Spatio-Temporal Modelling Spatially Aggregated Data, In case of Malaria in Southern Ethiopia

PHD THESIS

BY
YONAS SHUKE KITAWA

HAWASSA UNIVERSITY

HAWASSA, ETHIOPIA
31st October, 2023

Spatio-Temporal Modelling Spatially Aggregated Data, In case of Malaria in Southern Ethiopia

Ph.D Thesis

BY
Yonas Shuke Kitawa

A thesis submitted to the Department of Statistics at Hawassa University
In Partial Fulfilment of the Requirements For *The Degree of Doctor of Philosophy in Applied Statistics*

31st October, 2023

Approval Sheet I

This is to certify that the thesis titled "***Spatio-Temporal Modelling Spatially Aggregated Data, In case of Malaria in Southern Ethiopia***" submitted in partial fulfilment of the requirement for the degree of Doctor of Philosophy in Applied Statistics to the Department of Statistics, Hawassa University, and is the record of original research carried out by ***Yonas Shuke Kitawa, ID.No: PhdAps/005/10***, under my supervision and no part of the thesis has been submitted for another degree or diploma. The assistance and help received during this investigation have been duly acknowledged. Therefore, I recommended that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

----------	----------	----------
Name of Main Supervisor	Signature	Date
----------	----------	----------
Name of Co-Supervisor	Signature	Date

Approval Sheet II

We, the undersigned, members of the Board of Examiners of the final open defence by Yonas Shuke Kitawa have read and evaluated his thesis titled "***Spatio-Temporal Modelling Spatially Aggregated Data, In case of Malaria in Southern Ethiopia***" and Examined the candidate. This is therefore to certify that the thesis has been accepted in partial fulfilment of the requirement of the degree of Doctor of Philosophy in Applied Statistics.

----------	----------	----------
Name of Department Head	Signature	Date
----------	----------	----------
Name of Main Supervisor	Signature	Date
----------	----------	----------
Name of Co-Supervisor	Signature	Date
----------	----------	----------
Name of Chair Person	Signature	Date
----------	----------	----------
Name of External Examiner	Signature	Date
----------	----------	----------
Name of internal Examiner	Signature	Date

Declaration

This thesis has been submitted to the Department of Statistics at Hawassa University in partial fulfillment of the requirements for a degree of Doctor of Philosophy in Applied Statistics. I hereby declare that this Ph.D. thesis is my original work and has not been submitted to any other institution and anywhere for the award of any academic degree, diploma, or certificate. All sources of materials used for this thesis have been fully acknowledged.

----------	----------	----------
Name of student	Signature	Date

Place: Hawassa University, Hawassa, Ethiopia

ACKNOWLEDGMENTS

Foremost, I acknowledge the most powerful God for giving me the strength, wisdom, courage, and ability to complete this research work.

My deepest gratitude is to my supervisor Zeytu Gashaw (Ph.D.) for his continuous support, encouragement, and motivation. I would like to thank Emanuele Giorgi, Olatunji Johnson, and Diggle, Peter for their initial insight into the areas of spatial statistics and their professional help in the beginning. My sincere thanks also go to Professor Arnoldo Frigessi, for his unreserved support and professional help during difficult times. My special thanks also go to Hawassa University's Department of Statistics for entire support and help.

Finally, I would like to thank all my family and friends who enrich my life and continue to support my work. Special thanks to my parents, my brothers, my sisters, and my wife for her unwavering belief in me.

DEDICATION

I would like to dedicate this thesis to my entire family and my wife without whose support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

List of Papers

This thesis includes an introduction to the Ph.D. thesis and three research papers. It includes background about Spatial and spatial-temporal modelling in spatially aggregated data, with the application to predicting malaria risk at the district level in space and time.

Paper I: Space-time modelling of Monthly malaria incidence for seasonal associated drivers and early epidemic detection in Southern Ethiopia. *Published at Malaria journal; https://doi.org/10.1186/s12936-023-04742-9*

Authors: Yonas Kitawa and Zeytu Asfaw.

Paper 2: Understanding the Importance of Spatial Correlation in Identifying Spatio-temporal Variation of Disease Risk, in the Case of Malaria Risk Mapping in Southern Ethiopia: *Published at the journal of Scientific African, 22, e01926. https://doi.org/10.1016/j.sciaf.2023.e01926.*

Authors: Yonas Kiawa, Olatunji Johnson, Emanuele Giorgi and Zeytu Asfaw.

Paper 3: Multivariate spatiotemporal modeling of aggregated malaria count of the genus *P. falciparum* and *P. vivax*; A Case Study on Malaria Risk Mapping in Southern Ethiopia. *Under Review at the Journal of BMC-Medical Research Methodology*

Authors: Yonas Kitawa and Zeytu Asfaw.

ABSTRACT

Malaria is a major public health concern worldwide, particularly in Sub-Saharan Africa. Ethiopia accounts for 1.7% of cases and 1.5% of global deaths, with seasonal and unstable transmission patterns. The COVID-19 pandemic has increased morbidity, making early detection crucial for mitigating the negative effects of malaria. Thus disease risk mapping is important to identify areas with elevated risks to assist police decisions. However, in such areas, the data sets are mostly available in aggregated form. Spatially aggregated data is often expressed as disease cases or average measurements from districts, often focused on administrative convenience rather than knowledge of the aetiology of the disease. When the underlying results in the process of disease outcomes are thought to be spatially continuous, a spatially continuous model should be discovered rather than typically employed spatially discrete models. This thesis is composed of three papers; in the first paper, We have developed an early warning system for malaria that provides a signal whenever the incidence of malaria exceeds certain thresholds. Then, we identified areas with some elevated risks based on temporal trends. Our findings could assist public health offices in identifying areas for early intervention and guiding healthcare resource allocation, allowing its limited resources to be utilized to the greatest extent possible. In addition, we explored the significance of incorporating environmental variables in forecasting malaria risk in Southern Ethiopia. In the second paper, the researcher examined a spatiotemporal model to highlight disease risk change over time. Four models were considered to be the most appropriate for the problem of spatially aggregated data in a small area. According to the findings of the

study, the spatiotemporal spatially discrete approximation to the log-Gaussian Cox process (SPSDLGCP) gives credible and computationally efficient estimates of disease risk on both spatially continuous and aggregated scales. Furthermore, we have discovered that including spatial correlation is critical when modelling the spatiotemporal variation of disease risk in the region. In the third paper, taking into account areal units as the vertices of a graph and neighbour interactions as a collection of edges, we proposed a graph-based optimization approach that could be used to estimate either a static or a temporally changing neighbourhood matrix to better capture the spatial correlation in the data set. When compared to the commonly utilized border-sharing rule, the strategy yielded better inference. In particular, using a temporal varying waiting matrix for modelling the aggregated count of two dominating *Plasmodium* species; *P. falciparum* and *P. vivax*; a clearer picture of the distribution of the incidence in the region was provided. We have also highlighted regions where either species poses unusual threats. Finally, we identified major climate variables linked with malaria risk in the region, used a spline function to integrate the non-linear relationship between climatic factors and malaria risk, and investigated the delayed climatic impact on malaria risk. From August 2013 to June 2019, we considered aggregated malaria counts in 149 districts located in Southern Ethiopia. The findings of our study have the potential to assist the SNNPRS public health department and other stakeholders in defining areas for early intervention as well as regulating the distribution of limited resources for the healthcare facility to the greatest extent possible. Furthermore, by looking for a more toiled approach for estimating the waiting matrix for multivariate cases, including additional risk factors that were not identified, risk discontinuities, and the problem of zero inflation, one can improve the existing methods.

CHAPTER 1. INTRODUCTION

Over the past few years, malaria incidence and deaths have decreased globally but remain a serious public health issue in several parts of the world (Bhatt et al., 2015; WHO., 2019; Feachem et al., 2019). It affects people from all socioeconomic groups and is a major source of illness and death in numerous countries with low or middle incomes, notably in sub-Saharan Africa. There are around 87 countries and provinces where there is a risk of malaria transmission, which are home to around half of the world's population. For instance, in 2020, malaria is thought to have contributed to an estimated 241 million clinical episodes and 627,000 fatalities (WHO, 2021). African regions account for an estimated 95% of cases and deaths (WHO., 2019) from which Ethiopia accounts for an estimated 1.5% cases and 1.7% of malaria fatalities and cases worldwide (WHO, 2022).

Malaria is highly seasonal and unstable in Ethiopia, with epidemic-prone transmission patterns in various parts of the country. Approximately 52% of the country's population is at risk of this disease (WHO, 2022), which might be associated with the presence of favorable topography and climate for malaria transmission (Zhou et al., 2004; Tigu et al., 2021). The population of all age groups is at risk of infection (Tegegne et al., 2022), which might additionally be associated with low herd immunity.

Furthermore, associated with the COVID-19 pandemic, the morbidity of malaria patients has increased with delayed treatment (Schubert et al., 2021). This poses an increased threat for individuals with severe malaria because, despite the availability of effective medications

(Dondorp et al., 2005), *P. falciparum* malaria can advance to catastrophic results if treatment is delayed (Schubert et al., 2021). As a result, early detection of severe malaria infections is critical to mitigating such effects.

P. falciparum and *P. vivax* are two well-known parasites that cause malaria in Ethiopia, where they are thought to be responsible for 60% and 40% of cases, respectively. Due to the large expansion of malaria detection and treatment as well as the use of vector control techniques, the trend in malaria infections and deaths considerably declined between 2000 and 2015 (Deribew et al., 2017; Taffese et al., 2018). However, studies suggest that the number of malaria cases in some regions of the nation has stabilized or even grown between 2015 and 2018 (WHO., 2019). For instance, an increase in confirmed malaria cases was noted during July and August 2019 in Ethiopia's Southern Nations and Nationalities Regional State compared to the same months in 2018. Altitude, temperature, humidity, and rainfall have all been linked to malaria risk by either *Plasmodium* species (Seyoum et al., 2017; Abeku et al., 2004; Midekisa et al., 2015; Lyon et al., 2017). Age, wealth level, population migration, and proximity to places of mosquito breeding are some of the risk factors for *P. vivax* infection (Chirombo et al., 2020). Where the two epidemics coexist, *P. vivax* infection drops more slowly than that of *P. falciparum* infection (Deribew et al., 2017), and both remain the leading cause of malaria in Ethiopia (Taffese et al., 2018).

Healthcare systems may become overburdened by the development of infectious diseases like malaria, which can have catastrophic effects on people's health, the economy, and society on various levels (Cevik, 2023; Tefera et al., 2020; Hailu et al., 2017). Understanding the dynamics of the disease spread through disease risk modelling is crucial, and it may offer important insights into disease prevention and control Kaye et al. (2021). Each illness has

its own pathogen, set of symptoms, and mode of transmission. Malaria transmission by *Plasmodium* species is not different, and its risk varies across space and time which might be associated with climatic factors in the country (Nigussie et al., 2022).

There are significant health disparities between the communities due to inequality in social, economic, environmental, climatic, and other factors. At the same time, initiatives aimed at enhancing public health and health emergencies may have an impact on temporal patterns. Health inequality is the term used for the spatial variance in illness risk, with developed nations often having lower disease risks than more underprivileged ones (Wilkinson, 1997). According to the World Health Organization (WHO., 2019), health inequalities are a major public health concern that has gained political traction in Ethiopia. By modelling disease count data at the lower administrative level in a small area, such inequalities can be quantified, allowing for the resolution of policy-relevant questions like which regions show heightened risks and rising risk patterns in comparison to their nearby locations; and does a health concern affects vulnerability to illness in different districts equally?

Numerous statistical approaches have been suggested to model spatiotemporal heterogeneity in disease risk, with (Besag, 1974; Besag et al., 1991a; Knorr-Held and Best, 2001; Diggle et al., 1998) being among the most well-liked. Identification of clusters of regions demonstrating heightened risks compared to their nearby locations is a prominent purpose of modelling small-area data, and several methods have been proposed, including (Besag et al., 1991a; Wakefield and Kim, 2013; Knorr-Held and Best, 2001) in a spatial context. On the other hand, (Lee and Lawson, 2016; Lee et al., 2018) expanded this by applying it to a spatiotemporal context, while for predicting geographically restricted temporal patterns, several models have been proposed (e.g. (MacNab and Dean, 2001; Midekisa et al., 2012)).

When making predictions for a small geographical area, data sets are typically available in the aggregated form with sub-geographic districts or areas of interest. This is typically done to protect patient confidentiality and address ethical issues with data use. They are mostly reported as either many cases or average measurements derived from the subdivision of the study region into administrative districts. The subdivision is typically based on administrative effectiveness as opposed to knowledge of the aetiology of a particular disease, of course. Administrative convenience is often used as the basis for partitioning rather than knowledge of the aetiology of any given disease. For instance, in Ethiopia, several geographic units such as kebeles, districts, zones, or regions are utilized to provide health data. Depending on the specific requirements or interests of the community, such divisions are employed to give information.

Malaria risk maps have been produced and are being utilized in the development of a strategy, focusing on, measurement, tracking, and campaigning procedures for making decisions. To attain the 2030 targets, the High Burden High Impact programs and the Global Malaria Initiatives 2016-2030 emphasized the significance of improving tailored approaches to malaria control, including in areas with high incidence. However, special attention is appropriate in regions like SNNPRS which consists of districts with diverse risks but not well investigated. In contrast, various maps were produced in the country's northern regions, particularly in the Amhara regional state, where the risk is still very high (Nigussie et al., 2022; McMahon et al., 2021; Tegegne et al., 2022). However, there is insufficient research in the SNNPRS that identifies spatiotemporal variation in malaria risk. Few of the articles (Dabaro et al., 2021; Aliyo et al., 2023) were specifically focused on certain districts or health posts and did not provide an overall picture of the incidents in the region. To begin, we need to ensure regional coverage to balanced against the district level to have an accurate picture of disease

risk. This helps in identifying 1) regions where a mix of further and creative actions may be acceptable, and 2) given the relatively limited resources, malaria control must be prioritized in specific districts or through the use of specialized strategies.

Thus, the produced malaria risk maps might be utilized to assist in malaria decision-making across districts, zones, cities/towns, regions, and countries. Furthermore, it is critical to develop maps using data and approaches that policymakers understand, are confident in, and are willing to use. We hope, the method we have considered here is very helpful to monitor malaria risk in the area and could be accessed in the form of shiny maps or visualization tools. As Newman stated, "While the supply of research information is important, it is likely to be utilized to inform policy if it is accessed, valued, and understood by policymakers" (Newman et al., 2012).

Such spatially or spatiotemporally aggregated data sets can potentially be modelled through several methods, starting with small area estimation. When constructing a standard spatial or spatiotemporal model, it is assumed that the result depends on a number of both explained and unaccounted-for components. As a consequence, the outcome was modelled as a combination of factors that were explained and those that were not. The explained component consists of observable quantities, while the unexplained component can be described as an unknown stochastic process that changes in space or time. The unexplained part could potentially be modelled as either a spatially continuous (Diggle et al., 2007; Diggle and Giorgi, 2019; Diggle et al., 2013) or a spatially discrete variation (Besag et al., 1991b; Leroux et al., 2000), or by linking the two (Lindgren et al., 2011; Simpson et al., 2012; Johnson et al., 2019).

Furthermore, when used for malaria risk mapping in Southern Ethiopia, it is uncertain which

of these methodologies will best provide the estimation of the risk, help to identify elevated risks and predict future trends. As a result, the thesis's overarching goal is to develop spatiotemporal models to assist in identifying spatiotemporal variation of malaria risk in SNNPRS, with a particular emphasis on:

- Developing district-level malaria early warning system to forecast malaria risk in each district (this could be done to get a clearer picture regarding the seasonality of malaria risk in each district rather than considering common seasonality trends in the region)
- To point out the importance of including spatial correlation while determining the spatiotemporal variation of the disease risk (to understand how malaria risk distributions are affecting people from geographically connected districts through spatiotemporal modelling)
- Looking at neighbourhood while determining spatial correlation in cases of identifying malaria risk by multiple *plasmodium* species (Multivariate spatio-temporal modelling to detect malaria risk by either species)

1.1 Statistical Model for Spatially Aggregated Data In small area

A subset of geographical data known as spatially aggregated data consists of observations that are connected to a set of K nearby but separate areal units, like districts or kebeles, zones, or regions. Each observation pertains to a whole area, therefore measurements are often summaries, like the average of incidences in the area. Such data sets have recently grown more and more accessible as a result of the advancement of computers and the establishment of databases like Surveillance Epidemiology in settings with low resources. Despite some of the shortcomings, these databases offer information on a collection of K areal units over T subsequent times, resulting in a rectangle array of $K \times T$ spatiotemporal observations.

Several methodologies have been developed to model such data sets in a small area aimed at quantifying the degree of uncertainty (Lee and Lawson, 2016)), predicting the impact of a risk factor on the outcome (Wakefield, 2007), and detecting regions of adjacent geographic areas with elevated disease risk (Lawson, 2018). Proper handling of spatial and temporal autocorrelation in which data from nearby space-time units have more similar values than those from a distance apart (Tobler, 1970) is important. Temporal autocorrelation arises when the majority of the data is about the same people throughout study time, as opposed to spatial autocorrelation, which can be caused by neighbourhood, clustering, and unknown confounding effects. A number of approaches have been presented for this inquiry in the small area, but the optimal model relies on a scientific understanding of the data-generating process and computational challenges, as it was detailed below.

1.1.1 Conditional Auto-regressive (CAR) Model

After being established by (Besag, 1974), conditionally autoregressive (CAR) approaches have seen a sharp rise in usage over the last few years. This rise in popularity is due to their practical application in the Markov chain Monte Carlo (MCMC) techniques for fitting particular types of spatial models. Given data gathered over a range of geographic units, such as census tracts, counties, states, or districts, CAR models are ideally suited for estimating disease incidence and death rates. The CAR model can be applied as a likelihood for areal data, a probability model for illness risk that cannot be observed, or a probability model for values that were not completely observed. The Conditional Auto-Regressive (CAR) model, which evaluates the temporal dependency between the various realizations in addition to the geographical dependence between sites, is required for the analysis of data gathered in such spatial location and in a time interval. For a non-overlapping discreet set of areal units $k = 1, \ldots, K$; disease count Y_{it} are sometimes measured for each district at $t = 1, \ldots, T$ consecutive periods. Conditionally on a zero-mean Gaussian process $S = (S_1, \ldots, S_t)$, Y_{it} is mutually independent Poisson random variables ($Y_{it} \sim m_{it} * Poisson(\lambda_{it})$) with expectations:

$$log(\lambda_{it}) = d_{it}^T \beta + S_{it} \tag{1.1}$$

where m_{it} is the total population count at each district used as an offset otherwise, disease risks are spatially homogenous and β is a vector of regression coefficients associated with spatiotemporal referenced factors; d_{it}^T. The accuracy matrix for S_{it}, which is Gaussian is then specified to generate spatially discrete models. Symmetric $K \times K$ non-negative adjacency matrix $W = (w_{ij})$ controls spatial autocorrelation, where w_{ij} reflects the spatial proximity between areal units (S_{it}, S_{jt})). W is commonly considered to be binary, with $w_{kj} = 1$ if areal units (S_{it}, S_{jt})) have a common boundary (i.e., are geographically near) and zero otherwise.

Although this binary W specification based on sharing a boundary is the most widely used for such data, there are also additional ways for specifying such a waiting matrix (Bivand et al., 2008). The precision matrix for the Gaussian process S_{it} is then specified for developing spatially discrete models. Several specifications of Eq. 1.1 were present based on the number of response variables considered for the analysis as well as ways of undertaking S_{it}, see (Lee et al., 2018) for further understanding of the alternatives.

1.1.2 Spatio-temporal Geostatistics

Geostatistics is a statistical model and technique for assessing data that is spatially discrete yet describes unobserved spatially continuous phenomena. Model-based geostatistics (MBG) is a branch of spatial statistics that offers techniques for inference on a continuous surface utilizing spatially discrete data (Diggle et al., 1998, 2013). The approach has been frequently applied more and more in disease-mapping purposes (Diggle and Giorgi, 2016; Johnson et al., 2019), with a special emphasis on regions with limited resources where illness registries are either geographically absent or incomplete. Suppose Y_{it} represents the number of disease counts observed from potential locations $i = 1, \ldots, K$ and time $t = 1, \ldots, T$. Conditionally on a spatiotemporal process S_{it} and unstructured random effects Z_{it}, the outcomes Y_{it} are mutually independent Poisson distributed variables with expectation λ_{it} given as:

$$log(\lambda_{it}) = d_{it}^{T}\beta + S_{it} + Z_{it} \tag{1.2}$$

where d_{it}^{T} is a vector of location and time-specific covariates such as climatic factors, and β is the regression coefficients for these covariates. S_{it} is a spatiotemporal random effect used to capture the spatiotemporal correlation between districts, while the Z_{it}; nugget effect are taken to be uncorrelated zero-mean Gaussian variables with variance τ^2. The nugget

effect represents the unstructured residual variation which can either be a small-range spatiotemporal variation, an excess of zero counts, or behavioural variation between co-located districts. The goal of most geostatistical analyses is to predict S_{it} at an unobserved location. The spatiotemporal random effects S_{it}, can be expressed as a stationary and isotropic Gaussian process with zero means, variance σ^2 and correlation function given as:

$$corr(S_{it}, S_{it}^{'}) = \rho(u, v, \phi)$$

where u is the Euclidean distance between points x and $x^{'}$, v is the difference between successive time points $t, t^{'}$, and ϕ is the scale of spatialtemporal correlation.

