A statistical modelling framework for mapping malaria seasonality

Michele Nguyen\*; Rosalind E. Howes\textsuperscript{1}; Timothy C.D. Lucas\textsuperscript{1}; Katherine E. Battle\textsuperscript{1}; Ewan Cameron\textsuperscript{1}; Harry S. Gibson\textsuperscript{1}; Jennifer Rozier\textsuperscript{1}; Suzanne Keddie\textsuperscript{1}; Emma Collins\textsuperscript{1}; Rohan Arambepola\textsuperscript{1}; Su Yun Kang\textsuperscript{1}; Chantal Hendriks\textsuperscript{1}; Mauricette Andriamananjara Nambinisoa\textsuperscript{2}; Daniel J. Weiss\textsuperscript{1}\textsuperscript{†}; Peter W. Gething\textsuperscript{1}\textsuperscript{†}

\textsuperscript{1} Malaria Atlas Project, Big Data Institute, University of Oxford, Oxford, UK.
\* Corresponding author: M. Nguyen (michele.nguyen@bdi.ox.ac.uk).
\‡ Joint senior authors.
\textsuperscript{2} National Malaria Control Programme, Antananarivo, Madagascar.

Abstract:

Many malaria-endemic areas experience seasonal fluctuations in cases because the mosquito vector’s life cycle is dependent on the environment. While most existing maps of malaria seasonality use fixed thresholds of rainfall, temperature and vegetation indices to find suitable transmission months, we develop a spatiotemporal statistical model for the seasonal patterns derived directly from case data.

A log-linear geostatistical model is used to estimate the monthly proportions of total annual cases and establish a consistent definition of a transmission season. Two-component von Mises distributions are also fitted to identify useful characteristics such as the transmission start and end months, the length of transmission and the associated levels of uncertainty. To provide a picture of “how seasonal” a location is compared to its neighbours, we develop a seasonality index which combines the monthly proportion estimates and existing estimates of annual case incidence. The methodology is illustrated using administrative level data from the Latin America and Caribbean region.

Keywords:
Seasonality; Spatiotemporal Statistics; Geostatistics; Infectious diseases; Malaria

1. Introduction:

Malaria is a disease caused by the \textit{Plasmodium} parasite and remains a major cause of child mortality in sub-Saharan Africa (World Health Organisation 2018). Like that of many other infectious diseases, malaria transmission exhibits seasonality across endemic areas. Understanding location-specific seasonal characteristics is useful for maximising the impact of interventions, developing early warning systems as well as improving models relating indicators of transmission and disease (Stuckey et al. 2014).

To this end, maps of malaria seasonality have been developed. By using thresholds on environmental factors, one can determine the months suitable for transmission (Cairns et al. 2012, Gemperli et al 2006). Since seasonal malaria chemoprevention has been shown to be most effective when delivered over three months, these maps can be useful for targeting such interventions (Cairns et al. 2012).

Despite their functionality, the threshold-based maps have several limitations. Although the environment is a key driver of seasonality, there are other contributors such as migration (Martinez 2018). The same environmental factors could also affect different areas differently: rain, for example, can both create and wash away mosquito breeding sites depending on the local topology and rainfall intensity (Martinez 2018). Using environmental thresholds does not allow for other potential drivers or account for the variation of responses.

Another class of seasonality maps relates to concentration indices. To quantify the distribution of malaria cases in each district over a year, Mabaso et al. (2005) used Markham’s concentration index which was previously used to determine rainfall concentrations. Their concentration maps from the
case numbers estimated using a Bayesian spatiotemporal regression model displayed clearer spatial patterns than those derived from raw case numbers. Spatiotemporal models smooth out idiosyncratic deviations to enable us to focus on the main seasonal trend. They are also useful for relating the seasonality to input covariates and account for unknown spatiotemporal effects.

In this paper, we present a modelling framework for a cohesive and evidence-based analysis of malaria seasonality. Using a spatiotemporal geostatistical model, we obtain maps of various seasonality measures including the number of transmission periods in a year, as well as start and end months of each transmission season. Unlike previous work, we also present the uncertainty associated with each map. A seasonality index from the rainfall literature is adapted to give a visual impression of both the distribution and magnitude of malaria cases over a year. The methodology is illustrated using administrative level data from the Latin America and Caribbean (LAC) region.

2. Methodology:

2.1 Original seasonality index

As suggested by Feng et al. (2013), “how seasonal” a location \( j \) is can be expressed as the product of an entropy measure \((D_j)\) and the relative amplitude \((R_j/R_{\text{max}})\):

\[
S_j = D_j \times \frac{R_j}{R_{\text{max}}},
\]

where \( D_j = \sum_{i=1}^{12} p_{i,j} \log \left( \frac{p_{i,j}}{q_i} \right) \).

