zero cases, and secondly, because of the lack of nationally representative prevalence surveys in low burden countries (Sturrock et al., 2016, 2014). Routine surveillance data of malaria case counts, often aggregated over administrative regions defined by geographic polygons, is becoming more reliable and more widely available (Sturrock et al., 2016) and recent work has focussed on methods for estimating high-resolution malaria risk from these data (Sturrock et al., 2014; Wilson and Wakefield, 2017; Law et al., 2018; Taylor et al., 2017; Li et al., 2012). However, the aggregation of cases over space means that the data may be relatively uninformative, especially if the case counts are aggregated over large or heterogeneous areas, because it is unclear where within the polygon, and in which environments, the cases occurred. This data is therefore often under-powered for fitting flexible, non-linear models as is required for accurate malaria maps (Bhatt et al., 2017, 2015). A model that combines point surveys and aggregated surveillance data, and therefore leverages the strength of both, has great potential.

One approach for combining these data is to use prevalence point-surveys to train a suite of machine learning models, and then use predictions from these models as covariates in a model trained on polygon-level incidence data. This process of stacking models has proven effective in many realms however typical stacking uses a single dataset on a consistent scale (Sill et al., 2009; Bhatt et al., 2017). Here we propose training the level zero machine learning models on point-level, binomial prevalence data and stacking these models with a polygon-level, Poisson incidence model.

2. Methodology

We used two data sources that reflect *Plasmodium falciparum* malaria transmission; point-prevalence surveys and polygon-level, aggregated incidence data. We selected Colombia and Madagascar as case examples as they both have fairly complete, publicly available, surveillance data at a finer geographical resolution than admin 1. The prevalence survey data were extracted from the Malaria Atlas Project prevalence survey database using only data from 1990 onwards (Bhatt et al., 2015; Guerra et al., 2007). For Colombia we used all points from South America (n = 522) while for Madagascar we used only Malagasy data (n = 1505). We chose these geographic regions based on a trade-off between wanting a large sample size but wanting data from geographically similar areas. The prevalence points were then standardised to an age range of 2–10 using

the model from (Smith et al., 2007). The polygon incidence data were collected from government reports and standardised using methods defined in Cibulskis et al. (2011). This standardisation step accounts for missed cases due to lack of treatment seeking, missing case reports, and cases that sought medical attention outside the public health systems (Battle et al., 2016). For reports where cases were not reported at the species level, national estimates of the ratio between *P. falciparum* and *P. vivax* cases were used to calculate *P. falciparum* only cases. To minimise temporal effects we selected, for each country, one year of surveillance data. We used annual surveillance data from 2015 for Colombia (952 municipalities) and data from 2013 for Madagascar (110 districts) as these years had the most data in each case.

We considered an initial suite of environmental and anthropological covariates, at a resolution of approximately 5×5 kilometres that included the annual mean and log standard deviation of land surface temperature, enhanced vegetation index, malaria parasite temperature suitability index, elevation, tasseled cap brightness, tasseled cap wetness, log accessibility to cities, log night lights and proportion of urban land cover (Weiss et al., 2015). Tasseled cap brightness and urban land cover were subsequently removed as they were highly correlated with other variables. The covariates were standardised to have a mean of zero and a standard deviation of one. These covariates were used for both the machine learning models and the polygon-level models. Raster surfaces of population for the years 2005, 2010 and 2015, were created using data from WorldPop (Tatem, 2017) and from GPWv4 (NASA, 2018) where WorldPop did not have values. Population rasters for the remaining years were created by linear interpolation.

For each country we fitted five models via *caret* (Kuhn et al., 2017): elastic net (Zou and Hastie, 2012), Random Forest (Wright and Ziegler, 2015), projection pursuit regression (Friedman and Stuetzle, 1981), neural networks (Venables and Ripley, 2002) and boosted regression trees (Ridgeway et al., 2017). Our response variable was prevalence and we weighted the data by sample size (i.e. the number of people tested for malaria in each survey. For each model we ran five-fold cross-validation to select hyperparameters using random search for Random Forest and boosted regression trees and grid search for the other models. Predictions from these models were then made across Colombia and Madagascar respectively. These predictions were finally inverse logit transformed so that they are on the linear predictor scale of the top level model.