1.1.3 Log Gaussian Cox Process

(Cox, 1995) spatiotemporal Cox processes are point processes having a random density of points in space and time. Given that it is a non-negative valued stochastic process with intensity λ_{it}, the process is an inhomogeneous Poisson process with λ_{it}. Log-Gaussian Cox process (LGCP) is also a Cox process for which an exponentially transformed Gaussian process is used to represent an underlying intensity field $\lambda_i t$ of positive real values throughout the full domain $R_i t$. This constrains λ_{it} to be positive. The intensity field is then utilized to parameterize a Poisson point process, which is a stochastic method for locating points in space and time. If we assume that $log(\lambda_{it}) = S_{it}$ is a Gaussian process, we get the log-Gaussian Cox process (LGCP); see (Cox, 1995; Moller et al., 1998) for additional information.

(Diggle et al., 2013) developed spatial and spatiotemporal model for aggregated count; conditional on S_{it}, the observed realization Y_{it}; are mutually independent Poisson variables with expectation:

$$\int_{R_i} m_{it} exp(d_{it}\beta + S_{it}) dx \tag{1.3}$$

LGCP offers a way to overcome the difficulty of combining information at many spatial scales since it is not restricted to any particular division of the geographical area of study. The necessity to infer the unobserved locations for each of the reported episodes inside each area, however, significantly increases the computation effort.

Motivated by (Diggle et al., 2013), (Johnson et al., 2019) introduced the spatially discrete approximation to the Log-Gaussian Cox process (SDALGCP) as a technique for studying spatially aggregated data. In Chapter 3, we'll go into more detail about the SDALGCP model for spatiotemporal aggregated outcome data.

1.2 Spatio-temopral Exploratory Analysis

Similar to other statistical analyses, the important starting point for any spatial study is exploratory spatial data analysis (ESDA). ESDA approaches are concerned with analyzing spatial information for spatial autocorrelation and heterogeneity.

1.2.1 Variogram

In geostatistical analysis, a variogram is used to characterize the degree of geographical dependence of a stochastic process S_{it}. The theoretical semivariogram for the process $U_{it} = S_{it} + Z_{it}$ is defined as follows:

$$\begin{aligned}\gamma((x,t),(x,t)^{'}) &= \frac{1}{2}Var\{U(x,t) - U(x^{'},t^{'})\} \\ &= \frac{1}{2}E[\{U(x,t) - U(x^{'},t^{'})\}^2] \\ &= \tau^2 + \sigma^2(1 - \rho(u,v,\phi))\end{aligned}$$

Where τ^2 is the nugget effect, the sill denotes the parameter σ^2, and the practical range $3 * \phi$ denotes the distance u at which the correlation function $\rho(u, v)$ decays to 0.05. In practice, after fitting a non-spatial model to the data, the empirical variogram is employed to assess for the presence of residual spatial correlation in the residuals. The empirical variogram is useful for describing geographical dependence and estimating the correlation structure of the underlying process.

1.2.2 Moran's I Statistics

Moran's I and Geary's C (Moran, 1950) are statistics used to quantify the degree of spatial linkage between areal units. They are also areal unit counterparts of the empirical estimates for the correlation function and variogram, respectively. Moran's I take the form:

$$I = \frac{k \sum_k^K \sum_r^K w_{kr}(y_k - \bar{y})(y_r - \bar{y})}{(\sum_{r \neq k} w_{kr})(\sum_k^K (y_k - \bar{y})^2)}$$

Geary's C takes the form:

$$C = \frac{(k-1) \sum_k^K \sum_r^K w_{ij}(y_k - y_r)^2}{(\sum_{r \neq k} w_{kr})(\sum_k^K (y_k - \bar{y})^2}$$

C is never negative, and has a mean of 1 for the null model; low values (i.e., between 0 and 1) indicate positive spatial correlation.

1.3 Mapping of disease risk

Maps are crucial for comprehending and visualizing the spatiotemporal distribution of disease risk. In particular, maps support monitoring and surveillance of infectious illnesses when the location of persons is linked to disease and commonly serve as a basis for the development and execution of preventative and control strategies. Maps showing the spatiotemporal

illness distribution using region-specific summary measurements are essential components of an EWS (Abeku et al., 2004; Diggle and Giorgi, 2019; Besag, 1974).

A subset of spatial statistics called model-based geostatistics (MBG) enables us to describe geographic variation in disease risk while accounting for diverse sources of uncertainty.

1.3.1 Early warning system (EWS)

An infectious disease early warning system (EWS) is a systematic approach for identifying, monitoring and forecasting illness risk early. The goal is to empower individuals, groups, neighborhoods, and other organizations to take preventative actions to limit the health and socioeconomic impacts of a disease epidemic (Midekisa et al., 2012). An infectious disease EWS comprises risk identification, warning dissemination and communication, and responsibility allocation (Colborn et al., 2018).

1.3.2 Exceedance Probability

Although they typically do not convey much information, especially when the focus is on providing information on the degree of uncertainty, portraying a standard error map is a more natural way to represent measurement uncertainty (Colborn et al., 2018; Giorgi et al., 2018). When making decisions, for example, the goal is to consistently identify locations when the risk of illness is more than or less than a relevant policy threshold. The exceedance probability (EP) map is a more appropriate technique to represent significant uncertainty in this context. Let $\hat{\lambda}_{it}$ be the projected disease risk at i^{th} location at time t, and the EP expression is:

$$Pr(\hat{\lambda}_{it} > c/data)$$

where c is a predefined limit. Generally speaking, EP values close to one signal that disease risk is most likely to be higher than c, whereas EP values close to zero imply the opposite. Finally, EP values around 0.5 indicate the most ambiguous scenario, with disease risk equally likely to be at or below c.

1.3.3 Spatiotemporal smoothing

Understanding the geographical and temporal spread of disease is required for the formulation of an immediate and well-informed EWS. This requires spatiotemporal data at small spatiotemporal scales. The small number of participants is one of the challenges of data sets at smaller spatiotemporal scales, particularly during the early phases of an outbreak. Because of the scarcity of incidents, risk assessments may become erroneous due to excessive sample heterogeneity, resulting in a large range of misunderstandings of disease risk. Accounting for spatiotemporal dependency and heterogeneity in the occurrences might sometimes assist in addressing this challenge (Corpas-Burgos and Martinez-Beneito, 2021).

Borrowing strength from nearby neighbours over distance or time underpins spatiotemporal smoothing of disease risk. This is connected to the (Tobler, 1970) law of geography, stating that "neighbouring districts are more similar to one another than places separated by greater distances". It is an important consideration in small area estimate (SAE) (Handayani et al., 2018). When information from nearby districts is provided, both the information about the districts and between districts is increased. This additional information often lowers noise and extreme values. As a consequence, a smoothed risk map reflects the distribution of disease risk with more precision while decreasing bias (Handayani et al., 2018).

However, over the past few decades, many statistical models have been developed to predict

such risk with better precision while maintaining geographical resolution (Lee et al., 2018; Waller and Carlin, 2010; Martínez-Beneito and Botella-Rocamora, 2019). Such models in the small area usually include random effects using conditional autoregressive (CAR) priors (Besag, 1974; Wakefield, 2007; Leroux et al., 2000). In those modelling approaches, the geographical area of interest is divided into non-overlapping administrative units like districts, where the observed aggregated disease count is available for each geographical unit. Due to the higher variability of disease risk in each district in different periods, it is very important to consider models that smooth the spatial risk surface. However, the disease risks are spatially continuous (Diggle et al., 1998) for which over-smoothing the disease risk by borrowing information from nearby districts sometimes may limit the identification of high-risk areas. Sometimes, due to over-smoothing in discrete models and lack of significance of spatial correlation for districts located distant apart, modelling districts individually can bring gain in efficiency.

1.4 Structure of the thesis

This thesis addresses some problems with spatially aggregated data, with an emphasis on Southern Ethiopia's malaria risk mapping. It is composed of three papers.

- In the first paper, we discuss the problem of developing an early warning system to observe the distribution of incidence at various districts and forecast therein to get a preliminary understanding of the illness trend in the area (MacNab and Dean, 2001; Midekisa et al., 2012). Here, we focus on developing models for each district and examining how seasonality and temporal trends vary between districts to determine the variability of the incidences across spaces and time.

- In the second paper, we talk about how crucial it is to account for spatial correction when identifying spatiotemporal variation in malaria risk. As it's generally accepted that spatially aggregated data use models that presuppose a spatially discrete process, this is the case. Nonetheless, we contend that a spatially continuous model should be taken into account where there is a scientific basis for doing so, particularly when the disease outcome's underlying data generation mechanism is known to operate in a spatially continuous manner. In light of these specifications, we compare various models for spatially aggregated data. Here, we look forward to the CAR model from small area estimation (SAE) as suggested by (Rushworth et al., 2017) and another continuous model by (Diggle and Giorgi, 2019) suggestion on prediction by Discrete and Continuous models in various settings. The performance of predictions is then used to compare the models.

- In the third paper, we attempted to map malaria by various parasite species by multivariate spatiotemporal as specified by (Lee et al., 2022) and taking the idea of specification of waiting matrix from (Lee et al., 2021) and (Jack et al., 2019) using graph-based optimization for neighbourhood estimation. We use the aggregated count data collected at the district level from southern Ethiopia to illustrate each paper. Here, we undertook an application of graph-based optimization for neighbourhood estimation by (Lee et al., 2021) to model malaria by multiple parasite species through multivariate spatiotemporal.

We illustrate all the papers using a data set of the aggregated count at the district level from southern Ethiopia.

1.5 References

Abeku, T., Vlas, S. D., Borsboom, G., Tadege, A., Gebreyesus, Y., Gebreyohannes, H., Alamirew, D., Seifu, A., Nagelkerke, N., and Habbema, J. (2004). Effects of meteorological factors on epidemic malaria in ethiopia: a statistical modelling approach based on theoretical reasoning. *Parasitology*, 128(Pt 6):585–593.

Aliyo, A., Golicha, W., and Fikrie, A. (2023). Pastoral community malaria prevention practice and associated factors among households in three districts of the borena zone, southern ethiopia. *Health Services Research and Managerial Epidemiology*, 10:23333928221144555.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J., York, J., and Mollie, A. (1991a). Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*, 43:1–20.

Besag, J., York, J., and Mollié, A. (1991b). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

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):207–211.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., and Pebesma, E. J. (2008). *Applied spatial data analysis with R*, volume 747248717. Springer.

Cevik, S. (2023). Peering through the fog of uncertainty: Out-of-sample forecasts of post-pandemic tourism. Technical report, International Monetary Fund.

Chirombo, J., Ceccato, P., Lowe, R., Terlouw, D. J., Thomson, M. C., Gumbo, A., Diggle, P. J., and Read, J. M. (2020). Childhood malaria case incidence in malawi between 2004 and 2017: spatio-temporal modelling of climate and non-climate factors. *Malaria journal*, 19(1):1–13.

Colborn, K. L., Mueller, I., and Speed, T. P. (2018). Joint modeling of mixed plasmodium species infections using a bivariate poisson lognormal model. *The American journal of tropical medicine and hygiene*, 98(1):71.

Corpas-Burgos, F. and Martinez-Beneito, M. A. (2021). An autoregressive disease mapping model for spatio-temporal forecasting. *Mathematics*, 9(4):384.

Cox, D. (1995). Some statistical methods connected with series of events. *Royal Statistical Society*, 17(2):129–164.

Dabaro, D., Birhanu, Z., Negash, A., Hawaria, D., and Yewhalaw, D. (2021). Effects of rainfall, temperature and topography on malaria incidence in elimination targeted district of ethiopia. *Malaria journal*, 20(1):1–10.

Deribew, A., Dejene, T., Kebede, B., Tessema, G. A., Melaku, Y. A., Misganaw, A., Gebre, T., Hailu, A., Biadgilign, S., Amberbir, A., et al. (2017). Incidence, prevalence and mortality rates of malaria in ethiopia from 1990 to 2015: analysis of the global burden of diseases 2015. *Malaria journal*, 16(1):1–7.

Diggle, P. and Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *American Statistical Association*, 111(515):1096–1120.

Diggle, P. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman & amp; Hall/CRC Interdisciplinary Statistics. Chapman and Hall/CRC Press.

Diggle, P., Ribeiro, P., and Geostatistics, M.-b. (2007). Springer series in statistics. *Springer. Djelouah K., Frasheri D., Valentini F., D'Onghia AM and Digiaro M.(2014). Direct tissue blot immunoassay for detection of Xylella fastidiosa in olive trees. Phytopathologia Mediterranea*, 53(3):559–564.

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Applied Statistics*, 47(3):299–350.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Dondorp, A., Nosten, F., Stepniewska, K., Day, N., White, N., et al. (2005). Artesunate versus quinine for treatment of severe falciparum malaria: a randomised trial. *Lancet (London, England)*, 366(9487):717–725.

Feachem, R. G., Chen, I., Akbari, O., Bertozzi-Villa, A., Bhatt, S., Binka, F., Boni, M. F., Buckee, C., Dieleman, J., Dondorp, A., et al. (2019). Malaria eradication within a generation: ambitious, achievable, and necessary. *The Lancet*, 394(10203):1056–1112.

Giorgi, E., P. Diggle, R. W. S., and Noor, A. M. (2018). Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys. *International Statistical Review*, 86(3):571–597.

Hailu, A., Lindtjørn, B., Deressa, W., Gari, T., Loha, E., and Robberstad, B. (2017). Economic burden of malaria and predictors of cost variability to rural households in south-central ethiopia. *PLoS One*, 12(10):e0185315.

Handayani, D., Folmer, H., Kurnia, A., and Notodiputro, K. A. (2018). The spatial empirical bayes predictor of the small area mean for a lognormal variable of interest and spatially correlated random effects. *Empirical Economics*, 55:147–167.

Jack, E., Lee, D., and Dean, N. (2019). Estimating the changing nature of scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3):1061–1080.

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

Kaye, A. D., Okeagu, C. N., Pham, A. D., Silva, R. A., Hurley, J. J., Arron, B. L., Sarfraz, N., Lee, H. N., Ghali, G. E., Gamble, J. W., et al. (2021). Economic impact of covid-19 pandemic on healthcare facilities and systems: International perspectives. *Best Practice & Research Clinical Anaesthesiology*, 35(3):293–306.

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85.

Lawson, A. B. (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.

Lee, D. and Lawson, A. (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in glasgow. *The annals of applied statistics*, 10(3):1427.

Lee, D., Meeks, K., and Pettersson, W. (2021). Improved inference for areal unit count data using graph-based optimisation. *Statistics and Computing*, 31(4):1–17.

Lee, D., Robertson, C., and Marques, D. (2022). Quantifying the small-area spatio-temporal dynamics of the covid-19 pandemic in scotland during a period with limited testing capacity. *Spatial statistics*, 49:100508.

Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the CARBayesST package. *Journal of Statistical Software*, 84(9):1–39–350.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Statistical models in epidemiology, the environment, and clinical trials, chapter estimation of disease rates in small areas: A new mixed model for spatial dependence. *Springer-Verlag, New York*, pages 179–191.

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.

Lyon, B., Dinku, T., Raman, A., and Thomson, M. C. (2017). Temperature suitability for malaria climbing the ethiopian highlands. *Environmental Research Letters*, 12(6):064015.

MacNab, Y. C. and Dean, C. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57(3):949–956.

Martínez-Beneito, M. A. and Botella-Rocamora, P. (2019). *Disease mapping: from foundations to multidimensional modeling*. CRC Press.

McMahon, A., Mihretie, A., Ahmed, A. A., Lake, M., Awoke, W., and Wimberly, M. C. (2021). Remote sensing of environmental risk factors for malaria in different geographic contexts. *International journal of health geographics*, 20(1):1–15.

Midekisa, A., Beyene, B., Mihretie, A., Bayabil, E., and Wimberly, M. C. (2015). Seasonal associations of climatic drivers and malaria in the highlands of ethiopia. *Parasites & vectors*, 8(1):1–11.

Midekisa, A., Senay, G., Henebry, G., Semuniguse, P., Wimberly, and Michael, C. (2012). Remote sensing-based time series models for malaria early warning in the highlands of ethiopia. *Malar J*, 11(165).

Moller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scand J Stat*, 25(3):451–482.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Newman, K., Fisher, C., and Shaxson, L. (2012). Stimulating demand for research evidence: What role for capacity-building? *IDS Bulletin*, 43(5):17–24.

Nigussie, T. Z., Zewotir, T. T., and Muluneh, E. K. (2022). Detection of temporal, spatial and spatiotemporal clustering of malaria incidence in northwest ethiopia, 2012–2020. *Scientific reports*, 12(1):3635.

Rushworth, A., Lee, D., and Sarran, C. (2017). An adaptive spatio-temporal smoothing model for estimating trends and step changes in disease risk. *Royal Statistical Society C*, 66:141–157.

Schubert, L., Thurnher, P. M. M., Machold, P. K., Tobudic, P. S., and Winkler, P. S. (2021). Pandemic-related delay of falciparum malaria diagnosis in a traveller leading to cerebral malaria. *Journal of Travel Medicine*, 28(8):taab159.

Seyoum, D., Yewhalaw, D., Duchateau, L., Brandt, P., Rosas-Aguirre, A., and Speybroeck, N. (2017). Household level spatio-temporal analysis of plasmodium falciparum and plasmodium vivax malaria in ethiopia. *Parasites & vectors*, 10(1):1–11.

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29.

Taffese, H., Hemming-Schroeder, E., Koepfli, C., Tesfaye, G., Lee, M., Kazura, J., G.Y, Y., and Zhou, G. (2018). Malaria epidemiology and interventions in ethiopia from 2001 to 2016. *Infect Dis Poverty*, 7(103).

Tefera, D. R., Sinkie, S. O., and Daka, D. W. (2020). Economic burden of malaria and associated factors among rural households in chewaka district, western ethiopia. *ClinicoEconomics and Outcomes Research*, pages 141–152.

Tegegne, E., Alemu Gelaye, K., Dessie, A., Shimelash, A., Asmare, B., Deml, Y. A., Lamore, Y., Temesgen, T., Demissie, B., and Teym, A. (2022). Spatio-temporal variation of malaria incidence and risk factors in west gojjam zone, northwest ethiopia. *Environmental Health Insights*, 16:11786302221095702.

Tigu, F., Gebremaryam, T., and Desalegn, A. (2021). Seasonal profile and five-year trend analysis of malaria prevalence in maygaba health center, welkait district, north-west ethiopia. *Journal of Parasitology Research*, 2021:1–7.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–40.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–83.

Wakefield, J. and Kim, A. (2013). A bayesian model for cluster detection. *Biostatistics*, 14(4):752–765.

Waller, L. A. and Carlin, B. P. (2010). Disease mapping. *Chapman & Hall/CRC handbooks of modern statistical methods*, 2010:217.

WHO. (2019). World malaria report 2019. *Geneva: World Health Organization.*

WHO (2021). World malaria report 2021.

WHO (2022). *World malaria report 2022: World Health Organization and others.* World Health Organization.

Wilkinson, R. G. (1997). Socioeconomic inequalities in morbidity and mortality in western europe. *The Lancet*, 350(9076):516–517.

Zhou, G., Minakawa, N., Githeko, A. K., and Yan, G. (2004). Association between climate variability and malaria epidemics in the east african highlands. *Proceedings of the National Academy of Sciences*, 101(8):2375–2380.

Paper I

Space-time modelling of Monthly malaria incidence for seasonal associated drivers and early epidemic detection in Southern Ethiopia

Yonas Shuke Kitawa[1*]: Department of Statistics, College of Natural and Computational Science, Hawassa University, Hawassa, Ethiopia

Zeytu Gashaw Asfaw[2*]: Department of Bio-statistics and Epidemiology, School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia,

CHAPTER 2. Space-time modelling of Monthly malaria incidence for seasonal associated drivers and early epidemic detection in Southern Ethiopia

Abstract

Background: Although Ethiopia has made great strides in recent years to reduce the threat of malaria, the disease remains a significant issue in most districts of the country. It constantly disappears in parts of the areas before reappearing in others with erratic transmission rates. Thus, developing a malaria epidemic early warning system is important to support the prevention and control of the incidence.
Methods: Space-time malaria risk mapping is essential to monitor and evaluate priority zones, refocus intervention, and enable planning for future health targets. From August 2013 to May 2019, the researcher considered an aggregated count of genus *Plasmodium falciparum* from 149 districts in Southern Ethiopia. Afterwards, a malaria epidemic early warning system was developed using a model-based approach, which helped to chart the disease's spread and future management.
Results: Risk factors like precipitation, temperature, humidity, Enhanced vegetation index, distance from coastal area and nighttime light are significantly associated with malaria with different rates across the districts. Districts in the southwest, including Selamago, Bero, and Hamer, had higher rates of malaria risk, whereas in the south and centre like Arbaminch and Hawassa had moderate rates. The distribution is inconsistent and varies across time and space with the seasons.
Conclusion: Despite the importance of spatial correlation in disease risk mapping, it may occasionally be a good idea to generate epidemic early warning independently in each district to get a quick picture of disease risk. A system like this is essential for spotting numerous inconsistencies in lower administrative levels early enough to take corrective action before outbreaks arise.