Here, \( p_{i,j} \) is the average proportion for month \( i \) and \( q_i = 1/12 \), for \( i = 1, \ldots, 12 \). So, \( D_j \) quantifies how different the monthly proportions are from a uniform distribution over the year. In the context of malaria, \( R_j \) can represent the annual case or parasite incidence (API; the number of cases per 1000 people in a year) at location \( j \) and \( R_{\text{max}} \) the maximum API over the region. Since \( S_j \) separates the timing and amplitude aspects of seasonality and there are existing maps of API in the literature, we can focus on modelling the monthly proportions at each location, \( p_{i,j} \).

By modelling proportions instead of the number of cases as done by Mabaso et al. (2005), we bypass the need to estimate catchment populations. This is useful when working with health facility data, which is becoming increasingly available in multiple countries. In addition, we restrict our analysis to dynamic environmental covariates like temperature while omitting static factors such as elevation.

2.2 Spatiotemporal monthly proportion model

From incidence data, we compute monthly median case counts at each location over a set number of years and obtain the monthly proportions by dividing the medians by their annual sum.

The following spatiotemporal model is used to estimate monthly proportions over our study region:

\[
\log(p_{i,j}) = X_{ij}^T \beta + \phi_{i,j} + \epsilon_{i,j},
\]

where \( X_{ij} \) is a \( m \)-dimensional covariate vector including an intercept, \( \beta \in \mathbb{R}^m \) is the corresponding parameter vector and \( \epsilon \sim N(0, \sigma^2) \) denotes independent, identically distributed noise. The spatiotemporal Gaussian field \( \phi \) is constructed as follows:

\[
\phi_{i,j} = \begin{cases} 
\xi_{i,j} \text{ for } i = 1, \\
\alpha \phi_{i,j-1} + \xi_{i,j} \text{ for } i = 2, \ldots, 12,
\end{cases}
\]
and \( \xi_{i,j} \) correspond to zero-mean Gaussian innovations which are temporally independent but spatially coloured with a Matérn covariance. Note that rescaling of the estimated proportions is required to ensure that they sum to one at each location. This is consistent with the fact that some locations are more or less sensitive to the variation in the underlying factors.

Before fitting the model, we exclude outliers \((\log(p_{i,j}) \leq -11 \text{ for LAC})\) to model prototypical seasonal behaviour. We also compute monthly medians for our covariates over the study years and standardise them. The covariates we consider are rainfall from the Climate Hazards Group Infrared Precipitation and Station data (CHIRPS), enhanced vegetation index (EVI), daytime land surface temperature \((\text{LST}_{\text{Day}})\), diurnal difference in land surface temperature \((\text{LST}_{\text{delta}})\), night-time land surface temperature \((\text{LST}_{\text{night}})\), tasseled cap brightness (TCB), tasseled cap wetness (TCW) as well as the temperature suitability indices for \textit{Plasmodium falciparum} and \textit{Plasmodium vivax} \((\text{TSI}_{\text{Pf}}\) and \(\text{TSI}_{\text{Pv}}\)). The data sources are detailed elsewhere (Kang et al. 2018). To account for delayed and accumulated responses to these environmental variables, we also test them at 1-3 month lags.

We set aside 30% randomly selected sites for validation. Working with the rest of the data, we reduce the set of covariates and account for multicollinearity by iteratively computing the variance inflation factors (VIFs) and removing the covariates with the highest VIF value until all the remaining covariates have VIF values less than 10. Next, we fit the model in R using integrated nested Laplace approximation (INLA) (Lindgren et al. 2011, Kang et al. 2018). Backwards regression using the Deviance Information Criterion (DIC) is used to select the best parsimonious model.

2.3 Deriving seasonality statistics

After checking that the model performs well on both the training and test data in terms of coverage probabilities and root mean squared errors, we refit the chosen model using all of the data. Seasonality statistics are derived for each posterior sample of the location-specific monthly proportions.

We regard a location as potentially seasonal if its entropy \(D_j > 0\). If this is satisfied, we fit a rescaled, two-component von Mises \((\text{R2vM})\) density to the monthly proportions. Rewriting the month in a year as a random variable on a circle, \(\theta = \frac{2\pi i}{12}\) where \(i = 1, ..., 12\), the R2vM function is defined as:

\[
f(\theta; s, \omega, \mu_1, \kappa_1, \mu_2, \kappa_2) = s[\omega f_1(\theta; \mu_1, \kappa_1) + (1-\omega)f_2(\theta; \mu_2, \kappa_2)]
\]

\[
\text{where } f_k(\theta; \mu_k, \kappa_k) = \frac{1}{2\pi I_0(\kappa_k)} \exp(\kappa_k \cos(\theta - \mu_k)),
\]

and \(\mu_k\) and \(\kappa_k\) are the mean and concentration parameters of the \(k^{th}\) component respectively \((k = 1, 2)\). Here, \(I_0\) is the modified Bessel function and \(\omega\) is a probability weight. The scale parameter \(s > 0\) modulates between the continuous density function and monthly proportions over discrete months.