The top level model was a disaggregation regression model (Sturrock et al., 2014; Wilson and Wakefield, 2017; Law et al., 2018; Taylor et al., 2017; Li et al., 2012). This model is defined by a likelihood at the level of the polygon with covariates and a spatial random field at the pixel-level. Values at the polygon-level are given the subscript a while pixel level values are indexed with b.

The polygon case count data, y_a is given a Poisson likelihood

$$y_a \sim \text{Pois}(i_a \text{pop}_\text{a})$$

where i_a is the estimated polygon incidence rate and pop_a is the observed polygon population-at-risk. This polygon-level likelihood is linked to the pixel level prevalence

$$i_a = \frac{\sum(i_b \text{pop}_b)}{\sum \text{pop}_b}$$

$$i_b = \text{p2i}(p_b)$$

where p2i is from a model that was published previously (Cameron et al., 2015) which defines a function

$$\text{p2i} : f(P) = 2.616P - 3.596P^2 + 1.594P^3.$$

The fact that the model passes through prevalence space ensures that the predictions from the machine learning models can be appropriately scaled. The linear predictor of the model is related to prevalence by a typical logit link function and includes an intercept, β_0, covariates, X with regression parameters β, a spatial, Gaussian, random field, $u(s, \rho, \sigma_u)$, and an *iid* random effect, $v_j(\sigma_v)$.

$$p_b = \text{logit}^{-1}\left(\beta_0 + \beta X + u(s, \rho, \sigma_u) + v_j(\sigma_v)\right)$$

The Gaussian spatial effect has a Matérn covariance function and two hyperparameters: ρ, the nominal range (beyond which correlation is < 0.1) and σ_u, the marginal standard deviation. The *iid* random effect models both missing covariates and extra-Poisson sampling error.

Finally, we complete the model by setting priors on the parameters β_0, β, ρ and σ_u and σ_v. We assigned ρ and σ_u a joint penalised complexity prior (Fuglstad et al., 2018) such that $P(\rho < 1) = 0.00001$ and $P(\sigma_u > 1) = 0.00001$. This prior encoded our *a priori* preference for a simpler, smoother random field. We set this prior such that the random field could explain most of the range of the data if required.
We assigned σ_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 polygon-level cases observed, and an independently derived upper and lower bound for the case counts using the approach defined in (Cibulskis et al., 2011). We found that an *iid* effect with a standard deviation of 0.05 would be able to account for the discrepancy between the assumed Poisson error and the independently derived error. Finally, we set regularising priors on the regression coefficients $\beta_i \sim \text{Norm}(0, 0.4)$. The models were implemented and fitted using Template Model Builder (Kristensen et al., 2016) in R (R Core Team, 2018).
We compared the performance of the models with three sets of covariates, X. Firstly, we used the environmental and anthropogenic covariates, centered and standardised. Secondly, we used the predictions from the machine learning models. Finally we combined these two sets of covariates.
To compare the three models we used two cross-validation schemes. In the first, polygon incidence data was randomly split into six cross-validation folds. In the second, polygon incidence data was split spatially into three folds (via k-means clustering on the polygon centroids). This spatial cross-validation scheme is testing the models' ability to make predictions far from data where the spatial random field is not informative. Our primary performance metric was correlation between observed and predicted data.