Keywords: Malaria Epidemic Early Warning (MEWS), spatial time series, Malaria, Disease mapping, Monte Carlo maximum likelihood, *P. falciparum*

2.1 Introduction

The risk of malaria has considerably decreased during the past few years in various parts of the world. Notwithstanding the recent signal of re-emergency (WHO, 2022), several regions with moderate to high prevalence are successful in reducing the burden of malaria. The development of drug-resistant parasites, the urgency of other pandemics, and the deterioration of programmes designed for malaria control, as stated by (Cohen et al., 2010), have all been factors in the recent comeback of malaria. The COVID-19 pandemic re-emergency had a major impact on the world's population in general and Africa in particular. There is an urgent need to speed up national eradication efforts to meet the 2030 countries' elimination target (WHO, 2015) by concentrating on high-burden areas and considering places with a signal of an outbreak.

Since 1958 in Ethiopia, severe serious malaria epidemics have occurred for approximately 5 to 8 years in most lowland and some highland areas up to 2,500 metres of elevation (Negash et al., 2005). Meteorological, environmental, and socioeconomic factors like; rainfall, temperature, humidity, and others are associated with such risk (Nigussie et al., 2022; Teklehaimanot et al., 2004). There is evidence of malaria resurgence from district to district and over time, as described in various literature, including (Tessema et al., 2020; Taffese et al., 2018; Nigussie et al., 2022) and (Deress and Girma, 2019). Around 60 % of the population and 75 % of the country's territory are at risk of malaria, with *Plasmodium falciparum* accounting for roughly 65–75% of all cases that have been documented (WHO, 2022; Girum et al., 2021). With an unstable seasonal transmission of incidence occurring from September to mid-December, immediately after the main rainy season, and a minor transmission season occurring between March and May (Seyoum et al., 2017; FMHE, 2020), the country

is thought to have low to moderate malaria transmission intensity. Within a specific geographic area, the transmission is seasonal and changes over space and time, according to (Abeku et al., 2004). This may be related to climate changes that are favourable to parasite development that significantly impacted malaria transmission. Disease risk mapping is therefore useful to identify districts with increased risks (Diggle and Giorgi, 2019), as a population of all age groups is at risk with an estimated prevalence of 1.3%.

The COVID-19 (WHO, 2022) emergency and other health care needs which are vying for scarce resources and various political instability in Ethiopia, leads to the re-establishment of the system, regardless of whether the illness risk is declining. To enhance public health decision-making for the monitoring and prevention of malaria epidemics, it is important to build an effective malaria epidemic early warning system. By prioritizing prone locations and times that are most at risk, such a system helps with public health decisions (Harris et al., 2020). On the other hand, using those approaches to cluster districts also supports determining the seasonality of the risk rather than using commonly specified patterns throughout the country (Midekisa et al., 2012, 2015).

To predict disease risks and identify regions that demonstrate atypical outbreaks, many researchers have established malaria epidemic early wake-up calls (Midekisa et al., 2012; Colborn et al., 2018; Organization, 2001). Such tools are designed to identify at-risk districts so that preventative measures can be taken before outbreaks begin (Colborn et al., 2018). Many techniques use present or projected climatic conditions to forecast the risk of malaria in the upcoming weeks and months (Midekisa et al., 2012). Due to the complexity of forecasting with an areal model, using only those covariates does not determine a clear picture of the seasonality of malaria in the area as the seasonality is often noticed irrespective

of the climatic conditions of the area.

In combination with the variability of climatic conditions across the districts in the country, developing a malaria epidemic early warning system in lower administrative levels independently by including seasonality parameters helps to understand the heterogeneity of malaria in the area. Such approaches help to see the variability and understand seasonality across each district and then cluster districts based on temporal trends. Also, the approach gives stakeholders in each district to update their plan taking into account district-level heterogeneity in addition to what is happening in nearby areas. While developing an early warning system, (Midekisa et al., 2012) takes into account log-transformed malaria cases as responses and fits ARIMA and SARMA models. Yet, transforming aggregated count into a continuous measurement has its drawbacks (O'Hara and Kotze, 2010). However, the drawback of modelling aggregated count as a continuous measurement by transformation was first noted in (O'Hara and Kotze, 2010). This is because generalized linear models are one of the better modelling alternatives for such data sets. However, when dealing with zero counts, which are common in spatial data, a log transformation of counts has additional drawbacks (O'Hara and Kotze, 2010).

It is a good idea to begin clustering the disease risk at each administrative district based on the temporal trend and estimating the seasonality therein to gain initial insight into the current picture of disease risk. These models are crucial for locating and predicting risky areas so that timely preventive action may be taken, which is sometimes challenging to achieve using areal models. On the other hand, the spatial correlation is weaker to describe the relationship between surrounding districts when the distance between them is quite great. Decision-makers must therefore be capable of comprehending the complex dynamics not just

in space but also in time utilizing an administrative-level epidemic early warning model to forecast threats in the next weeks or months.

One can incorporate covariates to account for seasonality when developing such models, however, others consider seasonality components on a yearly or monthly basis in addition to covariates. Yet, there is no sufficient evidence to suggest which possible alternative best helps to identify the seasonality of malaria in the area which significantly varies with space and time. Using the total number of malaria cases in Southern Ethiopia, this study aims to develop an epidemic early warning system for malaria in each administrative district. The development of such a system could involve modelling temporal correlation as a Matérn process (Gneiting et al., 2006; Diggle and Giorgi, 2016). This method, which may be built by modelling malaria count as a Poisson linear mixed model, is crucial to detecting the slow underlying process, re/emergency of the risk in the area. Its objectives include:

- Identifying environmental and climatic factors associated with malaria risk in each district
- Clustering districts based on the temporal trend
- Develop a malaria epidemic early warning system that helps to detect districts with unusual risks and trends over time.

2.2 Data and Methodology

2.2.1 Study Setting

The study was carried out in Southern Nation Nationalities and People Regional State (SNNPRS), in Ethiopia Fig. 2.1. The region is located between 6° 03' 31.03" North latitude and 36° 43' 38.28" East longitude. As a result of the region's proximity to the equator, temperatures can vary from 10 degrees Celsius in high-altitude areas (4207 meters above sea level) to 28 degrees Celsius in the lowlands (360 meters above sea level). The region also has a high mean annual rainfall of 400 to 2200 mm. The SNNPRS is a vast land with unblemished topographical features, including the Great Rift Valley, mountains, forests, and plains. Since 2012, 149 rural districts have reported weekly malaria surveillance to the SNNPRS Public Health Institute.

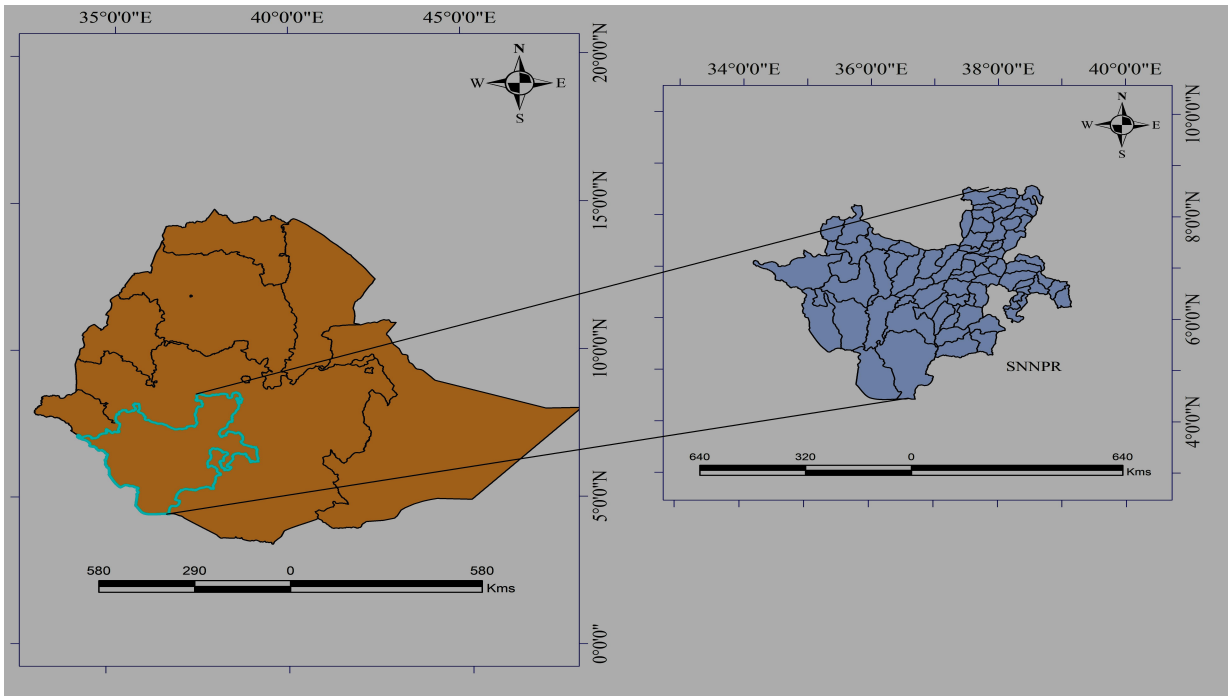

Figure 2.1 Study area map showing districts in the Southern nation nationalities and people regional state (SNNPRS), Ethiopia in 2013.

2.2.2 Data

The data set was obtained from the Ethiopian Public Health Institute. It consists of reported malaria counts of genus *P. falciparum* from August 2013 to May 2019 for the districts found in Southern Ethiopia. The district-level population data set was taken from the demographic department of the SNNPRS finance office and is projected based on 2007 Ethiopian census data (CSA, 2007). Monthly temperature (^0C) and total precipitation (mm) is extracted from weather and climate data provided at 2.5 minutes or ($\sim 21km^2$) spatial resolution (worldclim.org/data/monthlywth.html and worldclim21.html). The average monthly relative humidity is derived from ECMWF Medium-Range Weather Forecasts from ERA-Interim global atmospheric reanalysis. Finally, nighttime light (NTL) is obtained from NOAAs (National Centers for Environmental Information), Visible Infrared Imaging Radiometer Suite (https://ngdc.noaa.gov/eog/ viirs/index.html) available at approximately 100m at the equator.

2.2.3 Statistical Model

Disease risk can occasionally change across time or space, or perhaps both. When data sets are collected over a large geographic area or an extended period, it is sometimes possible to predict that the characteristics of the process S_{it} could vary between districts (Giorgi et al., 2018). The model can then be fitted individually for each district to assess the distribution of the incidence (Midekisa et al., 2012). Suppose y_{it} denote the monthly aggregated malaria counts of genus *P. falciparum* from i^{th} districts $i = 1, ..., 149$ at time $t = 1, ..., 70$ in months. Conditional on S_{it}, the aggregated count y_{it} in each administrative district are mutually independent Poisson random variables with expectation $m_{it}\lambda_{it}$ given as:

$$\log\{\lambda_{it}\} = d_{it}^{\top}\beta + S_{it} \tag{2.1}$$

Where d_{it} is a vector of space-time referenced explanatory variables with associated regression coefficients β, λ_{it}, λ_{it} is a malaria incidence rat and m_{it} is an offset representing the population at risk at each administrative district. Assuming S_{it}, temporal continuous Gaussian process in i^{th} districts, Eq. 2.1 can be re-expressed as:

$$log(\lambda_{it}) = \alpha_0 + \alpha_1 cos\left(\frac{2\pi it}{12}\right) + \alpha_2 sin\left(\frac{2\pi it}{12}\right) + \alpha_3 cos\left(\frac{2\pi it}{4}\right) + \alpha_4 sin\left(\frac{2\pi it}{4}\right) + S_{it} \quad (2.2)$$

Sometimes, it is also possible to account for seasonality through the covariates only or a combination of both. Thus, by considering the covariates also, Eq. 2.2 can be expressed as:

$$log(\lambda_{it}) = \alpha_0 + \alpha_1 cos\left(\frac{2\pi it}{12}\right) + \alpha_2 sin\left(\frac{2\pi it}{12}\right) + \alpha_3 cos\left(\frac{2\pi it}{4}\right) + \alpha_4 sin\left(\frac{2\pi it}{4}\right) + D_{it}^{T}\beta + S_{it} \quad (2.3)$$

Finally, the model in Eq.2.2 and Eq.2.3 were fitted using aggregated count data at each i^{th} district in Southern Ethiopia and then compared the prediction performance. where the temporal random effect; S_{it} assumed to follow stationary and isotropic Gaussian process with variance σ^2 and correlation between successive time assumed to be exponential with scale parameter ϕ and shape parameter κ given as:

$$corr(S_t, S_t^{'}) = \rho(t, t^{'}, \theta)$$

Where $\theta = (\sigma^2, \phi)$. The annualized linear combination of the sine and cosine functions and quarterly a year were used to model the malaria seasonality in addition to covariates for which higher incidence was observed mainly from September to December, following the main rainy seasons, and from March to May, after minimal rainy seasons (EFDR, 2019). These seasonality components were incorporated following observed malaria trends and some literature on the malaria pattern in the country (FMHE, 2020; WHO., 2019). Then, using the following formula, the forecasts can be generated for each i^{th} district at time t.

$$\lambda_i(t) = \int_t \hat{\lambda_{it}} dt$$

where $i = 1, 2, \ldots, 149; t = 1, 2, ..., 70$, each integer identifies a month, from August 2013 to May 2019. Also, $\hat{\lambda}(it)$ is the mean of the predictive distribution of intensity at time t for each district. After that, the integrals can be approximated using the MCMC method and then forecast the incidence for the next 12 months using predicted incidence.

2.2.4 Cross-validation

First, all of the data sets have been divided into training and test sets to evaluate the model's effectiveness in predicting future outcomes. After that, the model's performance was evaluated in terms of how well they were able to forecast incidence and estimate the accompanying uncertainty for the 12 months. Finally, by holding out the case reports of 12 months from June 2018 to May 2019 that are accessible, the model is fitted to the remaining data set i.e. from August 2013 to May 2018. Using the root-mean-square error, mean absolute error, and coverage probability, models' ability to predict outcomes for each of the 12 months are presented.

$$RMSE_t = \sqrt{\frac{1}{149}\sum_{i=1}^{149}\left(\lambda_{it}^{emp} - \hat{\lambda}_{it}\right)^2}$$

$$MAE_t = \frac{1}{149}\sum_{i=1}^{149}\left|\lambda_{it}^{emp} - \hat{\lambda}_{it}\right|$$

$$CP_t = \frac{1}{149}\sum_{i=1}^{149} I\left(\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}\right),$$

where λ_{it}^{emp} is the true observed incidence of i^{th} district in the test set at time $t = 1, \ldots, 41$; $\hat{\lambda}_{it}$ is the predicted mean incidence; and $I\left(\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}\right)$ is an indicator function that

takes value 1 if $\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}$ and 0 otherwise, with $\hat{\lambda}_{it}^{0.025}$ and $\hat{\lambda}_{it}^{0.975}$ corresponding to the quantiles; 0.025 and 0.975 of the predictive or posterior distribution for λ_{it}, respectively.

Finally, a variogram was also used to validate the compatibility of the fitted correlation function to the data by simulating 60,000 data sets under the fitted model. Variogram-based graphical validation was used to check fitted temporal correlation in each district using a PrevMap R package (Giorgi and Diggle, 2017). Furthermore, RMSE is used to evaluate the prediction performance of models with and without covariates. One can further see the variogram-based validation in (Giorgi et al., 2018; Giorgi and Diggle, 2017). PrevMap package in R (Giorgi and Diggle, 2017) was used to analyze the data.

2.3 Results

As shown in Fig. 2.2, the incidence of malaria changes with time and space, with an increased incidence observed in the western parts of the region. Additionally, the occurrence varies over time, with 2013, 2017, and 2018 respectively seeing the highest numbers of cases. Space-time modelling is useful for comprehending such variability and forecasting future trends.

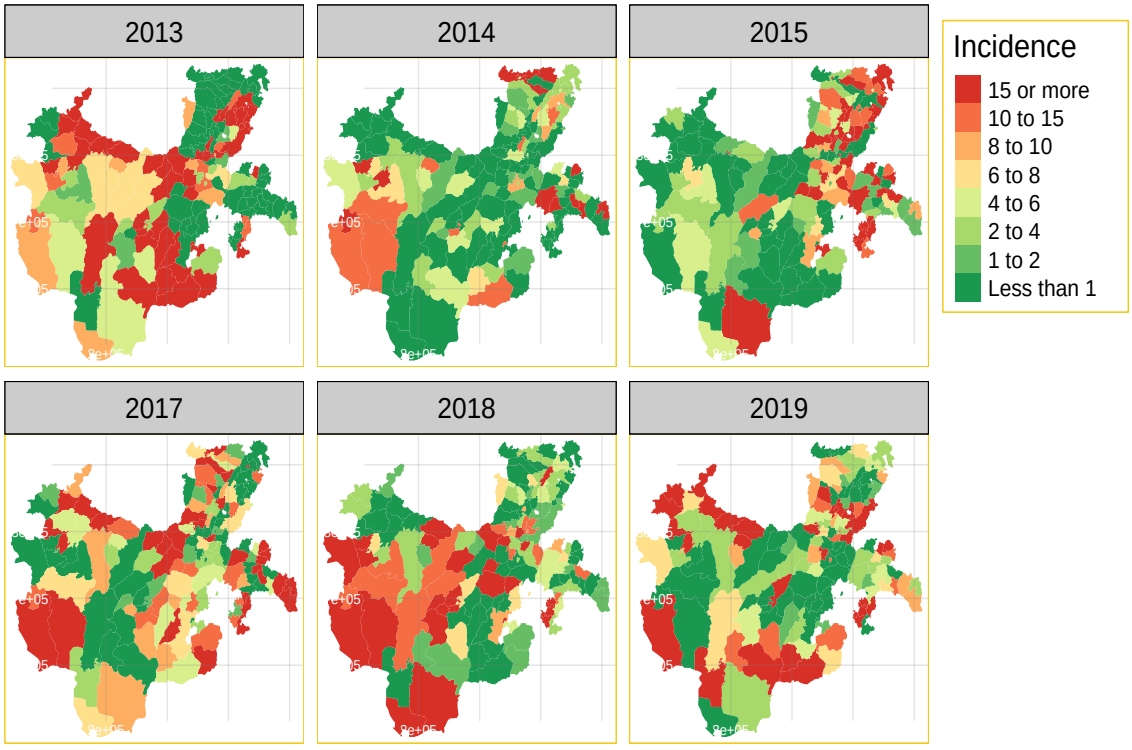

Figure 2.2 Distribution of yearly aggregated observed incidence of *P. falciparum* in all districts of Southern Ethiopia from August 2013 to May 2019 per 1000 population of all age groups

The distribution of the incidence varies over the districts with higher incidences observed in districts found in the western region and moderate incidences found in the districts found in

central areas of the region as was shown in Fig. 2.2.

For some of the randomly chosen districts, the plot of the observed incidences and residuals over time as depicted in Fig. 2.3 was provided. The illustration shows that the incidence distribution pattern alters with time and space. In particular districts like Bero and Daramalo, there is a signal of an increase in incidence as demonstrated in Fig. 2.3. On the other hand, the distribution of residuals further revealed that incidences vary across districts. As a result, space-time modelling is important to get further insight into the distribution of the incidences in each district of the region. According to the outcome in the lower panel

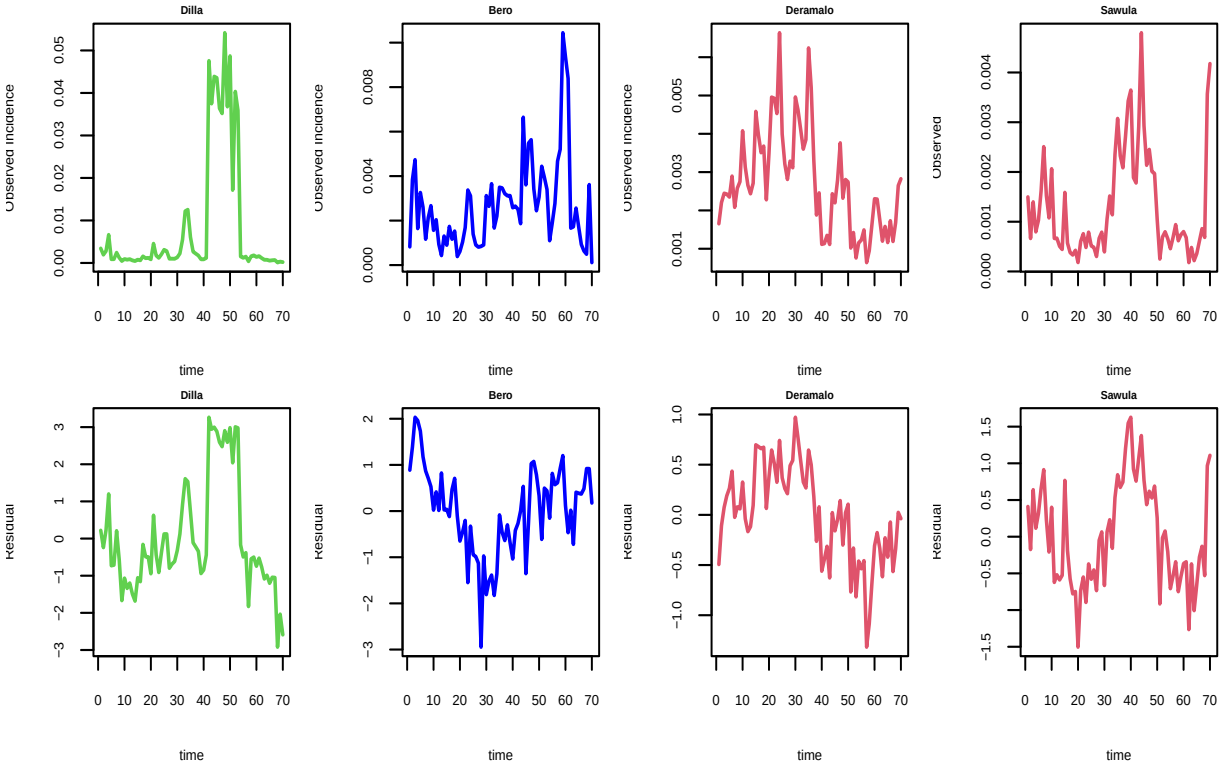

Figure 2.3 Distribution of observed incidence [upper panel] and residuals from generalized linear mixed model [lower panel] for selected districts from August 2013 to May 2019 by *P. falciparum* in Southern Ethiopia

of Fig. 2.3, it is wise to model the residuals in each administrative district over time since

they fluctuate in distribution. There is a signal of increment of the incidences in districts like Deguna-Fanigo, Bero, and Sawula despite the residuals varying over time. The residuals from the non-spatial regression model show the existence of temporal correlation as it was shown in the upper panel of Fig. 2.4. Also, the incidence is temporally correlated, as it was depicted in the variogram at various time bins in the upper panel of Fig. 2.4. This is because the observed incidences are out of the confidence bound as shown in the upper panel of Fig. 2.4 indicating the existence of temporal correlation which decreases as increases in time. Thus, developing a district-level malaria epidemic early warning system is a smart place to start identifying variability as well as epidemics throughout space and time.