Using a circular distribution provides a continuous curve between the months of January and December. Using a two-component von Mises density, in particular, is convenient for identifying the peaks of the bimodal distribution since these correspond to the mean parameters (Pewsey et al. 2013). Although an arbitrary number of von Mises components can be used, we only use two because areas with seasonal malaria transmission typically have one or two main seasons (Stuckey et al. 2014).

Based on the fitted R2vM curve, we identify the transmission periods by marking the months where the curve is at or above \(1/12\). In this way, we can also obtain the start, end and length of each season. If there is only one transmission period, we fit a rescaled, one component von Mises density to the monthly proportions and estimate the seasonality statistics based on this instead.

To obtain the uncertainty associated with the derived statistics, we summarise the results from 100 posterior samples of the monthly proportions. By looking at the proportion of times a location is deemed bimodal or unimodal, we can obtain the majority decision as well as the degree of certainty.
Based on this, we can analyse the uncertainty in the estimated seasonal characteristics. For the start, end and length of the transmission, we can obtain the means and standard deviations.

2.4. Adjusted seasonality index

The seasonality index introduced in Section 2.1 does not work well for bimodal distributions since the entropy does not account for the two peaks and only considers the overall distribution in a year which typically appears as more even than a unimodal one. To better reflect the degree of seasonality, we adapt the entropy for bimodal distribution at location \( j \) using the fitted R2vM function as follows:

\[
\tilde{D}_j = \omega D_j^{(1)} + (1 - \omega) D_j^{(1)},
\]

where

\[
D_j^{(k)} = \sum_{i=1}^{12} f_k(\theta; \mu_k, \kappa_k) \log_2 \left( \frac{f_k(\theta; \mu_k, \kappa_k)}{q_i} \right),
\]

for \( k = 1,2 \) and the terms are as defined before. For consistency, we also base the adjusted seasonality index for unimodal distributions on the fitted one component von Mises density.

3. Results:

We illustrate our method using case data from the LAC region and restrict our analysis to *Plasmodium vivax* (Pv), the dominant malaria species there. We study the smallest administrative units available over each area and use their centroid coordinates as their point locations.

For the years 2009-2016, we can compute monthly median case counts for 1 ADMIN1 (state) unit and 567 ADMIN2 (municipalities) units in Brazil, 458 ADMIN2 units in Colombia, 21 ADMIN1 units in Venezuela, 1 ADMIN1 units in Panama and the ADMIN0 (national) level for Ecuador, Suriname and Paraguay. The model chosen by backwards regression was refitted to all of the data since both training and test coverage probabilities (90.244% and 90.204%) are close to the target of 95% and the root mean squared errors (0.0211 and 0.0187) are comparably low.

Table 1 shows the parameter estimates for the chosen monthly proportion model. Under a 5% significance level, we observe a positive relation between monthly proportions and EVI but negative relations with CHIRPS, CHIRPS_lag1 and EVI_lag2. The negative relation with rainfall could be explained by the increase in mosquito breeding habits when river recede during the dry season (Valle & Lima 2014) Intense rain could also reduce treatment seeking rates and hence the number of recorded malaria cases.

Figure 1 shows the estimates of the seasonality index which were computed using 2016 Pv API estimates and 100 realisations of the estimated monthly proportions. The strongest seasonality based on the mean index is observed in Colombia, near its borders with Brazil and Peru. It was also found that most of the LAC region experiences only one transmission season per year. As an example, we present the map of the mean start month of the first season and its standard deviation in Figure 2. These are calculated via the circular definitions and transformed back into months via multiplication with \( \frac{12}{2\pi} \) (Pewsey et al. 2013). The mean start months are reasonably contiguous across space as expected. Such maps of seasonality characteristics will be useful for policymakers for planning interventions.
| Term             | Median  | 95% CI       | Term       | Median  | 95% CI       |
|------------------|---------|--------------|------------|---------|--------------|
| Intercept        | -2.205  | (-2.276, -2.133) | LST_delta_lag1 | -0.017  | (-0.116, 0.082) |
| CHIRPS_lag3      | 0.014   | (-0.010, 0.038) | TCB_lag1   | -0.026  | (-0.085, 0.033) |
| CHIRPS_lag1      | -0.065  | (-0.093, -0.036) | EVI        | 0.092   | (0.019, 0.165) |
| TCB_lag3         | 0.014   | (-0.042, 0.071) | EVI_lag2   | -0.124  | (-0.203, -0.044) |
| TSL_Pv_lag3      | -0.070  | (-0.157, 0.017) | obs.var ($\sigma^2$) | 0.188  | (0.173, 0.203) |
| LST_night        | 0.035   | (-0.080, 0.149) | spde.var.nom ($\sigma^2$) | 1.167  | (1.057, 1.288) |
| CHIRPS_lag1      | -0.041  | (-0.070, -0.012) | spde.range.nom ($\kappa$) | 5.350  | (4.954, 5.750) |
| LST_delta_lag3   | 0.094   | (-0.005, 0.193) | AR.rho ($\alpha$) | 0.908  | (0.892, 0.922) |

Table 1: Posterior medians and 95% credible intervals (CIs) for the parameters of the refitted model.