3. Results

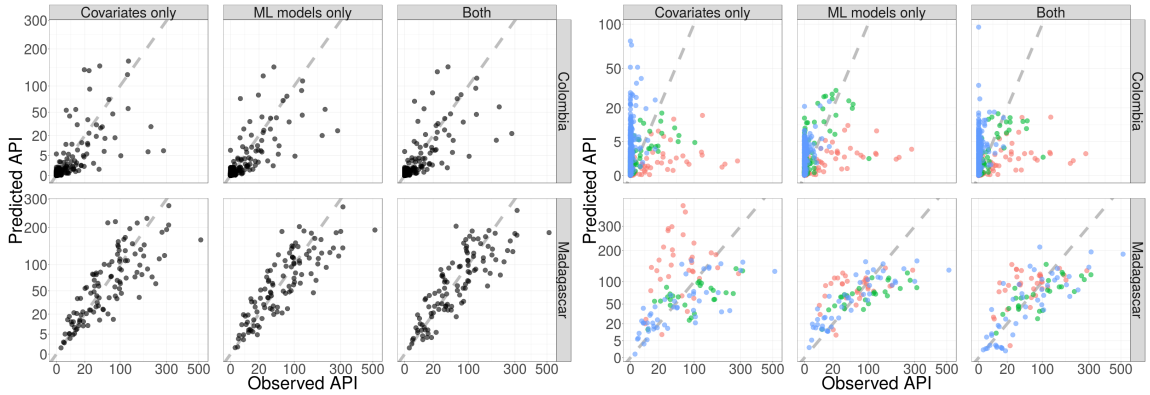
Figure 1 shows the model performance under random and spatial cross-validation for both Madagascar and Colombia. The poor model performance in Colombia under spatial cross-validation indicates that the covariates alone cannot explain malaria incidence in this area. For all other models that use the machine learning predictions as covariates, correlations between observed and predicted data of 0.54 – 0.76 were achieved (Table 1). Input data and mapped out-of-sample predictions of the best performing model, in Colombia, are shown in in Figure 2.

Table 1: Pearson correlations between observed and predicted values.

Cross-validation scheme	Country	Covariates	ML	Covs + ML
Random	Colombia	0.45	**0.55**	0.54
Random	Madagascar	0.70	**0.76**	0.75
Spatial	Colombia	0.05	**0.18**	0.10
Spatial	Madagascar	0.22	**0.63**	0.61

The model using only machine learning predictions as covariates was the best performing model in both countries and both cross-validation schemes (Table 1). As expected, models performed better in the random cross-validation scheme than the spatial cross-validation scheme. The difference between the covariate only model and the machine learning predictions only model was greater in the spatial cross-validation scheme than in the random cross-validation. The improvement in performance between the worst and best models was always smaller than the difference between the random and spatial cross-validation schemes.
Predictive performance of machine learning models was similar, with Random Forest performing best in Madagascar and neural networks, Random Forests and elastic net performing equally well in Colombia (Table 2). The means (across folds) of the regression coefficients (i.e. the weights of the machine learning models in the level zero model) from the polygon-level models that used only predictions from machine learning models as covariates can also be seen in Table 2. The estimated regression parameters are similar between the random and spatial cross-validation schemes. However, the best performing machine learning models do not have the largest estimated regression coefficients as would be expected if prevalence and incidence were completely correlated. Also of note is that some models were estimated to have a negative relationship with incidence (conditional on the inclusion of predictions from other machine learning models).

5. Conclusions



(a) Random cross-validation (b) Spatial cross-validation

Figure 1: Observed data against predictions for cross-validation hold-out samples on a square root transformed scale. a) Six-fold random cross-validation. b) Three-fold spatial cross-validation with folds indicated by colour.