Notable malaria cases by genus *P. falciparum* were observed in 2013, 2015, and 2018 across the districts. On the other hand, several districts like Boricha, Loko Abaya, Wenago, Humbo, Konta, Hamer, Basketo, Yemi, and Surima showed peak cases in the year 2013. On the other hand, only some of the districts in Sidama region, Hadiya and Gambata tamboura Zone exhibited any significant malaria incidence. An increase in malaria cases was also observed from 2018 to 2019 in several districts of the region. Therefore it should be very important to model the incidence to anticipate the variation over space and time.

2.3.1 Model Comparison and Validation

To validate our models, 1) a variogram-based approach to determine whether the fitted correlation function was compatible with the data was incorporated. Several variograms from the fitted model had been simulated, and then compared to the estimated empirical variogram derived from the data. The estimated empirical variogram, as shown in the lower panel of Fig. 2.4, completely falls within the 95% confidence interval of the simulated

empirical variograms, confirming that the adopted correlation function is appropriate with our data.

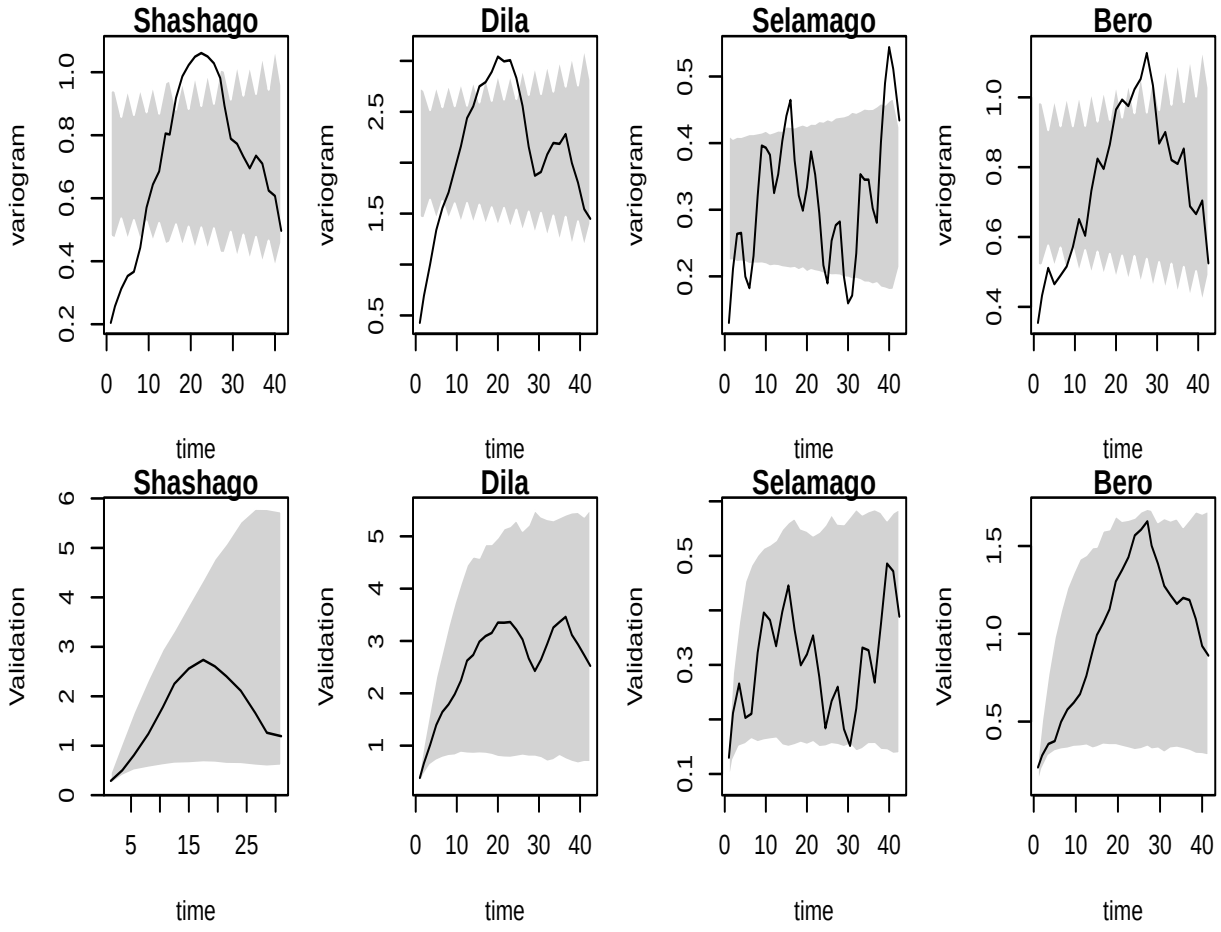

Figure 2.4 The figures show the outcomes of the Monte Carlo methods applied to test the temporal independence hypothesis (upper panels) and the data compatibility hypothesis (bottom panels) for each district. The 95% confidence region is represented by the shaded areas for each hypothesis. The solid lines represent the empirical variogram for each time bin. Using *P. falciparum* count data from Southern Ethiopia, the theoretical variograms derived using the least squares (solid lines) and maximum likelihood (dashed lines) approaches are displayed in the lower panel.

2), then the researcher considered out-of-sample validation by taking 12-month data as a test set as shown in Table. 2.1 was considered. The result indicates that the method produces a small prediction error with a coverage probability around 75% on average which is not much further apart from 95% confidence interval. Thus, the fitted model is important in predicting disease risk with minimum prediction error and coverage probability closer to 95%. Finally, the researcher evaluates the importance of including covariates in predicting

Table 2.1 Summary of out-of-sample accuracy: root mean square error (RMSE): Mean absolute error (MAR) and coverage probability (CV) obtained for the 12-month validation set data averaged to all districts

Valid	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
RMSE	0.046	0.022	0.058	0.023	0.013	0.017	0.018	0.039	0.019	0.015	0.015	0.016
MAE	0.012	0.009	0.012	0.009	0.007	0.008	0.009	0.011	0.008	0.008	0.009	0.009
CVP	78.174	75.919	69.154	65.396	74.416	75.919	66.899	72.161	70.658	71.409	69.154	67.651

space-time variation of disease risk. This was done by comparing the prediction performance of models with and without covariates. The result indicates that including covariates is very important in predicting disease as it was shown in Table. 2.2 with smaller RMSE. 2.2. Adding climatic variables to the model significantly improved the model fit in the majority of the districts with smaller RMSE, see Table.2.2.

Table 2.2 Comparison of models using: root mean square error (RMSE): obtained with covariates (1) and without covariates (2) for some selected Districts

Districts	*RMSE1*	*RMSE2*	*Districts*	*RMSE1*	*RMSE2*
Wolkite	0.036	0.06	**Sheshage**	0.0053	0.0054
Dila	0.0092	0.0095	**Deguna-Fanigo**	0.01	0.012
Selamago	0.025	0.03	**Dasenech**	0.012	0.02
Arbaminch	0.018	0.038	**Bero**	0.05	0.06
Hawassa	0.002	0.003	**Benatsemay**	0.035	0.041

Table 2.3 Parameter estimates of the models and their 95% confidence interval for some selected Districts in the Region

Districts	*Parameter*	*Estimate*	*Districts*	*Estimate*
Wolkite	β_0	-6.95 (-9.26, -4.64)*	**Sheshage**	-7.13 -11.12,-3.14)*
	$\beta_1 sin(2*pi*t/12)$	0.101 (0.065, 0.138)*		0.142 (0.109, 0.174)*
	$\beta_2 cos(2*pi*t/12)$	0.417 (0.367, 0.467)*		-0.224 (-0.264, -0.183)*
	$\beta_3 sin(2*pi*t/4)$	-0.178 (-0.360, 0.003)		-0.163 (-0.331, 0.006)
	$\beta_4 cos(2*pi*t/4)$	0.134 (0.099, 0.169)*		)0.085 (-0.058,0.228
	$\beta_5 log.precipitation$	0.223 (0.105, 0.341)*		0.307 (0.207, 0.407)*
	$\beta_6 temperature$	0.222 (0.163, 0.282)*		0.368 (0.247, 0.488)*
	$\beta_7 humidity$	0.123 (0.093, 0.154)*		0.186 (0.161, 0.211)*
	$\beta_8 NTL$	0.011 (0.010, 0.012)		-4.860 (-6.525, -3.195)
	σ^2	0.933 (0.012, 1.853)		0.890 (0.369, 1.412)
	ϕ	5.806 (5.640, 5.972)		6.976 (6.832, 7.120)
Dilla	β_0	-6.31 (-14.03, 1.42)	**Selamago**	-4.425 (-16.326, 7.476)
	$\beta_1 sin(2*pi*t/12)$	-0.097 (-0.679, 0.484)		0.089 (-0.143, 0.321)
	$\beta_2 cos(2*pi*t/12)$	0.410 (0.339, 0.482)		0.141 (-0.180, 0.463)
	$\beta_3 sin(2*pi*t/4)$	-0.039 (-0.065, -0.012)		0.104 (0.072, 0.137)
	$\beta_4 cos(2*pi*t/4)$	0.071 (0.043, 0.099)		0.278 (0.130, 0.426)
	$\beta_5 log.precipitation$	0.292 (0.170, 0.713)		0.776 (0.266, 2.015)
	$\beta_6 temperature$	0.750 (0.488, 1.011)		0.132 (0.112, 0.152)
	$\beta_7 humidity$	0.133 (0.091, 0.175)		0.009 (-0.009, 0.027)
	$\beta_8 NTL$	0.000 (-0.001, 0.002)		3.495 (-1.148, 8.138)
	σ^2	2.326 (1.967, 2.684)		0.223 (0.060, 0.385)
	ϕ	5.939 (5.783, 6.095)		1.524 (1.056, 1.991)
Bero	β_0	-9.494 (-15.047, -3.941)	**Dasenech**	-7.890 (-13.250, -2.531)
	$\beta_1 sin(2*pi*t/12)$	-0.154 (-0.486, 0.177)		0.422 (0.374, 0.470)
	$\beta_2 cos(2*pi*t/12)$	0.330 (0.030, 0.630)		0.125 (0.062, 0.188)
	$\beta_3 sin(2*pi*t/4)$	-0.066 (-0.20, 0.069)		-0.014 (-0.227, 0.198)
	$\beta_4 cos(2*pi*t/4)$	0.011 (-0.134, 0.156)		-0.119 (-0.364, 0.126)
	$\beta_5 log.precipitation$	0.634 (0.245, 1.022)		2.099 (0.507, 3.691)
	$\beta_6 temperature$	0.083 (0.012, 0.155)		0.192 (0.015, 0.370)
	$\beta_7 humidity$	-0.005 (-0.021, 0.010)		0.023 (-0.006, 0.052)
	$\beta_8 NTL$	-1.200 (-2.395, -0.004)		-0.625 (-8.857, 7.607)
	σ^2	0.768 (0.475, 1.061)		1.024 (0.414, 1.633)
	ϕ	5.393 (5.239, 5.548)		2.863 (2.592, 3.134)
Arbaminch	β_0	-4.755 (-7.924, -1.585)	**Hawassa**	-6.455 (-8.646, -4.264)
	$\beta_1 sin(2*pi*t/12)$	-0.329 (-0.577, -0.080)		0.334 (0.132, 0.535)
	$\beta_2 cos(2*pi*t/12)$	-0.070 (-0.097, -0.044)		0.085 (-0.156, 0.327)
	$\beta_3 sin(2*pi*t/4)$	0.106 (-0.002, 0.215)		0.020 (-0.082, 0.123)
	$\beta_4 cos(2*pi*t/4)$	0.118 (0.014, 0.223)		0.084 (-0.005, 0.172)
	$\beta_5 log.precipitation$	0.411 (0.256, 0.567)		0.605 (0.197, 1.012)
	$\beta_6 temperature$	0.389 (0.290, 0.487)		0.012 (-0.040, 0.063)
	$\beta_7 humidity$	-0.037 (-0.055, -0.018)		-0.008 (-0.024, 0.007)
	$\beta_8 NTL$	0.021 (0.020, 0.021)		0.000 (0.000, 0.001)
	σ^2	0.538 (0.140, 0.936)		0.784 (-1.121, 2.689)
	ϕ	7.497 (7.375, 7.619)		17.129 (17.037, 17.220)

The model parameters' maximum likelihood estimates are shown in Table 2.3 which indicates that the incidence varies throughout the year in some of the districts and seasonally in other districts. Precipitation, temperature, humidity, enhanced vegetation index, and nighttime light are significantly associated with the incidence in the majority of the district. For instance, an increase in precipitation is significantly associated with 0.223 (0.105, 0.341), 0.307 (0.207, 0.407), 0.292 (0.170, 0.713), 0.776 (0.266, 2.015), 2.099 (0.507, 3.691), 0.411 (0.256, 0.567), 0.634 (0.245, 1.022), 0.605 (0.197, 1.012) increase in malaria risk in Weleikite, Shashago, Dila, Selamago, Dasenech, Arbaminch, Bero and Hawassa districts respectively. Whereas, an increase in temperature, humidity, and nighttime light is also significantly associated with malaria risk in some of the districts Table.2.3.

When combined with the distribution of residual in Fig. 2.3, variance $sigma^2$, which is significant in the majority of the districts, shows the variability of the incidence over time. For the districts shown in Table.2.3, it is estimated that the practical range of the temporal correlation is $log(20) \times \hat{\phi}$, i.e. $log(20) \times 1.524 \approx 4.6$ months and $log(20) \times 17.13 \approx 51.3$ months respectively in Selamago and Hawassa. The practical range is defined as the months beyond which the temporal correlation is below 0.05. In Selamago and Hawassa, respectively, the 95% confidence interval for the practical range ranges from 3.16 to 6 months and between 50 and 52 months.

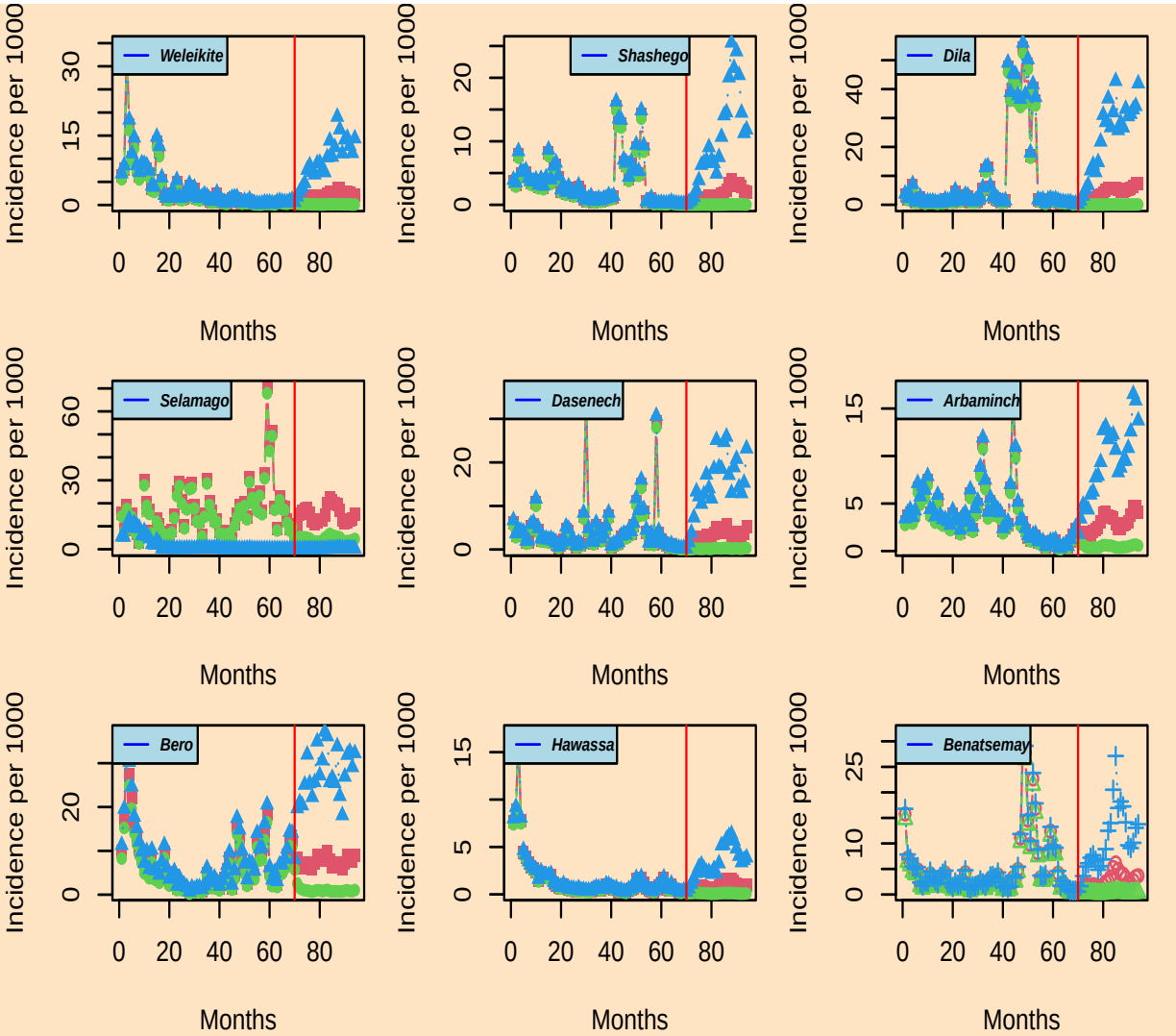

Figure 2.5 Prediction results for the districts of Weleikite, Shashego, Dila, Selamago, Duguna-fango, Dasenech, Arbaminch, Bero, and Hawassa overall months from 2013 and forecast for the 24 months. The plot shows predictive inference and associated 95% confidence interval for *P. falciparum*

Even though no significant outbreaks were observed in the study area during the study period, there were times when notable incidences were observed from September to November 2013 in the Woleikite, November to December 2016 in Shashego, January to March 2018 in Dila, June to August 2018 in Selamago, October to November 2015 and January to February 2017 in Arbamich and early 2013 in Hawassa city Fig.2.5. The right panel of Fig.2.5 indicates the forecast for the next 24 months. The forecast indicates, the re-emergence of the incidence in the area as indicated in Fig.2.5. In general, the decreasing trend of the incidence was observed till the end of 2018 with an unstable rate, and there is a signal of re-emergency starting from 2019. For the last 24 months of the time series, the forecasted incidences were provided, and the result indicates a signal of the outbreak for most districts in the region.

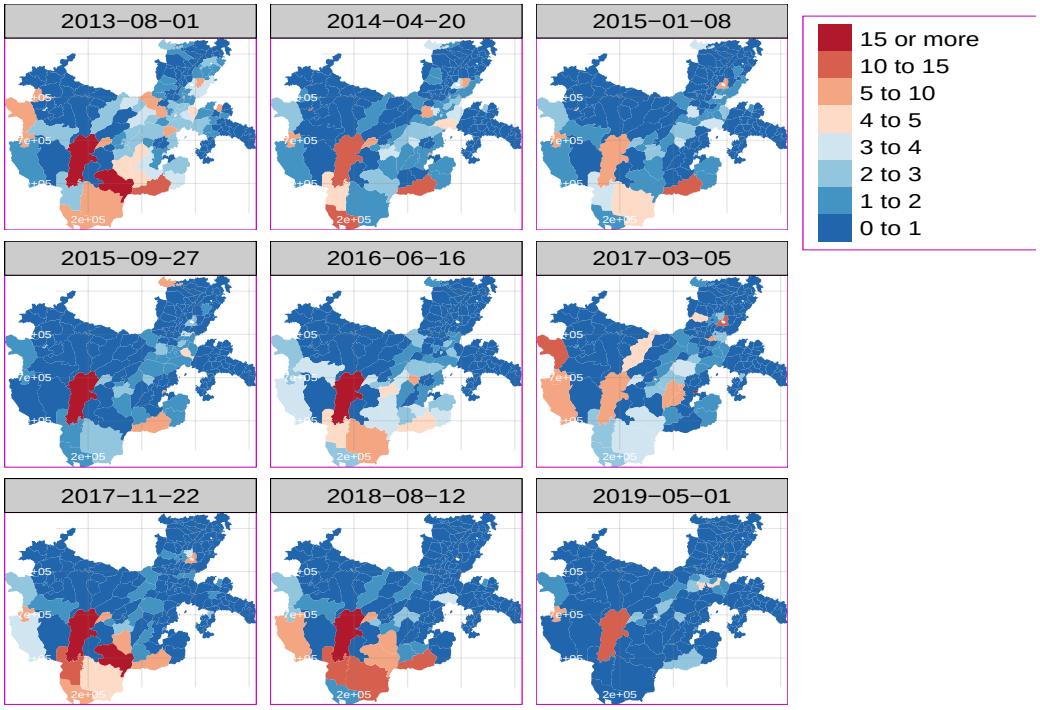

Figure 2.6 Prediction incidence for the selected months of the years for all districts for *P. falciparum* incidence per 1000 population

The variation in incidences throughout time and space is depicted in the prediction map in Fig.2.6. The southwest's Bero, Selamago, Hamer, and Dasenech districts also saw higher incidences. Moreover, a considerable number of districts in the region's northwest and centre have a moderate to high incidence of malaria. Districts in the north and southeast, on the other hand, exhibited a lower incidence.