Figure 1: (a) The mean adjusted seasonality index estimates over the LAC region and (b) the standard deviation where square roots have been applied to highlight the spatial heterogeneity. The white regions denote areas with no Pv API data.

Figure 2: (a) Mean start month and (b) associated standard deviation for the first transmission season in the LAC region.
4. Discussion and Conclusion:

Using a spatio-temporal statistical model, we related malaria seasonality to environmental drivers and derived useful information on transmission seasons. Although our focus has been on mapping these, our methodology has other use cases. For example, by multiplying the monthly proportion estimates with the API, we can compute the monthly case or parasite incidence (MPI) at each location without developing a separate model for MPI itself. In addition, the different stages of the methodology can be used separately: if we have previously estimated monthly incidence, we can use the monthly proportion concept and algorithm to derive seasonal statistics.

Centroids of the smallest administrative units were used as point locations in the LAC data analysis. To avoid this approximation, continuous autoregressive (CAR) models can be used in place of the geospatial autoregressive process in our model. However, this has its own drawbacks (Valle and Lima 2014). The correlation between neighbouring units does not depend explicitly on the distance between them which is unnatural when we have units of varying sizes. Furthermore, the relation between malaria incidence and unit-representative covariates is likely to be weaker than at the point level.

The monthly proportion model identifies the dominant relationship between malaria cases and the environment in our study region. A key assumption is that this and the resultant seasonal pattern remain constant at least for the time period in our data. Since this may not be the case with climate change, there is a need to update models and investigate extensions to deal with varying relations.

As more countries adopt the District Health Information Software 2 (DHIS2) for instant recording of cases, using case data to establish seasonality patterns will be increasingly feasible and desirable. Currently, work is being done to apply this methodology to health facility case data from Madagascar.

References:
Cairns, M., Roca-Feltrer, A., Garske, T., Wilson, A. L., Diallo, D. Milligan, P. J., Ghani, A. C. & Greenwood, B. M. (2012), ‘Estimating the potential public impact of seasonal malaria chemoprevention in African children’, Nature communications 3, 881.
Feng, X., Porporato, A. & Rodriguez-Iturbe, I. (2013), ‘Changes in rainfall seasonality in the tropics’, Nature Climate Change 3(9), 811-815.
Gemporler, A., Sogoba, N., Fondjo, E., Mabaso, M., Bagayoko, M., Briët, O. J. T., Anderegg, D., Liebe, J., Smith, T. & Vounatsou, P. (2006), ‘Mapping malaria transmission in West and Central Africa’, Tropical Medicine & International Health 11(7), 1032-1046.
Kang, S. Y., Battle, K. E., Gibson, H. S., Ratsimbosoa, A., Randrianarivelosoa, M., Ramboarina, S., Zimmerman, P. A., Weiss, D. J., Cameron, E., Gething, P. W. & Howes, R. E. (2018), ‘Spatio-temporal mapping of Madagascar’s Malaria Indicator Survey results to assess Plasmodium falciparum endemicity trends between 2011 and 2016’, BMC Medicine 16(1), 71.
Lindgren, F., Rue, H. and Lindström, J. (2011), ‘An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(4), 423-498.
Mabaso, M. L. H., Craig, M., Vounatsou, P. & Smith, T. (2005), ‘Towards empirical description of malaria seasonality in southern Africa: the example of Zimbabwe’, Tropical Medicine & International Health 10(9), 909-918.
Martinez, M. E. (2018), ‘The calendar of epidemics: seasonal cycles of infectious diseases’, PLoS Pathogens 14(11), e1007327.
Pewsey, A., Neuhäusser, M. & Ruxton, G.D. (2013), Circular Statistics in R, Oxford University Press.
Stuckey, E. M., Smith, T. & Chitnis, N. (2014), ‘Seasonally dependent relationships between indicators of malaria transmission and disease provided by mathematical model simulations’, PLoS Computational Biology 10(9), e1003812.
Valle, D. & Lima, J. M. T. (2014), ‘Large-scale drivers of malaria and priority areas for prevention and control in the Brazilian Amazon region using a novel multi-pathogen geospatial model’, Malaria Journal 13(1), 443.
World Health Organization (2018), World Malaria Report 2018, Geneva.