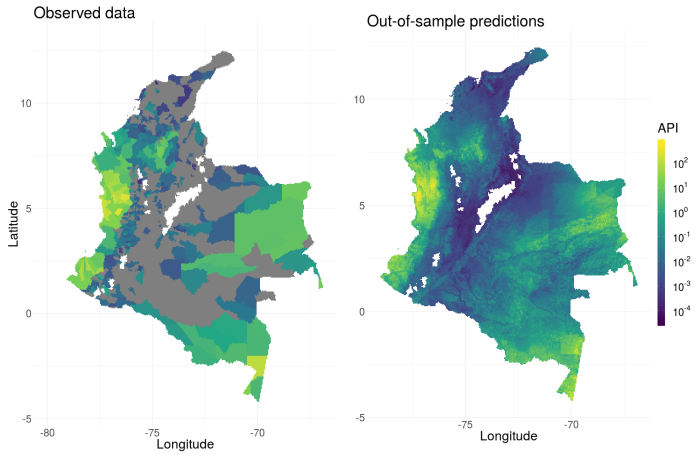


Figure 2: Left: Observed data for Colombia (grey for zero incidence). Right: Out-of-sample predictions for the random cross-validation, machine learning only model. For each cross-validation fold, predictions are made for the held out data which are then combined to make a single surface.

Table 2: Machine learning model results and means of fitted parameters (i.e. model weights) across cross-validation folds of the machine learning predictions only model.

	Madagascar			Colombia		
Model	ML RMSE	Random CV $\bar{\beta}_i$	Spatial CV $\bar{\beta}_i$	ML RMSE	Random CV $\bar{\beta}_i$	Spatial CV $\bar{\beta}_i$
nnet	0.113	0.031	0.025	**0.058**	-0.250	-0.246
RF	**0.100**	0.337	0.350	**0.058**	0.782	0.742
gbm	0.109	**0.450**	**0.402**	0.066	**0.835**	**0.775**
enet	0.116	0.326	0.307	**0.058**	-0.563	-0.369
ppr	0.110	-0.233	-0.204	0.059	0.210	0.166

Overall, our experiments suggest that using predictions from machine learning models trained on prevalence points provides more accurate predictions than using environmental covariates when fitting disaggregation models of malaria incidence. This increased performance comes despite the data being on different scales, the data being measurements of different aspects of malaria transmission and despite the imperfect model we have used to translate between the two scales. However, the reduced model accuracy in the spatial cross-validation schemes, relative to the random cross-validation scheme, highlights that better spatial coverage of data would improve predictions more than the improved model we have suggested.

Due to the low power of typical aggregated incidence datasets, previous analyses using disaggregation regression used a small number of covariates (Sturrock et al., 2014). However, as models such as Random Forest and elastic net can robustly handle high dimensional data, future work could include many more covariates, potentially increasing predictive performance.

While the approach presented here is related to stacking, it differs in that we have not constrained the regression parameters to be positive nor included a sum-to-one constraint i.e. the result is not simply a weighted average of the level zero model predictions. We did not include these constraints because the base models and the meta model are trained on response data on different scales. However, future work could examine whether using a positive constraint on the regression parameters improves performance.

Another area of potential improvement is varying the data used to train the base level learners. Here we only used data from the region of interest. However, the global dataset is much larger than these subsets. Training some base level models on local data and some on the global dataset and then combining predictions from all these models has potential to further improve model performance.

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(1):20.

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(134):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(7572):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(12):e1001142.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.

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*, pages 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(9):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(1):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(1):17.

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(5):1–21.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2017). *caret: Classification and Regression Training*. R package version 6.0-76.

Law, H. C. L., Sejdinovic, D., Cameron, E., Lucas, T. C., Flaxman, S., Battle, K., and Fukumizu, K. (2018). Variational learning on aggregate outputs with Gaussian processes. *arXiv preprint arXiv:1805.08463*.

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(5):479–507.

NASA (2018). Gridded Population of the World (GPW), v4.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgeway et al. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.

Sill, J., Takács, G., Mackey, L., and Lin, D. (2009). Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*.

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):1–28.

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(1):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(8):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(1):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 (Statistics in Society)*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

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(1):68.

Wilson, K. and Wakefield, J. (2017). Pointless continuous spatial surface reconstruction. *arXiv preprint arXiv:1709.09659*.

Wright, M. N. and Ziegler, A. (2015). Ranger: a fast implementation of Random Forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Zou, H. and Hastie, T. (2012). *elasticnet: Elastic-Net for sparse estimation and sparse PCA*. R package version 1.1.