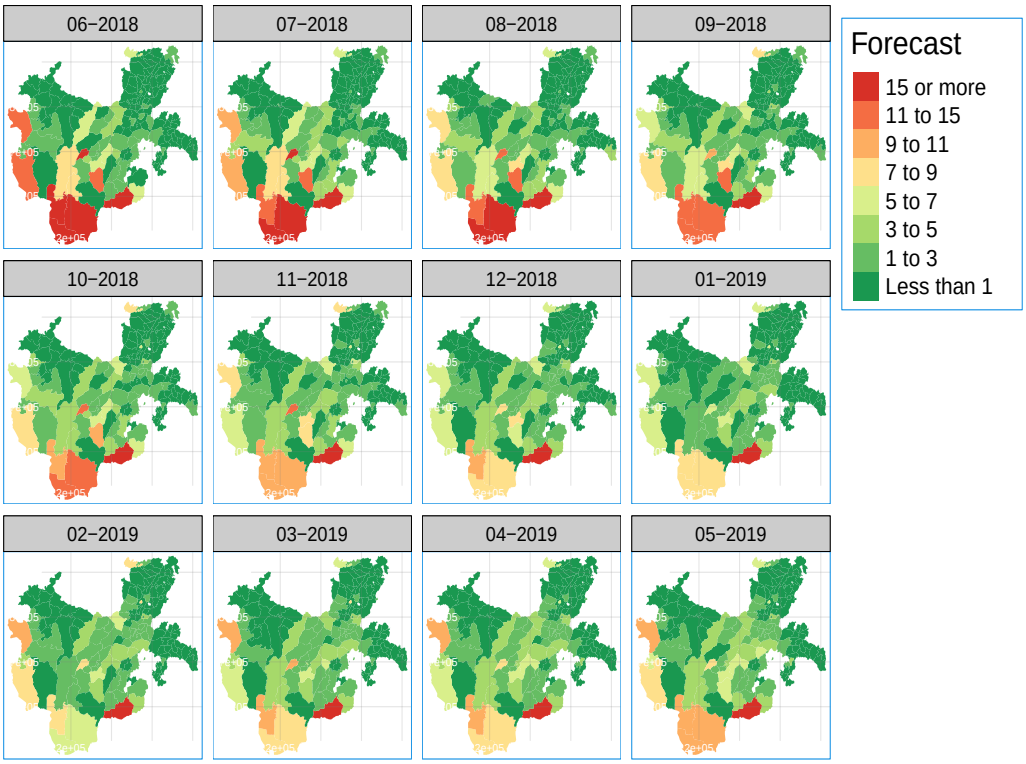

Figure 2.7 Forecasted map of *P. falciparum* incidence per 1000 population in the region from June 2018 to May 2019 for all districts

2.4 Discussion

According to our study results, it would be ideal to establish a malaria epidemic early warning system in each administrative district to better understand the geographic distribution of cases. As noted by (Dabaro et al., 2021) in one of the local administrative districts, the incidence varies among the districts. This highlights how establishing an early warning system for epidemics in each district enables the local administration to act right away to address the issue rather than constantly waiting for a solution from outside sources. Similarly to (WHO, 2022; FMHE, 2020) report, the temporal trend in the region varies over time from district to district as displayed in Fig.2.3. The transmission is not consistent; rather, it varies from district to district Fig.2.2, Fig.2.5, and Fig.2.6, with some exhibiting a similar pattern.

To forecast the pattern and reemergence of malaria as shown in Fig.2.7 across several districts, a time series model-based epidemic early warning system for malaria is a promising choice. Yet, as demonstrated in Table.2.2, incorporating variables dramatically improved prediction performance. However, if an outbreak is only briefly observed before going away, the value of EWSs as a forecasting tool for policymakers may be not used. The forecasts as shown in Fig.2.7 and Fig.2.5 indicate cases are emerging in some of the districts in the region. As model primarily forecasts a nonlinear increase as was shown in Fig.2.5, which is characteristic of the seasonal pattern of the illness risk and is consistent with (Midekisa et al., 2012; Nigussie et al., 2022).

Even though this study sheds light on the resurgence of the incidents in various districts, the applicability of its findings is frequently constrained by the poor quality of the available data (Ohrt et al., 2015). This is due to the effect of reporting bias affecting surveillance data in low-resource environments. When districts are spread out and the importance of the spatial

correlation is uncertain, time series modelling might be useful in identifying these situations. The modelling of such cases independently provides superior information regarding the incidences for specific districts, however, as some districts may show distinctive tendencies when compared to other districts.

Increases in rainfall have a considerable impact on malaria risk, as shown in Table.3.1, which is consistent with earlier research (Midekisa et al., 2012; Giorgi et al., 2021). This may be because mosquitoes breed near water bodies, where they are occasionally visible after a very strong rainstorm. This is typical in the majority of the districts in the area, as it has been widely discussed in various works of literature (Rodo et al., 2021). On the other hand, there is a strong correlation between the risk of malaria and the temperature. Malaria risk rises by 0.22 in Woleikte, 0.37 in Shashego, 0.750 in Dila, 0.132 in Selamago, 0.083 in Bero, 0.19 in Dasenech, 0.39 in Arbaminch, and 0.012 in Hawassa city with each degree of temperature increase. This implies that a rise in temperature and rainfall correlates with (Dabaro et al., 2021) and has a strong favourable impact on all districts with moderate to high malaria cases. In Woleikte, Shashego, and Dila districts, respectively, an increase in humidity is similarly linked to increases in malaria risk of 0.123, 0.186, and 0.133%; however, this association is not statistically significant in some of the other districts. On the other hand, a rise in nighttime light is adversely correlated with 0.86 in Shashsgo, 0.2 in Bero, and 0.625 in Dasenech increase in malaria risk. Yet there were also discovered positive connections in other districts as presented in Table. 3.4.

As a signal of re-emergency is noticed in some districts in the region, malaria risk prediction with greater accuracy is currently crucial in nations like Ethiopia as seen in Fig. 2.5. Someone might take into account a different option that aids in the detection of hotspots while

designing such a system. Because infectious diseases like malaria fluctuate with both space and time with notable influence of nearby areas, it may therefore be very important to build a system that considers spatial correlation into account.

Fitting more precise malaria models for the future, not only to revisit our findings but also to precisely address various questions about the underlying trend in districts and various malaria cases with multiple *Plasmodium* species could be important. A promising path has recently been opened up by the development of discrete and continuous spatial models for infectious disease dynamics (Diggle and Giorgi, 2019). Future research anticipates going into more detail on some of the important drivers and aspects that this study did not sufficiently explore regarding the prevalence of malaria in Southern Ethiopia. The results obtained highlight the value of dynamically identifying districts with elevated risks, but additional modelling is needed to fully understand the spatio-temporal variations of malaria risk in the region.

2.5 Conclusion

Space-time modelling of malaria risk is crucial for describing the aetiology of the disease and directing decision-making at the lower administrative levels. A re-emergency signal was seen a few months/years ago with an unstable rate despite the incidence having been on the decline for the past few years. Districts found in the southwest have detected higher incidence that varies with time. When incidence heterogeneity increases, it is important to address "bottlenecks" such as dealing with persistent foci, subsequent re/emergencies, and parasite development areas. In a changing climate, sustainable and adaptive plans should now be guided from an informed local level.

Declarations

Ethics Approval and consent to participate

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the College of Natural and Computational Science, Hawassa University (protocol code: RERC/030/12, and date of approval: July 20, 2022). The ethical review board, College of Natural and Computational Sciences of Hawassa University Research Ethics Review Committee (RERC) waived participant informed consent since the study was conducted using district-level monthly surveillance data.

Consent for publication

Not applicable

Availability of data and materials

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data-sharing policy of the Ethiopian Public Health Institute.

Competing interests

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

YS conceived the study idea, designed the study, carried out the statistical analyses, interpreted the results, and drafted the manuscript. ZG designed the study reviewed the manuscript provided technical support for interpreting results and reviewed it for intellectual content. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to the Ethiopian Public Health Institute worker for facilitating data collection and to Hawassa University for creating an opportunity to participate in the research work.

2.6 References

Abeku, T., Vlas, S.D., Borsboom, G., Tadege, A., Gebreyesus, Y., Gebreyohannes, H., Alamirew, D., Seifu, A., Nagelkerke, N., Habbema, J., 2004. Effects of meteorological factors on epidemic malaria in ethiopia: a statistical modelling approach based on theoretical reasoning. Parasitology 128, 585–593. URL: https://researchonline.lshtm.ac.uk/id/eprint/14632/.

Cohen, D., Person, M., Wang, P., Gable, C.W., Hutchinson, D., Marksamer, A., Dugan, B., Kooi, H., Groen, K., Lizarralde, D., et al., 2010. Origin and extent of fresh paleowaters on the atlantic continental shelf, usa. Groundwater 48, 143–158.

Colborn, K.L., Giorgi, E., Monaghan, A.J., Gudo, E., Candrinho, B., Marrufo, T.J., Colborn, J.M., 2018. Spatio-temporal modelling of weekly malaria incidence in children under 5 for early epidemic detection in mozambique. Scientific reports 8, 1–9. doi:10.1038/s41598-018-27537-4.

CSA, 2007. Central statistical authority, 2007 population and housing census of ethiopia. country level. Addis Ababa, Ethiopia .

Dabaro, D., Birhanu, Z., Negash, A., Hawaria, D., Yewhalaw, D., 2021. Effects of rainfall, temperature and topography on malaria incidence in elimination targeted district of ethiopia. Malaria journal 20, 1–10.

Deress, T., Girma, M., 2019. Plasmodium falciparum and plasmodium vivax prevalence in ethiopia: a systematic review and meta-analysis. doi:10.1155/2019/7065064.

Diggle, P., Giorgi, E., 2016. Model-based geostatistics for prevalence mapping in low-resource settings. American Statistical Association 111, 1096–1120. doi:10.1080/01621459.2015.1123158.

Diggle, P., Giorgi, E., 2019. Model-based Geostatistics for Global Public Health: Methods and Applications. Chapman & amp; Hall/CRC Interdisciplinary Statistics, Chapman and Hall/CRC Press.

EFDR, 2019. Ethiopian ministry of health report, malaria epidemiological profile doi:www.moh.gov.et/ejcc/en/malaria.

FMHE, 2020. Ethiopia malaria elimination strategic plan by federal ministry of health ethiopia: 2021-2025 .

Giorgi, E., Diggle, P., 2017. PrevMap: An R package for prevalence mapping. Journal of Statistical Software 78, 1–29. doi:10.18637/jss.v078.i08.

Giorgi, E., Fronterre, C., Macharia, P.M., Snow, V.A.A.R.W., Diggle, P., 2021. Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict. J. R. Soc. Interface 18. doi:10.1098/rsif.2021.0104.

Giorgi, E., P. Diggle, R.W.S., Noor, A.M., 2018. Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys. International Statistical Review 86, 571–597. doi:10.1111/insr.12268.

Girum, T., Shumbej, T., Shewangizaw, M., 2021. Burden of malaria in ethiopia, 2000-2016: findings from the global health estimates 2016. Trop Dis Travel Med Vaccines 5. doi:10.1186/s40794-019-0090-z.

Gneiting, T., Genton, M., Guttorp, P., 2006. Geostatistical space-time models, stationarity, separability and full symmetry. American Statistical Association 97, 590–600. doi:10.1201/9781420011050.ch4.

Harris, M.J., Hay, S.I., Drake, J.M., 2020. Early warning signals of malaria resurgence in kericho, kenya. Biology letters 16, 20190713.

Midekisa, A., Beyene, B., Mihretie, A., Bayabil, E., Wimberly, M.C., 2015. Seasonal associations of climatic drivers and malaria in the highlands of ethiopia. Parasites & vectors 8, 1–11.

Midekisa, A., Senay, G., Henebry, G., Semuniguse, P., Wimberly, Michael, C., 2012. Remote sensing-based time series models for malaria early warning in the highlands of ethiopia. Malar J 11. doi:10.1186/1475-2875-11-165.

Negash, K., Kebede, A., Medhin, A., Argaw, D., Babaniyi, O., Guintran, J., Delacollette, C., 2005. Malaria epidemics in the highlands of ethiopia. East African medical journal 82.

Nigussie, T.Z., Zewotir, T.T., Muluneh, E.K., 2022. Detection of temporal, spatial and spatiotemporal clustering of malaria incidence in northwest ethiopia, 2012–2020. Scientific reports 12, 3635.

Ohrt, C., Sturrock, K.R.H., Wegbreit, J., Lee, B., Gosling, R., 2015. Information systems to support surveillance for malaria elimination. Am J Trop Med Hyg. 93. doi:10.4269/ajtmh.14-0257.

Organization, W.H., 2001. Malaria early warning systems–concepts, indicators and partners. A framework for field research in Africa, Geneva 47.

O'Hara, R.B., Kotze, D.J., 2010. Do not log-transform count data. methods in ecology and evolution, 1, 118-122.

Rodo, X., Martinez, P., Siraj, A., Pascual, M., 2021. Malaria trends in ethiopian highlands track the 2000 'slowdown' in global warming. Nat Commun 12. doi:10.1038/s41467-021-21815-y.

Seyoum, D., Yewhalaw, D., Duchateau, L., Brandt, P., Rosas-Aguirre, A., Speybroeck, N., 2017. Household level spatio-temporal analysis of plasmodium falciparum and plasmodium vivax malaria in ethiopia. Parasites & vectors 10, 1–11.

Taffese, H., Hemming-Schroeder, E., Koepfli, C., Tesfaye, G., Lee, M., Kazura, J., G.Y, Y., Zhou, G., 2018. Malaria epidemiology and interventions in ethiopia from 2001 to 2016. Infect Dis Poverty 7. doi:10.1186/s40249-018-0487-3.

Teklehaimanot, H.D., Lipsitch, M., Teklehaimanot, A., Schwartz, J., 2004. Weather-based prediction of plasmodium falciparum malaria in epidemic-prone regions of ethiopia i. patterns of lagged weather effects reflect biological mechanisms. Malaria Journal 3. doi:10.1186/1475-2875-3-41.

Tessema, S.K., Belachew, M., Koepfli, C., Lanke, K., Huwe, T., Chali, W., Shumie, G., Mekuria, E.F., Drakeley, C., Gadisa, E., et al., 2020. Spatial and genetic clustering of plasmodium falciparum and plasmodium vivax infections in a low-transmission area of ethiopia. Scientific reports 10, 1–10.

WHO, 2015. Global technical strategy for malaria 2016-2030. World Health Organization.

WHO., 2019. World malaria report 2019. Geneva: World Health Organization doi:www.who.int/malaria/publications/world-malaria-report-2019.

WHO, 2022. World malaria report 2022: World Health Organization and others. World Health Organization.

Paper II

Understanding the Importance of Spatial Correlation in Identifying Spatio-temporal Variation of Disease Risk, in the Case of Malaria Risk Mapping in Southern Ethiopia

Yonas Shuke Kitawa[1*]: Department of Statistics, College of Natural and Computational Science, Hawassa University, Hawassa, Ethiopia
Olatunji Johnson[2]: Department of Mathematics, University of Manchester, Manchester, UK
Emanuele Giorgi[3]: CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK
Zeytu Gashaw Asfaw[4]: Department of Bio-statistics and Epidemiology, School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia,

CHAPTER 3. Understanding the Importance of Spatial Correlation in Identifying Spatio-temporal Variation of Disease Risk, in the Case of Malaria Risk Mapping in Southern Ethiopia

Abstract

Malaria remains a major health problem in developing countries despite a significant reduction in incidence in the last few years. Disease mapping thus helps to understand the spatial pattern and identify areas characterized by unusual risks. Several spatial models have been used to analyze the incidence of malaria. We aim to compare the predictive performance of these models and investigate the effect of ignoring spatial correlation. The reported malaria case counts of genus *plasmodium falciparum* in 149 districts of southern Ethiopia from January 2016 to May 2019 were analyzed using the spatial time series model (STS) that ignores spatial correlation, Spatio-temporal conditional autoregressive model (STCAR), Spatio-temporal geostatistical model (STG) and Spatio-temporal spatial discrete approximation to log Gaussian cox process (STSDALGCP). We assess the predictive performance of the models using root mean square error, mean absolute error, and coverage probability. We found that monthly average rainfall, temperature, humidity, and EVI are significantly associated with malaria risk. The spatial variation of malaria incidence changes with time, in particular, the high incidence was observed from November to December, months after heavy rainfall, and more pronounced in the southwest of the country. STSDALGCP gives a small prediction error and captures the uncertainties better than other models, while the STS model gives a high prediction error. Accounting for spatial correlation is crucial for disease risk mapping and leads to better prediction of disease risk. Since malaria transmission operates in a spatially continuous manner, a spatially continuous model should be considered when it is computationally feasible.

Keyword: *Disease mapping, geostatistics, log-Gaussian Cox process, STSDALGCP, Monte Carlo maximum likelihood, P. falciparum*

3.1 Introduction

Spatial modelling and prediction of disease risks have been extensively used to understand the spatial pattern of disease and identify areas characterized by unusual risks (Anderson et al., 2014; Diggle and Giorgi, 2016; WHO., 2019). The information provided by those analyses is crucial for health researchers as it gives insight into the aetiology of disease and supports monitoring prevention efforts. The predictions of the spatial variation of disease risk can now be made available in different spatial and temporal scales, (Wakefield, 2007) depending on the scale at which policy decisions and/or interventions are made. Disease mapping has been a major guiding tool for carrying out interventions and formulating policies, especially in low-resource settings where disease data (including malaria) are geographically incomplete.

Malaria remains a major cause of illness and death particularly in sub-Saharan Africa even though most countries have considerably reduced the burden via several interventions (Bhatt et al., 2015; Weiss et al., 2019). The interventions have resulted in over 44%, 50%, and 40% decline in mortality, infection prevalence, and clinical incidence, respectively worldwide from 2010 to 2019 (WHO., 2020). In Ethiopia, malaria is the leading cause of morbidity and mortality (Yeshiwondim et al., 2009; Taffese et al., 2018; Girum et al., 2021) with high transmission during the autumn to the spring season and varying levels of intensity in different districts of the country (Girum et al., 2021; Rodo et al., 2021). Some districts have experienced several outbreaks that have affected the health as well as economic development of many people (Yeshiwondim et al., 2009). Therefore, it is important to look at malaria incidence by the district over several years to identify areas of unusual risk to gain more understanding of the effectiveness of a malaria control program. Thus, developing a predictive model to identify the spatio-temporal variation of disease risk that enables efficient and

timely intervention is important.

Suppose the disease count data y_{it} is obtained from a discrete set of potential locations R_i within an area of interest A, at each of a sequence of times $t; t = 1, \ldots, T$. Modelling this kind of data falls under the general class of generalized linear mixed model (GLMM), where the model has two components namely, fixed effect and random effect. In particular, conditional on random effects S_{it}, the count y_{it} are mutually independent Poisson random variables with mean $m_{it}\lambda_{it}$ such that:

$$\log\{\lambda_{it}\} = d_{it}^{\top}\beta + S_{it} \tag{3.1}$$

where d_{it} is a vector of spatio-temporally referenced explanatory variables with associated regression coefficients β, λ_{it} is a malaria incidence rate and m_{it} is an offset representing the population at risk. S_{it} can be modelled using either a spatially discrete process or a spatially continuous process. A choice that depends on the scientific understanding of the data-generating process. The most common choice for district-level data is the spatially discrete process, a choice that is largely based on computational convenience and the data format. We refer the reader to the paper by (Diggle et al., 2013) for more discussion on this issue.

We employed four different models to analyze the data. The first model is a time series model that models each district separately. Therefore, treating each district-level data independently. We shall call this model STS denoting spatial time series. This approach was used by (Kitawa and Asfaw, 2023), (Midekisa et al., 2012),(Dabaro et al., 2021), and (Girond et al., 2017), among others, but a few to predict the incidence of malaria. The approach accounts for temporal correlation but ignores the spatial correlation between the districts, hence, loss of the potential benefits that might be gained through modelling malaria risk

using spatial, as well as temporal dependence (Colborn et al., 2018). The approach is sensible when the districts are distant apart, bordered by water and/ or other things for which including spatial correlation might have no significant effect. The other models are spatial models with different specifications of the random effect, S_{it} in Eq. 3.1.

In the second approach, we model S_{it}, as a spatial discrete process using the space-time extension of the conditional autoregressive (CAR) priors model (Rushworth et al., 2017). We refer to this model as STCAR. These models have been applied successfully in different geographical settings for modelling spatially aggregated data but, they depend on the neighbourhood structure of the districts (Besag et al., 1991; Anderson et al., 2014; Lee et al., 2018; Martínez-Beneito et al., 2008). This approach is computationally efficient and it has been widely used in many spatio-temporal disease mapping. However, the way neighbourhoods are defined is sometimes arbitrary unless all districts are similar in size, and shape and arranged in a regular pattern (Wakefield, 2007; Li et al., 2012a). Also, it assumes constant risk within the spatial units, despite disease described over space, hence, failing to account for the internal composition in terms of a particular location and population characteristics (Benjamin et al., 2018).

The third model considers S_{it} as a spatial continuous process using a model-based geostatistical (MBG) approach (Diggle and Giorgi, 2019), thereby assuming that the data set is observed at the centroid of the district. Therefore treating district-level data as point data. We refer to this approach as STG denoting a spatio-temporal geostatistical model. The natural way to model point-referenced data is to use the MBG method as it provides a principled way to map diseases (Diggle and Giorgi, 2019). However, representing a district by a point is often questionable as the correlation between the districts depends on the distance

between the centroids of the district without accounting for the shape of the district.

The fourth model takes S_{it}, as a spatial continuous process using a computationally discrete approximation to log-Gaussian Cox process (LGCP) (Johnson et al., 2019). This is an approximation of the LGCP method proposed by (Diggle et al., 2013) and (Taylor et al., 2015). This approach accounts for the size and shape of the spatial index set which is one of the limitations of the CAR model. Another attractive attribute of this approach is the reduction of the computational cost in LGCP (Diggle et al., 2013) through discretizing continuous regions using various spatial scales.

Several models (Li et al., 2012b; Konstantinoudis et al., 2020; Paige et al., 2020; Utazi et al., 2021) have been developed to compare the prediction performance of models, but they focus only on some particular class of models. The computational issue as integrated nested Laplace approximations (INLA) (Bivand et al., 2015; Illian et al., 2012) helps to obtain posterior estimates more quickly than MCMC methods (Diggle and Giorgi, 2019, 2016; Christensen, 2004) and precision of estimates (Joe, 2008) play a great role in the selection of appropriate models. Furthermore, considering spatial proximity in determining the spread of geographic processes, such as fluctuations in disease incidence, one can look at discrete or continuous spatial models.

The majority of disease mapping research is based on the conditional autoregressive (CAR) prior distribution (Besag, 1974), which involves smoothing the disease risk locally by borrowing information from surrounding districts. Over-smoothing, on the other hand, can occasionally inhibit the detection of high-risk locations or hot spots. Similarly, geostatistical modelling is often utilized in disease risk mapping (Diggle and Giorgi, 2019). This is particularly apparent in malaria risk mapping, where statistical models that explicitly account

for space and time have been employed to map essential parameters over enormous spatial extents to compensate for data sparsity.

However, there is still no model that is the ideal choice in all scenarios. Some models are especially well suited for some circumstances but not for others due to the specific characteristics of each unique data set, such as the existence of excessive zeroes, discontinuities, or lack of spatial dependence among units. Understanding the variety of models that are accessible enables one to choose at least one suitable model for a particular situation (Wong et al., 2023). Thus, here we look forward to choosing models that are especially suited for predicting disease risk using spatially aggregated data in southern Ethiopia.

Thus, the overachieving aim of this study is: 1) to develop a predictive model that helps to understand the spatio-temporal variation of malaria risk at the district level in Southern Ethiopia; 2) to compare the prediction performance of models; and 3) to understand the effect of ignoring the spatial correlation in predicting disease risk.

The structure of the paper is as follows. Section 2 describes statistical models used to analyze spatially aggregated disease count data. More specifically, the spatial time series (STS) model, spatio-temporal conditional autoregressive model (STCAR), spatio-temporal geostatistical model (STG), and spatio-temporal spatially discrete approximation to log Gaussian Cox process (STSDALGCP) including the procedures for parameter estimation and spatial prediction. Section 3 gives the result of the analysis using the four models. We conclude with a discussion in Section 4.

3.2 Malaria data and the predictors

The data set used in the study is reported malaria case counts of genus *P. falciparum* y_{it}, aggregated over 149 districts in southern Ethiopia from January 2016 to May 2019.

The data was obtained from the Ethiopian Public Health Institute. We also assemble environmental and demographic data as predictors of malaria incidence. The district-level population data were obtained from the regional demographic department and it is projected based on the 2007 Ethiopian census data (CSA, 2007). Satellite data sets of rainfall are obtained from TAMSAT (Tropical Applications of Meteorology using Satellite data and ground-based observations) at 4 km spatial resolution, average minimum temperature (^0C), average maximum temperature (^0C), and total precipitation (mm) were obtained from monthly weather and climate data provided at 2.5 minutes or ($\sim 21km^2$) spatial resolution (worldclim.org/data/monthlywth.html and worldclim21.html). Average relative humidity was derived from ECMWF Medium-Range Weather Forecasts from ERA-Interim global atmospheric reanalysis. Finally, the Enhanced vegetation index (EVI) was obtained from a Moderate-resolution Imaging Spectroradiometer (MODIS) available at $1km^2$ spatial resolution (lpdaac.usgs.gov/products/mod13a3v006/). The satellite data for the areal models, i.e. for models 2 and 4 were aggregated to the district level through averaging points failing inside each polygon or district, and for other models, i.e. model 1 and 3, we have considered covariates at the centroid of each district.

3.3 Statistical models for spatially aggregated data

For the analysis of geographically aggregated data, several statistical models are frequently employed. Discrete spatial models and continuous spatial models are the two basic categories

under which this class of models can be divided. The way the random effects are modelled is where these two types of models diverge most. For example, discrete spatial models that assume a fixed spatial domain for disease risk captures the spatial correlations between districts using an areal random effect. The continuous Gaussian process random effect is used to account for the spatial correlation between districts in continuous spatial models when the disease risk is assumed to have a continuous spatial domain. Moreover, one can fit a discrete model and can make continuous inferences (Benjamin et al., 2018). We direct the reader to the book by (Diggle and Giorgi, 2019) and the article by(Johnson et al., 2019) which provides the description and the connection between these approaches in more detail.

Here, we discuss in detail the four models used to analyze malaria incidence. Let y_{it} denote the reported malaria counts of genus *P. falciparum* at district i and time t, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$ corresponding to each month from January 2016 to December 2019, for T=41 and N=149. We modelled y_{it} using the model described in Eq. 3.1.

Model 1: Spatial Time Series (STS) model

The Spatial Time Series (STS) model is the model of the form in Eq. 3.1 developed for each of the i^{th} administrative districts in Ethiopia such that the random effect S_{it} is a temporal process modelled as Gaussian process. This model is similar to the model developed by (Midekisa et al., 2012) but the model reported counts as a continuous measurement using the seasonal autoregressive integrated moving average (SARIMA) model. On the other hand, by specifying the temporal correlation using the Gaussian process; see, for example, (Gneiting and Guttorp, 2010; Diggle and Giorgi, 2016) in Eq. 3.1, (Kitawa and Asfaw, 2023) developed a model in each administrative district following a modelling approach by (Midekisa et al.,

2012) and suggestions by (Giorgi et al., 2018) using spatial time series model. This model helps to understand the temporal distribution of the incidence in each district without accounting for the spatial correlation. This type of model is normally used in countries where there is no coordination among the districts and each district make policy decision using only its district data.

Taking into account d_{it} as a vector of the explanatory variable with associated regression coefficients β in Eq. 3.1 at each administrative district, S_{it} separately can be modelled, developing 149 separate models for our data. Hence, the temporal random effect in each district, S_{it} is modelled as a zero-mean stationary and isotropic Gaussian process with covariance function given as:

$$Cov(S_{it}, S_{it'}) = \sigma^2 \rho(v; \psi),$$

where $\rho(\cdot; \psi)$ is an exponential correlation function with temporal scale parameter ψ; $v = \|t - t'\|$ is the Euclidean distance between the time points t and t'; and σ^2 is the variance.

Model 2: Spatio-temporal Conditional Autoregressive (STCAR) model

The spatio-temporal Conditional Autoregressive (STCAR) model is the model of the form in Eq. 3.1 where the random effect S_{it} is used to account for residual spatio-temporal autocorrelation in the data set. There are several ways to specify the structure of S_{it} (Lee et al., 2018). We consider the STCAR model proposed by (Rushworth et al., 2017) which is the space-time extension of CAR prior proposed by (Leroux et al., 2000). This model outperforms other popular choices like BYM models (Lee, 2011). The model decomposes the density of the set of random effects $S = (S_1, \ldots, S_T)$ as

$$f(S_1, \ldots, S_T) = f(S_1) \prod_{t=2}^{T} f(S_t | S_{t-1})$$

where $S_t = (S_{1t}, \ldots, S_{Nt})$ denotes the vector of random effect for period t. The temporal correlation is induced by allowing S_t to depend on S_{t-1}, while the spatial correlation is modelled using the CAR prior (Leroux et al., 2000). The CAR prior specified for $f(S_t)$ induces spatial autocorrelation into the random effects at time t using the binary $N \times N$ adjacency matrix $\mathbf{W}$, which is based on the contiguity structure of the N districts.

The specification of the adjacency matrix $\mathbf{W}$ is based on the neighbourhood structure of the districts. Therefore, $w_{ij} = 1$ if districts R_i and R_j share a common border and zero otherwise. The joint prior distribution for S_i is given by $(S_1, \ldots, S_N) \sim N(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W}, \xi)^{-1})$, where $\mathbf{Q}(\mathbf{W}, \xi)$ is the precision matrix and $\xi \in [0, 1]$ is the spatial dependence parameter with $\xi = 0$ corresponds to independence and $\xi = 1$ intrinsic CAR prior (Besag et al., 1991). The precision matrix $\mathbf{Q}(\mathbf{W}, \xi)$ is modelled using the approach proposed by (Leroux et al., 2000), given as $Q(\mathbf{W}, \xi) = \xi[diag(\mathbf{W1}) - \mathbf{W}] + (1 - \xi)\mathbf{I}$, where $\mathbf{1}$ is a vector of ones and $\mathbf{I}$ is an identity matrix. Assuming CAR prior for each S_t, the temporal autocorrelation is induced in the random effect as; $S_t \sim N(\alpha S_{t-1}, \tau^2\mathbf{Q}(\mathbf{W}, \xi)^{-1})$ with $S_1 \sim N(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W}^{-1})$ for $t = 2, \ldots, T$, where $\alpha \in [0, 1]$ controls temporal auto-correlation. Naturally, a real data set reveals a spatially varying level of correlation and (Rushworth et al., 2017) argued to incorporate these varying correlations through modelling $w_{ij} = 1$ as unknown parameters. Consider $\mathbf{w}^+ = (w_{ij}/i \sim j)$ as a random variable which assumes multivariate Gaussian prior on the transformed scale; $\mathbf{v}^+ = \log\left\{\frac{\mathbf{w}^+}{(\mathbf{1}-\mathbf{w}^+)}\right\}$. Then, prior for $\mathbf{v}^+$ with a constant mean μ and variance ς^2 given as: $f\left(\mathbf{v}^+|\tau_w^2, \mu\right) \propto exp\left[-\frac{1}{2\tau_w^2}\sum_{v_{ij}\in\mathbf{v}^+}\left(v_{ij}-\mu\right)^2\right]$; where, $\tau_w^2 \sim Inverse - Gamma(a, b)$. As $\tau_w^2 \to 0$, the elements of $\mathbf{v}^+$ shrunk to μ and then the model becomes global smoothing mode as developed by (Rushworth et al., 2014). When τ_w^2 increases, there is spatial clustering or step changes around a common vertex.

Model 3: Spatio-temporal geostatistical model (STG)

Here we apply the spatio-temporal geostatistical model for point-referenced data to the district-level data. We assume that the data is located at the centroid of the district. This approach has been used by (Giorgi et al., 2018) in modelling historical malaria data in Cameroon. By denoting the centroid of district i as x_i and t as t_i, the spatio-temporal geostatistical (STG) model can be modelled as $\log(\lambda(x_i, t_i)) = d(x_i, t_i)^T\beta + S(x_i, t_i) + Z(x_i, t_i)$, where $Z(x_i, t_i) \sim (0, \tau^2)$ is added to Eq. 3.1 to account for non-spatial random effect and spatial variation on a scale smaller than the minimum observed distance between any two districts. The unobserved spatio-temporal random effect $S(x_i, t_i)$ is modelled as a Gaussian process with mean zero and the spatio-temporal covariance function given as:

$$Cov(S(x,t), S(x',t')) = \sigma^2\rho(x, x^{'}, t, t^{'}; \vartheta),$$

where $\rho(x, x^{'}, t, t^{'}; \vartheta)$ is a correlation function with a vector of parameter ϑ that regulates the scale of the spatial and temporal correlation. By assuming that $S(x,t)$ is a stationary and isotropic process as well as a separable process, we have:

$$\rho(x, x^{'}, t, t^{'}; \vartheta) = \rho(u, v; \vartheta) = \rho_s(u; \phi)\rho_t(v; \psi) = \exp\{-u/\phi\} \exp\{-v/\psi\},$$

where $u = \|x - x^{'}\|$ is the Euclidean distance between the centroid of the districts; $v = \|t - t'\|$ is the Euclidean distance between the time points and $\rho(\cdot;\cdot)$ is exponential correlation function with parameter ϕ and ψ that regulates the rate at which the spatial and temporal correlation gets close to zero with increasing distance u and time v, respectively.

Model 4: Spatio-temporal spatially discrete approximation to log Gaussian Cox process (STSDALGCP)

Motivated by (Diggle et al., 2013); (Johnson et al., 2019) introduced the spatially discrete approximation to the Log-Gaussian Cox process (SDALGCP) as a method for analyzing spatially aggregated data. By incorporating the time dimension here, the observed count y_{it}, conditional on the unobserved stochastic process S_{it} as independent Poisson random variables with expectation $m_{it}\lambda_{it}$, such that:

$$\begin{aligned} log(\lambda(x,t)) &= \int_{R_{it}} w(x,t)\left\{d^T(x,t)\beta + S(x,t)\right\}\, dx && (3.2)\\ &= \int_{R_{it}} w(x,t)d^T(x,t)\beta dx + \int_{R_{it}} w(x,t)S(x,t)\, dx && (3.3)\\ &= d_{it}^T\beta^* + S_{it}^* && (3.4) \end{aligned}$$

where $w_{it}(x,t)$ is the population weight at i^{th} location and time t, β^* is a vector of regression coefficients for the aggregate explanatory variables d_{it} and S_{it}^* is a Gaussian process. We use a weighting function $w_{it}(x,t)$ to account for the heterogeneous distribution of disease risk within a district. It is computed as: $w_{it}(x,t) = m(x,t)/m_{it}$, for $m_{it} = \int_{R_{it}} m(x,t)dx$, if $m(x,t)$ is available and $w_{it}(x,t) = 1/\left|R_{it}\right|$ otherwise.

Then the joint distribution of $S^* = (S_1^*, \ldots, S_N^*)$ is multivariate Gaussian with zero mean and covariance function given as:

$$\mathbf{Cov}(S_{it}^*, S_{it}^{*\prime}) = \sigma^2 \int_{R_{it}} \int_{R_{it}^{'}} w(x,t)w(x,t)'\rho(||(x,t) - (x,t)'||;\vartheta)\, dxdx' \qquad (3.5)$$

The variance of $S^*(x,t)$ depends on the size and shape of R_{it}, and the larger regions leading to smaller variances. Assuming a separable exponential correlation function, the correlation function in Eq. 3.5 can be expressed as:

$$\rho(||(x,t) - (x,t)'||;\vartheta) = \exp(-||t-t'||/\psi)\exp(-||x-x'||/\phi) \qquad (3.6)$$

where $||.||$ is the Euclidean distance, ϕ is scale parameter and ψ is the temporal parameter. For the collection of all the random effects; $S_t^* = (S_{1t}^*, ..., S_{Nt}^*)$ at time t; the first-order autoregressive relationship defines the overall spatio-temporal covariance structure as; $S_t^* = \varphi S_{t-1}^* + W_t, -1 < \varphi < 1$, where the temporal innovation W_t is modelled as a multivariate Gaussian distribution.

3.3.1 Parameter estimation and Spatial Prediction

The parameter estimation and prediction methods for the models discussed in the previous section are as follows. Model 2 is fitted by a fully Bayesian framework using the Markov chain Monte Carlo (MCMC) sampling method. The main advantage of using Bayesian inference is that parameter estimation and prediction can be carried out together. we specify weakly informative flat priors for: $\tau_w^2 \sim inverse - gamma(a = 1, b = 0.01)$, $\rho \sim uniform[0, 1]$, $\alpha \sim uniform[0, 1]$. The weak priors are used to allow values of the parameters to be informed by the data. We computed the incidence of malaria using the samples of the posterior distribution of the parameters. We used the R package `CARBayesST` (Lee et al., 2018) for the analysis. For models 1, 3, and 4, we use Monte Carlo Maximum likelihood (MCML). Let S_{it} for $i = 1, \ldots, N$ be a random effects associated with Y_{it}. Assume that Y_{it} conditionally on S_{it} are mutually independent random variables, then the likelihood of the vector of parameters; $\theta = \{\beta, \sigma^2, \phi, \psi, \tau^2\}$ is obtained from the marginal distribution by integrating out random effect as:

$$L(\theta) = [y|\theta] = \int [S, y|\theta] \, dS \tag{3.7}$$

where [.] is a shorthand notation for 'the distribution of'. The high dimension integral in Eq. 3.7 cannot be solved analytically. We approximate the integral using the Monte Carlo method. Taking initial parameters estimate θ_0 through Generalised linear model, variance

components using variogram; then estimate parameters using MCML method (Christensen, 2004) as:

$$\begin{aligned} L(\theta) &= \int [S, y|\theta] * \frac{[S, y|\theta_0]}{[S, y|\theta_0]}, dS \\ &= \int \frac{[Sy|\theta]}{[S, y|\theta_0]} * [S, y|\theta_0], dS \\ &\propto \int \frac{[S, y|\theta]}{[S, y|\theta_0]} * [S|y; \theta_0], dS \\ &= E\left\{ \frac{[S, y|\theta]}{[S, y|\theta_0]} \right\} \end{aligned}$$

We then generate B samples from $[S, y|\theta_0]$ say s_i and approximate the integrals as:

$$L_B(\theta) = \frac{1}{B} \sum_{b=1}^{B} \frac{[s_{(i)}|y, \theta]}{[s_{(i)}|y, \theta_0]}$$

Finally, we maximize $L_B(\theta)$ with respect to θ to obtain an estimate $\hat{\theta}_B$. We then set $(\hat{\theta}_B = \theta_0)$ and repeat the simulation until convergence. To reduce the computational burden in model 4, we restrict the maximization to a finite set of predefined values for ϕ. For a detailed description of the methods, see (Giorgi et al., 2018; Johnson et al., 2019). We use R package `PrevMap` (Giorgi and Diggle, 2017) for Model 1 and 3; and `SDALGCP` for model 4 (Johnson et al., 2018).

We have used plug-in prediction for spatio-temporal discrete or continuous predictions for models 1, 3, and 4, replacing the unknown parameter with their MCML estimates. Before fitting each model, the selection of appropriate variables from a complete list of variables is essential. We first considered the correlation matrix between covariates to select candidate variables and then also used AIC to select variables from the list initially in the GLM model. It compares different models by balancing underfitting (including only a few variables in the model) and overfitting (including many variables in the model). Including too few variables

often fails to capture the true relation and too many variables create an over-fitting problem. Thus, a trade-off between simplicity and adequacy of model fitting is therefore required and AIC can help to achieve this. Then, we used the selected variables to fit each of the candidate models from M[1-4].

3.3.2 Model comparison

We split the data set into five validation sets to evaluate the models' ability to predict incidence and quantify associated uncertainty. Here, we divided the data set into five separate, independent groups at random, each having 30 districts, with the fifth group holding 29 districts, for a total of 149 districts. Then, using the remaining 4 groups as a training set, we fit the model and assess how well it predicts the outcomes for the hold-out group. Root-mean-square-error (RMSE), mean absolute error (MAE), and 95% coverage probabilities (CP), which are derived from the anticipated incidence of malaria on the validation set, were compared with the predictions to describe the performance of the model.

These performance metrics, root-mean-square error (RMSE), mean-absolute-error (MAE), and 95% coverage probability (CP) for each set are calculated as follows:

$$RMSE_t = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\lambda_{it}^{emp} - \hat{\lambda}_{it}\right)^2}$$

$$MAE_t = \frac{1}{k}\sum_{i=1}^{k}\left|\lambda_{it}^{emp} - \hat{\lambda}_{it}\right|$$

$$CP_t = \frac{1}{k}\sum_{i=1}^{k} I\left(\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}\right),$$

where λ_{it}^{emp} is the true observed incidence of i^{th} district in the test set at time $t = 1, \ldots, 41$; $\hat{\lambda}_{it}$ is the predicted mean incidence; and $I\left(\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}\right)$ is an indicator function that

takes value 1 if $\hat{\lambda}_{it}^{0.025} < \lambda_{it}^{emp} < \hat{\lambda}_{it}^{0.975}$ and 0 otherwise, with $\hat{\lambda}_{it}^{0.025}$ and $\hat{\lambda}_{it}^{0.975}$ corresponding to the quantiles; 0.025 and 0.975 of the predictive or posterior distribution for λ_{it}, respectively and k is several districts in the test set. Finally, using data from June 2018 to May 2019 as a test set, we considered in-sample validation to compare M1 with the other three models. This is because M1 is fitted independently for every 149 districts, making it impossible for neighbouring districts to be predicted coherently. The performance of each model was then validated using the aforementioned matrices (RMSE, MAE, and CP).

3.4 Results

The map of the observed incidence rate of *P. falciparum* per 1000 population in southern Ethiopia for some selected months between 2016 to 2019 is shown in Figure 3.1. There is evidence of spatial correlation as there is a cluster of low incidence in the north and a high incidence in the south. This also means that improved prediction of the incidences can be achieved by applying a geostatistical model that exploits the correlations between the districts.

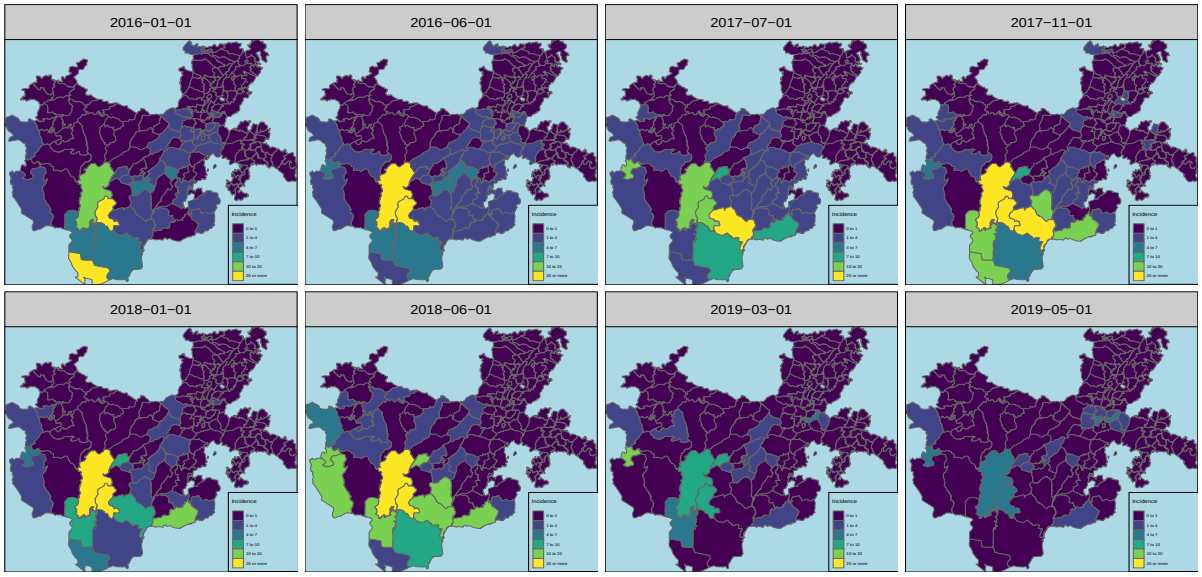

Figure 3.1 The observed incidence of *P. falciparum* per 1000 population for some selected months between 2016 and 2019 in Southern Ethiopia

The association between the malaria incidence of genus *P. falciparum* and the environmental and climatic predictors included are shown in Figure 3.1. We found that there is a mix of linear and non-linear relationships between the log of incidence of genus *P. falciparum* and the predictors as shown via the regression splines curve in Figure 3.2. The knots of each spline are then chosen by inspecting graphs for which the local maximum or minima were observed in the smoothed curve. Some notes on either end of the smoothed curve were removed. We also take the logarithm of the population density and precipitation to obtain a more linear relationship. We excluded maximum temperature, water vapor, precipitation, and elevation from the analysis after examining these variables for the presence of a "strong correlation" with other covariates.

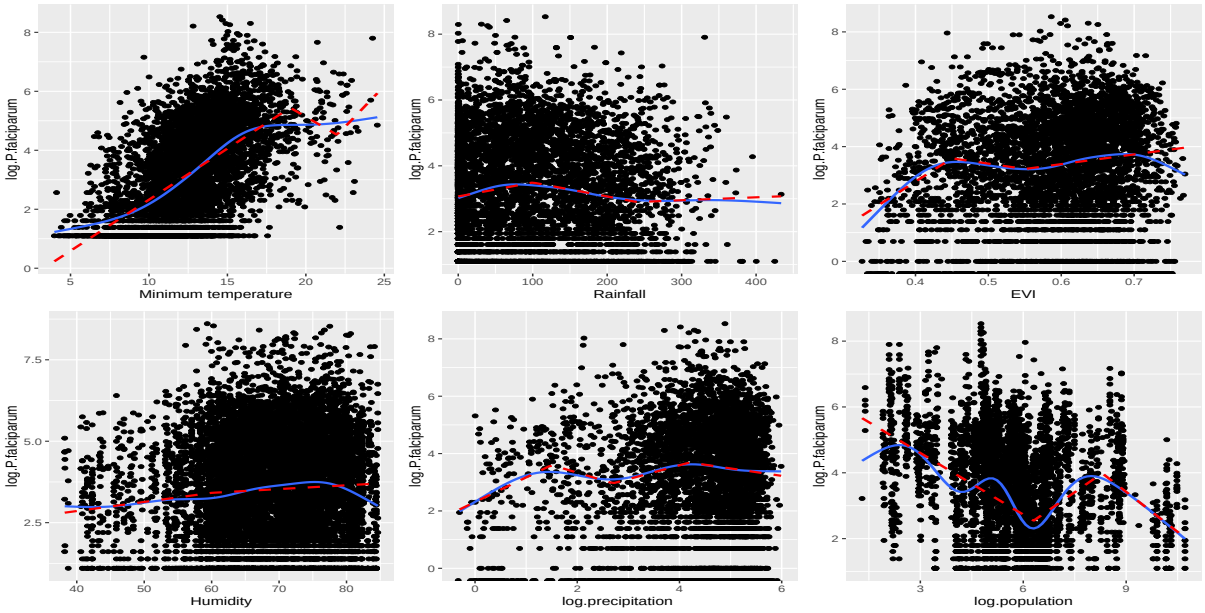

Figure 3.2 Scatter plot of the log. *P. falciparum* against temperature, Rainfall, log.precipitation, log.population density, Enhanced vegetation index (EVI), and Humidity. The solid blue line shows natural splines and the dashed red line shows linear splines

Table 3.1 Cross validation statistics for the Southern Ethiopia malaria count data using four models, M1-M4

Validation Measures	*STSM[M1]*	*STCAR[M2]*	*STG[M3]*	*STSDALGCP[M4]*
RMSE	0.028856141	0.008481092	0.01781151	0.016295230
MAE	0.0332017327	0.0031912349	0.02286517	0.012382210
CP	87.375838926	98.530201342	93.57494407	95.020134228

First, we investigated the significance of ignoring spatial correlation in disease risk mapping. We used data from July 2018 to May 2019 as a test set and another data set as a training set to identify models that better predict disease risk. The validation was then computed in the manner indicated in 3.1. The results show that 1) integrating spatial correlation in disease risk mapping is crucial, as all models consisting of spatial correlation had lower validation metrics such as RMSE when compared to a model that does not include spatial correlation. As it was indicated in Table. 3.1, model 4 (STSDALGCP) is a far better model than the others that include spatial correlation, with lower RMSE and MAE. Additionally, we provided the 95% CP, for which all models have CP values close to 95%.

Table 3.2 Cross validation statistics for the Southern Ethiopia malaria count data using three models, M2-M4

Validation Measures	*Set*	*STCAR[M2]*	*STG[M3]*	*STSDALGCP[M4]*
RMSE	1	0.004115	0.04148	0.03733
MAE		0.0017658	0.01986	0.01788
CP		71.95	73.82	81.20
RMSE	2	0.0024396	0.02873	0.02586
MAE		0.0012682	0.01863	0.01677
CP		79.51	79.35	87.28
RMSE	3	0.001904	0.01847	0.01662
MAE		0.0012565	0.009833	0.00885
CP		76.67	74.07	81.47
RMSE	4	0.0016115	0.01540	0.01386
MAE		0.0011072	0.008867	0.00798
CP		83.09	79.033	86.9360
RMSE	5	0.003659	0.037	0.0333
MAE		0.0015753	0.01362	0.01226
CP		81.67	76.79	84.47

Here, we provide an applied comparison of three proposals, spatio-temporal: CAR, geo-statistics, and discrete approximation to log Gaussian Cox process models, similar to the one proposed by (Wong et al., 2023). We demonstrated the three methods by estimating the incidence of malaria in southern Ethiopia. On these data, we compare the performance of the three approaches in terms of accuracy and computation time. According to the results of

the study, Model 2 (STCAR) has a very low validation matrix, RMSE, and MAE, as shown in Table. 3.2. This could be because spatial random effects can be introduced into these models using conditional autoregressive (CAR) priors, as detailed in (Besag et al., 1991; Leroux et al., 2000), by locally smoothing the risk by borrowing data from nearby regions thereby reducing random noise. Excessive smoothing, on the other hand, could hinder the detection of high-risk regions or hot spots since the discontinuities in the smooth risk become obscured.

Model 3 (STSDALGCP) on the other hand, is a considerably superior model with comparable lower RMSE compared to other models and a coverage probability closer to 95%. Also, STSDALGCP shows a higher coverage proportion which is closer to 95% followed by STG and STSM models. The coverage probabilities of STCAR (M2) is closer to 99% which might be due to over-smoothing. In terms of computational time, the STCAR model outperforms the others, closely followed by STSDALGCP. In terms of computational time, the STG model is the worst of the three.

Table. 3.1 and 3.2 show the relationship between the observed and predicted incidence for the four models. The result shows that there is a good agreement between the observed and the predicted incidence.

The fixed effect of our model has the following structure;

$$\mu_{ij} = \beta_0 + \beta_1 Rain.F_{it} + \beta_2 I\left\{(Rain.F_{it} - 100) * (Rain.F_{it} > 100)\right\} +$$

$$\beta_3 I\left\{(Rain.F_{it} - 300) * (Rain.F_{it} > 300)\right\} + \beta_4 Min.Temp_{it} +$$

$$\beta_5 Humidity_{it} + \beta_6 EVI_{it},$$

where Rain.F is the rainfall, Min.temp is the minimum temperature and EVI is the enhanced vegetation index, while the random effect is modelled using the class of models described in the method section.

The parameter estimates and their corresponding 95% confidence interval for models 2-4 are shown in Table 3.3. To obtain these results, we run 50,000 iterations of the MCMC and MCML algorithm with a burn-in of 10,000 samples and then retain every 8^{th} sample to reduce the autocorrelation of the Markov chain, resulting in 5000 samples for inference. For the STSDALGCP, We discretize ϕ using 1000 equally spaced values between 30,000 and 40,000.

We found that rainfall below 100 millimetres (mm) is associated with 0.4 % [0.2, 0.6] increased risk of malaria; rainfall between 100 mm and 300 mm is associated with 0.4 % decreased risk of malaria; and rainfall greater than 300mm is associated with 1.5% increase in the incidence of malaria. This indicates that malaria incidence in the country is non-linearly related to rainfall. Since rainfall increases up to 100 mm, the incidence increases and then slightly decreases when rainfall increases above 100mm. Then after, it starts to increase showing a non-linear pattern. This might be because, after heavy rainfall, favourable conditions will be created for mosquito breeding in the month from October to December- the higher incidence season and after moderate rainfall from March to April- the second higher incidence time

which coincides with (EFDR, 2019) report. High values of rainfall are positively associated with the incidence of malaria in the area.

Similarly, temperature and humidity are associated with an increase in malaria incidence whereas, EVI is negatively associated with an increase in malaria risk, i.e. an increase in EVI is significantly associated with a decline in malaria risk. For the random effect parameters, the estimated scale parameters for all models are significant. In particular, for the STSDALGCP model, the estimated practical range is 210 (206.937, 222.432) in km, which indicates that observations at about 210 km apart still correlate about 0.05; for the STG model, the practical range is 213.228 (213.196, 213.261) in km and temporal range is around 8 months.

Table 3.3 Parameter estimates of the models and their 95% CI based on the STCAR [M2], spatio-temporal geostatistical model [M3], and spatio-temporal spatially discrete approximation to log-Gaussian Cox process (STSDALGCP)[M4]

Model	*STCAR[M2]*	*STG[M3]*	*STSDALGCP[M4]*
β_0	-9.229 (-9.824, -8.769)	1.106 (-0.252, 2.463)	-4.593 (-5.319, -3.868)
β_1	0.005 (0.001, 0.007)	0.002 (0.000, 0.003)	0.004 (0.002, 0.006)
β_2	-0.006 (-0.019, 0.000)	-0.005 (-0.007, -0.002)	-0.008 (-0.011, -0.005)
β_3	0.026 (0.013, 0.083)	0.030 (0.005, 0.056)	0.018 (0.008, 0.028)
β_4	0.035 (0.019, 0.053)	0.045 (0.008, 0.083)	0.106 (0.089, 0.122)
β_5	0.002 (0.001, 0.005)	0.013 (0.008, 0.018)	0.012 (0.007, 0.018)
β_6	-0.011 (-0.435, 0.441)	-1.041 (-1.551, -0.532)	-3.484 (-4.013, -2.954)
σ^2		5.638 (5.569, 5.707)	20.689 (20.620, 20.759)
ϕ		71.076 (71.065, 71.087)	70.000 (68.979, 74.144)
ψ		8.400 (8.395, 8.405)	0.300 (0.259 0.341)
τ^2	0.845 (0.771, 0.922)	0.813 (0.255, 1.371)	
ξ_S	0.879 (0.855, 0.901)		
ξ_T	0.897 (0.881, 0.912)		
τ_w^2	291.540 (250.359, 340.146)		

The prediction maps for selected months of the year by all models are shown in Figure 3.3. The map shows the variability of malaria incidence across spaces and time, with a higher incidence occurring in December 2017 (which is among the highest malaria incidence seasons in the country), and the higher incidence was observed in south-western districts of the region including Konso, Hamer, Surma, and others. The incidence of malaria is seasonal and our finding agrees with (EFDR, 2019) report and it varies in space and time for which the higher incidence was observed after the heavier rainy season from October to December.

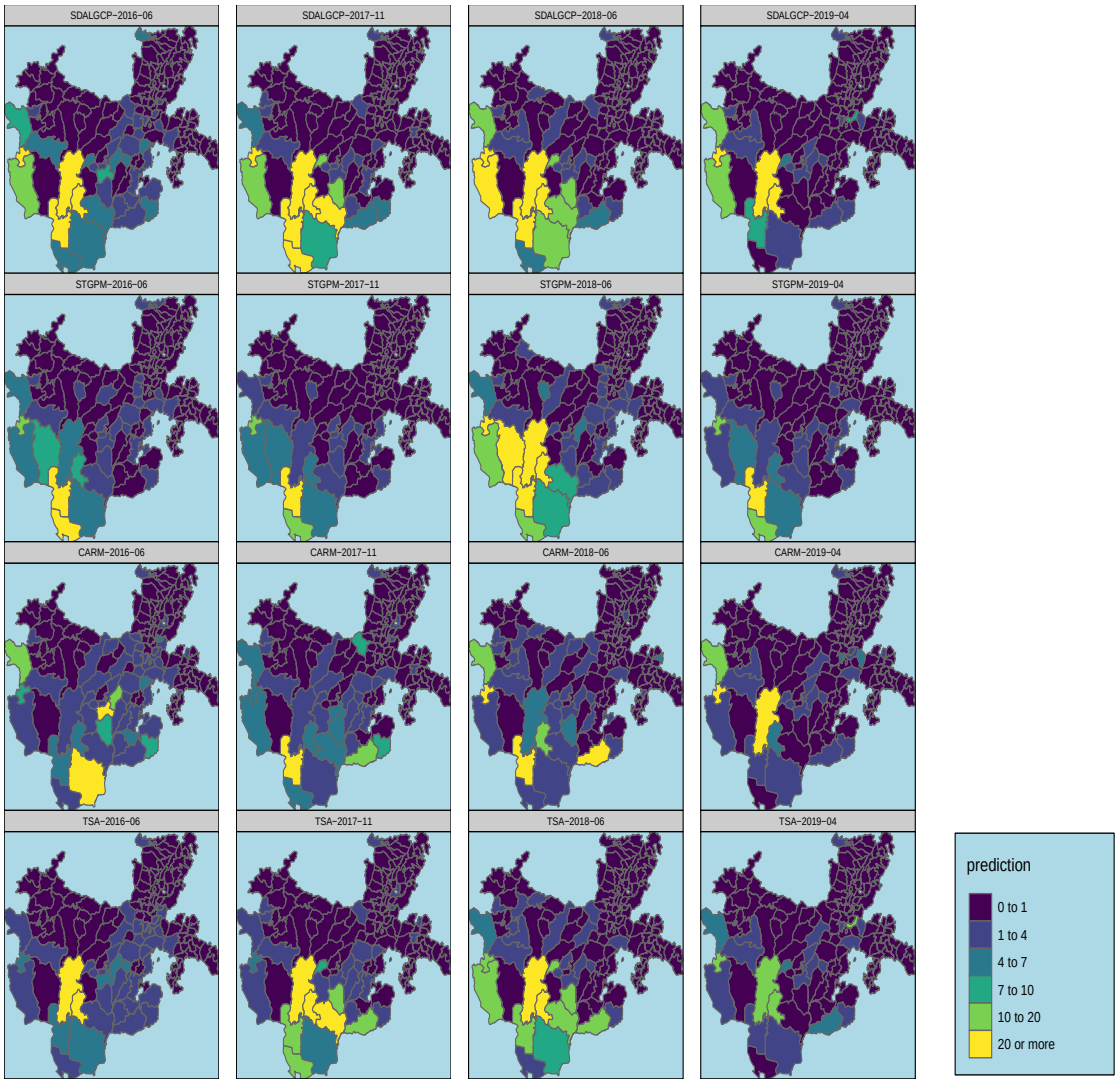

Figure 3.3 The predicted incidence of textitP. falciparum per 1000 population for four selected months using four models M[1:4]. The months are from January 2016 to May 2019

3.5 Discussion

In this study, we developed four spatial models to predict the incidence of malaria in Southern Ethiopia and compared the predictive performance of the models. While all models are reliable in accounting for the uncertainties in malaria risk, our results show that the spatio-temporal patterns of malaria incidence are better captured by the STSDALGCP model and this model can be used to identify areas characterized by unusual risks and to propose future control and intervention. STSDALGCP is not only better at predicting uncertainties but also provides smaller prediction errors in comparison to the other three models. The STSDALGCP and STG give a similar prediction of malaria incidence which could be because of the close connection between LGCP and geostatistical models as discussed in (Diggle et al., 2013).

The STCAR model has a lower prediction error and a coverage probability closer to 99%, predicting perfectly the observed incidence. These models typically include spatial random effects with conditional autoregressive (CAR) priors by smoothing the risk locally via borrowing information from nearby districts. However, extensive smoothing could hinder the detection of high-risk districts or hot spots. Despite the lower size of the validation metrics, this strategy does not appear to be a good choice for obtaining statistically precise estimates and finding high/low-risk areas while retaining geographical resolution utilizing administrative districts in Southern Ethiopia (Waller and Carlin, 2010).

Although all models are reliable in predicting the uncertainties, spatial time series models are poor in incorporating the uncertainty as they fail to capture important dependencies between variables, regions, and time simultaneously. This implies that including the spatial correlation is important in developing predictive models in space and time and ignoring it

can lead to higher prediction errors or consequent underestimation of the sizes of standard errors (Hoeting et al., 2006). A series of papers by (Gelfand et al., 2010) advocated that accounting for spatial correlation will give a better assessment of the risk factors, produce more accurate maps, and allow for an honest assessment of uncertainty.

The malaria incidence showed different distributional patterns during the study period. Figure 3.3 shows that high incidences of malaria were observed in districts located in the southwest and center regions of the country, indicating that these regions should remain foci for active malaria intervention. Climatic variables play a significant role in the intensity of the transmission in the country (Abeku et al., 2004; Teklehaimanot et al., 2004; Rodo et al., 2021; Girum et al., 2021), and the level and nature of the transmission vary across space and time. The inclusion of these variables improved the prediction performance of the model which is consistent with some of the previous works (Colborn et al., 2018; Giorgi et al., 2021; Bhatt et al., 2017). Our results support the scientific evidence that an increased level of rainfall, temperature, and humidity are associated with an increased risk of malaria because these factors create a favorable environment for mosquitoes to breed (Girum et al., 2021; Bhatt et al., 2017; Giorgi et al., 2021) whereas EVI is negatively associated with increased risk of malaria in the area supported by the study (McMahon et al., 2021). Yu et al. (2015) and (Giorgi et al., 2021) found that in most parts of Africa, as temperature, rainfall, and humidity increase, malaria incidence increases. However, the relationship is not always linear (Teklehaimanot et al., 2004; Rodo et al., 2021; Giorgi et al., 2021) and linear splines can be used to account for the non-linear relationship and to describe the relationship in an easily understandable form.

One way of extending this analysis would be to use a multivariate spatiotemporal LGCP

to jointly model the counts of multiple species like *p.vivax* and others by fitting joint models. This will allow us to borrow information across multiple species to improve prediction. Furthermore, future studies will improve the model by accounting for more informative risk factors that are not captured in the study, for example, literacy level, income, and distance from the water body. While the model performs fairly well at predicting district-level malaria incidences, its predictive accuracy can further be improved.

The study is without some limitations. One is the lack of important socio-economic predictors that would have improved the malaria incidence prediction. Two is the problem of underreporting which is a common problem in surveillance data in a low-resource setting. Three, aggregating the malaria cases at the district level can further increase the bias. Lastly, the susceptible population projection is an estimate from the demographic census that assumes the population is constant over ten years. This is because the census is carried out every 10 years.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data accessibility

The Malaria data sets considered for the study are available based on the consultation with the Ethiopian Public Health Institute on reasonable request. The covariate information is freely available online under the specified website.

3.6 References

Abeku, T., Vlas, S.D., Borsboom, G., Tadege, A., Gebreyesus, Y., Gebreyohannes, H., Alamirew, D., Seifu, A., Nagelkerke, N., Habbema, J., 2004. Effects of meteorological factors on epidemic malaria in ethiopia: a statistical modelling approach based on theoretical reasoning. Parasitology 128, 585–593. URL: https://researchonline.lshtm.ac.uk/id/eprint/14632/.

Anderson, C., Lee, D., Dean, N., 2014. Identifying clusters in Bayesian disease mapping. Biostatistics 15, 457–469. URL: https://doi.org/10.1093/biostatistics/kxu005, doi:10.1093/biostatistics/kxu005.

Benjamin, M.T., Andrade-Pacheco, R., JWS., H., 2018. Continuous inference for aggregated point process data. Royal Statistical Society 181, 1125–1150. doi:10.1111/1467-9876.00113.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological) 36, 192–225.

Besag, J., York, J., Mollie, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43, 1–20. doi:10.1007/BF00116466.

Bhatt, S., Cameron, E., Flaxman, S., Weiss, D., Smith, D., Gething, P., 2017. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. J R Soc Interface 14. doi:10.1098/rsif.2017.0520.

Bhatt, S., Weiss, D., Cameron, E., et.al, 2015. The effect of malaria control on plasmodium falciparum in africa between 2000 and 2015. Nature 526, 207–211. doi:10.1038/nature15535.

Bivand, R., Gómez-Rubio, V., Rue, H., 2015. Spatial data analysis with r-inla with some extensions, American Statistical Association.

Christensen, O., 2004. Monte carlo maximum likelihood in model-based geostatistics. Computational and Graphical Statistics 13, 702–718.

Colborn, K.L., Giorgi, E., Monaghan, A.J., Gudo, E., Candrinho, B., Marrufo, T.J., Colborn, J.M., 2018. Spatio-temporal modelling of weekly malaria incidence in children under 5 for early epidemic detection in mozambique. Scientific reports 8, 1–9. doi:10.1038/s41598-018-27537-4.

CSA, 2007. Central statistical authority, 2007 population and housing census of ethiopia. country level. Addis Ababa, Ethiopia .

Dabaro, D., Birhanu, Z., Negash, A., Hawaria, D., Yewhalaw, D., 2021. Effects of rainfall, temperature and topography on malaria incidence in elimination targeted district of ethiopia. Malaria journal 20, 1–10.

Diggle, P., Giorgi, E., 2016. Model-based geostatistics for prevalence mapping in low-resource settings. American Statistical Association 111, 1096–1120. doi:10.1080/01621459.2015.1123158.

Diggle, P., Giorgi, E., 2019. Model-based Geostatistics for Global Public Health: Methods and Applications. Chapman & amp; Hall/CRC Interdisciplinary Statistics, Chapman and Hall/CRC Press.

Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M., 2013. Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. Statistical Science 28, 542–563. URL: http://www.jstor.org/stable/43288435.

EFDR, 2019. Ethiopian ministry of health report, malaria epidemiological profile doi:www.moh.gov.et/ejcc/en/malaria.

Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P., 2010. A handbook ofspatial statistics. Boca Raton: Chapman and Hall/CRC Press .

Giorgi, E., Diggle, P., 2017. PrevMap: An R package for prevalence mapping. Journal of Statistical Software 78, 1–29. doi:10.18637/jss.v078.i08.

Giorgi, E., Fronterre, C., Macharia, P.M., Snow, V.A.A.R.W., Diggle, P., 2021. Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict. J. R. Soc. Interface 18. doi:10.1098/rsif.2021.0104.

Giorgi, E., P. Diggle, R.W.S., Noor, A.M., 2018. Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys. International Statistical Review 86, 571–597. doi:10.1111/insr.12268.

Girond, F., Randrianasolo, L., Randriamampionona, L., Rakotomanana, F., Randrianarivelojosia, M., Ratsitorahina, M., Brou, T.Y., Herbreteau, V., Mangeas, M., Zigiumugabe, S., et al., 2017. Analysing trends and forecasting malaria epidemics in madagascar using a sentinel surveillance network: a web-based application. Malaria journal 16, 1–11.

Girum, T., Shumbej, T., Shewangizaw, M., 2021. Burden of malaria in ethiopia, 2000-2016: findings from the global health estimates 2016. Trop Dis Travel Med Vaccines 5. doi:10.1186/s40794-019-0090-z.

Gneiting, T., Guttorp, P., 2010. Continuous parameter spatio-temporal processes, in handbook of spatial statistics (a. e. gelfand, p. j. diggle, m. fuentes and p. guttorp, eds.). CRC Press: Boca Raton, FL , 427–436.

Hoeting, J.A., Merton, R.A.D.A.A., Thompson, S.E., 2006. Model selection for geostatistical models. Ecological Applications 16, 87–98.

Illian, J.B., Sørbye, S.H., Rue, H., 2012. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). The annals of applied statistics 6, 1499–1530.

Joe, H., 2008. Accuracy of laplace approximation for discrete response mixed models 52, 5066–5074.

Johnson, O., Giorgi, E., Diggle, P., 2018. Sdalgcp: Spatially discrete approximation to log-gaussian cox processes for aggregated disease count data doi:CRAN.R-project.org/package-SDALGCP.

Johnson, O., Giorgi, E., Diggle, P., 2019. A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. Statistics in Medicine 38, 4871–4887. doi:10.1002/sim.8339.

Kitawa, Y., Asfaw, Z., 2023. Space-time modeling of monthly malaria incidence for seasonal associated drivers and early epidemic detection in southern ethiopia .

Konstantinoudis, G., Schuhmacher, D., Rue, H., Spycher, B.D., 2020. Discrete versus continuous domain models for disease mapping. Spatial and spatio-temporal epidemiology 32, 100319. doi:10.1016/j.sste.2019.100319.

Lee, D., 2011. A comparison of conditional autoregressive models used in bayesian disease mapping. Spatial and spatio-temporal epidemiology 2, 79–89.

Lee, D., Rushworth, A., Napier, G., 2018. Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the CARBayesST package. Journal of Statistical Software 84, 1–39–350. doi:10.18637/jss.v084.i09.

Leroux, B.G., Lei, X., Breslow, N., 2000. Statistical models in epidemiology, the environment, and clinical trials, chapter estimation of disease rates in small areas: A new mixed model for spatial dependence. Springer-Verlag, New York , 179–191.

Li, Y., Brown, P., Rue, H., al-Maini, M., Fortin, P., 2012a. Spatial modelling of lupus incidence over 40 years with changes in census areas. Journal of the Royal Statistical Society Series C 61, 99–115. doi:j.1467-9876.2011.01004.x.

Li, Y., Brown, P., Rue, H., al Maini, M., Fortin, P., 2012b. Spatial modelling of lupus incidence over 40 years with changes in census areas. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61, 99–115.

Martínez-Beneito, M., López-Quílez, A., Botella-Rocamora, P., 2008. An autoregressive approach to spatio-temporal disease mapping. Statistics in medicine 27, 2874–89.

McMahon, A., Mihretie, A., Ahmed, A.A., Lake, M., Awoke, W., Wimberly, M.C., 2021. Remote sensing of environmental risk factors for malaria in different geographic contexts. International journal of health geographics 20, 1–15.

Midekisa, A., Senay, G., Henebry, G., Semuniguse, P., Wimberly, Michael, C., 2012. Remote sensing-based time series models for malaria early warning in the highlands of ethiopia. Malar J 11. doi:10.1186/1475-2875-11-165.

Paige, J., Fuglstad, G.A., Riebler, A., Wakefield, J., 2020. Design- and model-based approaches to small-area estimation in a low- and middle-income country context: Comparisons and recommendations. Survey Statistics and Methodology URL: https://doi.org/10..1093/jssam/smaa011, doi:10.1093/jssam/smaa011.

Rodo, X., Martinez, P., Siraj, A., Pascual, M., 2021. Malaria trends in ethiopian highlands track the 2000 'slowdown' in global warming. Nat Commun 12. doi:10.1038/s41467-021-21815-y.

Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater london. Spatial and spatio-temporal epidemiology 10, 29—38. URL: https://doi.org/10.1016/j.sste.2014.05.001, doi:10.1016/j.sste.2014.05.001.

Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatio-temporal smoothing model for estimating trends and step changes in disease risk. Royal Statistical Society C 66, 141–157.

Taffese, H., Hemming-Schroeder, E., Koepfli, C., Tesfaye, G., Lee, M., Kazura, J., G.Y, Y., Zhou, G., 2018. Malaria epidemiology and interventions in ethiopia from 2001 to 2016. Infect Dis Poverty 7. doi:10.1186/s40249-018-0487-3.

Taylor, B.M., Davies, T.M., Rowlingson, B.S., Diggle, P.J., 2015. Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. Journal of Statistical Software 63, 1–48.

Teklehaimanot, H.D., Lipsitch, M., Teklehaimanot, A., Schwartz, J., 2004. Weather-based prediction of plasmodium falciparum malaria in epidemic-prone regions of ethiopia i.

patterns of lagged weather effects reflect biological mechanisms. Malaria Journal 3. doi:10.1186/1475-2875-3-41.

Utazi, C., Nilsen, K., Pannell, O., Dotse-Gborgbortsi, W., A.J.Tatem, 2021. District-levelestimation of vaccination coverage: Discrete vs continuous spatial models. Statistics in Medicine 40, 2197–2211.

Wakefield, J., 2007. Disease mapping and spatial regression with count data. Biostatistics 8, 158–83. doi:10.1093/biostatistics/kxl008.PMID:16809429.

Waller, L.A., Carlin, B.P., 2010. Disease mapping. Chapman & Hall/CRC handbooks of modern statistical methods 2010, 217.

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

WHO., 2019. World malaria report 2019. Geneva: World Health Organization doi:www.who.int/malaria/publications/world-malaria-report-2019.

WHO., 2020. World malaria report 2020: 20 years of global progress and challenges doi:apps.who.int/iris/handle/10665/337660.

Wong, S., Flegg, J.A., Golding, N., Kandanaarachchi, S., 2023. Comparison of new computational methods for geostatistical modelling of malaria. arXiv preprint arXiv:2305.01907 .

Yeshiwondim, A., Gopal, S., Hailemariam, A., Dengela, D., Patel, H., 2009. Spatial analysis of malaria incidence at the village level in areas with unstable transmission in ethiopia. Int J Health Geogr 8. doi:10.1186/1476-072X-8-5.

Yu, W., Mengersen, K., Dale, P., Ye, X., Guo, Y., Turner, L., Wang, X., Bi, Y., McBride, W.J., Mackenzie, J.S., et al., 2015. Projecting future transmission of malaria under climate change scenarios: challenges and research needs. Critical Reviews in Environmental Science and Technology 45, 777–811.

Appendix A: Discrete approximation to log Gaussian Cox process

We approximate Eq. 3.5 as a discrete sum over L_i and L_j randomly selected points in R_i and R_j to at time t give as:

$$\int_{R_{it}} \int_{R_{jt}} w_{it}(x,t) w_{jt}(x^{'}, t^{'}) \rho(||(x,t) - (x,t)'||; \vartheta)$$

$$\approx \frac{\sum_{k=1}^{L_{i,t}} \sum_{k=1}^{L_{j,t}} w_{it}(x_k, t) w_{jt}(x_k^{'}, t^{'}) \rho(||(x,t) - (x,t)'||; \vartheta)}{\sum_{k=1}^{L_{j,t}} \sum_{k=1}^{L_{j,t}} w_{it}(x_k, t) w_{jt}(x_k^{'}, t^{'})}$$

Where $w_{it}(x,t)$ is an equation with a positive coefficient and a domain of R_{it}, such that $\int_{R_{it}} w_{it}(x,t)dx = 1$. Then, by averaging its weight over R_{it}, we can roughly estimate the conditional log intensity of an LGCP as a separate constant. The weighting function $w_{it}(x,t)$is used to take into consideration the likelihood of a non-homogeneous variation in illness cases within a region R_{it}. The simplest choice for $w_{it}(x,t)$, assuming m(x,t) is available, would be to assign this equal to $\frac{m(x,t)}{m_i}$ with $mi = \int_{R_i} m(x,t)dx$. For instance, a greater number of incidents may cluster in more densely populated districts. A practical solution would be to determine $w_{it}(x,t) = \frac{1}{|R_{i,t}|}$ if m(x,t) is absent. We propose to draw each of the $x_k t$ and $x_(kt)$ in the aforementioned equation using a class of inhibitory processes referred to as (Diggle et al., 2013) to get a decent spatial coverage of R_{it} and R_{it}.

Appendix B: Results of Time Series Model for Randomly Selected Districts

Table 3.4 Parameter estimates and the corresponding 95% CI based in Each Districts using Time series model [M4]

Parameters	*point Estimate*	*Lower 95% CI*	*Upper 95% CI*
Kebena			
Intercept	-11.934	-14.354	-9.514
Rain.F	0.043	0.037	0.050
$I((Rain.F-100)*(Rain.F>100))$	-0.004	-0.013	0.005
$I((Rain.F-300)*(Rain.F>300))$	0.025	0.011	0.039
MIT	0.391	0.258	0.525
Humidity	0.027	0.013	0.041
EVI	-5.229	-7.276	-3.182
σ^2	1.880	0.900	2.860
ψ	8.382	8.284	8.479
Abeshege			
Intercept	-4.136	-8.486	0.215
Rain.F	0.082	0.074	0.090
$I((Rain.F-100)*(Rain.F>100))$	-0.012	-0.022	-0.001
$I((Rain.F-300)*(Rain.F>300))$	0.037	-0.149	0.222
MIT	0.284	0.076	0.491
Humidity	0.127	0.087	0.167
EVI	0.846	-0.505	2.196
σ^2	0.968	0.009	1.926
ψ	4.802	4.526	5.078
Ezha			
Intercept	-10.624	-17.499	-3.749
Rain.F	0.020	0.004	0.037
$I((Rain.F-100)*(Rain.F>100))$	-0.015	-0.044	0.015
$I((Rain.F-300)*(Rain.F>300))$	0.097	0.040	0.154
MIT	0.477	0.055	0.899
Humidity	0.037	-0.025	0.100
EVI	-4.016	-7.940	-0.093
σ^2	1.725	0.888	2.562
ψ	5.321	4.789	5.853
Kokir Gedebano			
Intercept	-3.392	-12.164	5.379
Rain.F	0.024	0.008	0.041
$I((Rain.F-100)*(Rain.F>100))$	-0.018	-0.042	0.006
$I((Rain.F-300)*(Rain.F>300))$	0.222	0.166	0.277
MIT	0.124	-0.285	0.534
Humidity	0.041	0.017	0.065
EVI	3.931	-7.223	- 0.640
σ^2	4.756	4.494	5.018
ψ	7.283	7.096	7.471
SODO WEREDA			
Intercept	-5.915	-11.598	-0.232
Rain.F	0.011	0.000	0.022
$I((Rain.F-100)*(Rain.F>100))$	-0.024	-0.041	-0.006
$I((Rain.F-300)*(Rain.F>300))$	0.084	0.042	0.126
MIT	0.312	0.049	0.575
Humidity	0.117	0.081	0.153
EVI	- 2.062	-4.047	-0.077
σ^2	1.400	0.725	2.074
ψ	3.884	3.600	4.168

Parameters	*point Estimate*	*Lower 95% CI*	*Upper 95% CI*
Meskan			
Intercept	-11.109	-16.755	-5.463
Rain.F	0.021	0.011	0.030
$I((Rain.F-100)*(Rain.F>100))$	-0.014	-0.029	0.000
$I((Rain.F-300)*(Rain.F>300))$	0.224	0.145	0.303
MIT	0.425	0.172	678
Humidity	0.010	-0.020	0.040
EVI	- 2.959	-4.685	-1.233
σ^2	1.654	0.888	2.420
ψ	6.370	6.147	6.593
Mareqo			
Intercept	19.935	13.318	26.553
Rain.F	0.032	0.018	0.046
$I((Rain.F-100)*(Rain.F>100))$	-0.022	-0.042	-0.002
$I((Rain.F-300)*(Rain.F>300))$	0.048	0.014	0.082
MIT	0.727	0.317	1.137
Humidity	0.017	-0.018	0.053
EVI	-2.413	- 2.759	-2.067
σ^2	10.158	9.959	10.357
ψ	10.169	10.085	10.253
Endegagn			
Intercept	0.735	-2.879	4.349
Rain.F	0.028	0.017	0.038
$I((Rain.F-100)*(Rain.F>100))$	-0.132	-0.145	-0.119
$I((Rain.F-300)*(Rain.F>300))$	5.623	5.614	5.633
MIT	0.513	0.328	0.698
Humidity	-0.066	-0.094	-0.038
EVI	-4.027	-6.752	-1.301
σ^2	2.480	1.602	3.358
ψ	11.855	11.772	11.939
Gumer			
Intercept	-6.587	-7.234	-5.940
Rain.F	0.014	0.014	0.015
$I((Rain.F-100)*(Rain.F>100))$	0.009	-0.002	0.020
$I((Rain.F-300)*(Rain.F>300))$	0.167	0.067	0.267
MIT	0.877	0.853	0.900
Humidity	0.100	0.095	0.104
EVI	-1.638	-2.375	-0.901
σ^2	0.471	0.107	0.835
ψ	9.152128	9.060	9.244
Cheha			
Intercept	-10.010	-16.253	-3.766
Rain.F	0.015	0.006	0.023
$I((Rain.F-100)*(Rain.F>100))$	0.009	-0.002	0.020
$I((Rain.F-300)*(Rain.F>300))$	0.167	0.067	0.267
MIT	0.235	0.019	0.451
Humidity	0.106	0.070	0.142
EVI	-8.124	-14.042	- 2.206
σ^2	5.823	5.455	6.191
ψ	14.112	14.047	14.177

Paper III

Multivariate Spatio-temporal Modeling of Aggregated Malaria Count of Genus *P. falciparum* and *P. vivax*; A Case Study on malaria risk mapping in Southern Ethiopia

Yonas Shuke Kitawa[1*]: Department of Statistics, College of Natural and Computational Science, Hawassa University, Hawassa, Ethiopia
Zeytu Gashaw Asfaw[2]: Department of Bio-statistics and Epidemiology, School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia

CHAPTER 4. Multivariate Spatio-temporal Modeling of Aggregated Malaria Count of Genus *P. falciparum* and *P. vivax*; A Case Study on Malaria Risk Mapping in Southern Ethiopia

Abstract

Background: Although malaria incidence has substantially fallen sharply over the past few years, the rate of decline varies by district, time, and malaria type. Despite this turn-down, malaria remains a major public health threat in various districts of Ethiopia. Consequently, the present study is aimed at developing a predictive model that helps to identify the spatio-temporal variation in malaria risk by multiple *Plasmodium* species.

Methods: We propose a multivariate spatio-temporal Bayesian model to obtain a more coherent picture of the temporally varying spatial variation in disease risk. The spatial autocorrelation in such a data set is typically modelled by random effects assigning a conditional autoregressive prior distribution. However, the autocorrelation considered in such cases depends on a binary neighbourhood matrix specified through the border-sharing rule. Over here, we propose a graph-based optimization algorithm for estimating the neighbourhood matrix that merely represents the spatial correlation by exploring the areal units as the vertices of a graph and the neighbour relations as the series of edges. Furthermore, we used aggregated malaria count in southern Ethiopia from August 2013 to May 2019.

Results: We recognized that precipitation, temperature, and humidity are positively associated with the malaria threat in the area. On the other hand, enhanced vegetation index, nighttime light (NTL), and distance from coastal areas are negatively associated. Moreover, nonlinear relationships were observed between malaria incidence and precipitation, temperature, and NTL. Additionally, lagged effects of temperature and humidity have a significant effect on malaria risk by either species. More elevated risk of *P. falciparum* was observed following the rainy season and unstable transmission of *P. vivax* was observed in the area. Finally, *P. vivax* risks are less sensitive to environmental factors than that of *P. falciparum*.

Conclusion: The improved inference was gained by employing the proposed approach in comparison to the commonly used border-sharing rule. Additionally, different covariates are identified including delayed effects, and elevated risks of either of the cases were observed in districts found in the central and western regions. As, malaria transmission operates in a spatially continuous manner, a spatially continuous model should be employed when it is computationally feasible.

keywords: Disease mapping, MSTCAR, Graph-based optimization algorithm, *p.falciparum*, *p.vivax*, Waiting Matrix

4.1 Introduction

Malaria continues to be an enormous public health problem, especially in sub-Saharan Africa, despite being preventable and treatable (Bhatt et al., 2015; Weiss et al., 2019). It not only contributes to the disease burden but also admittedly has devastating social and economic consequences, especially in sensitive areas where resources are severely limited. The difficulty continues to inflict recently, with an estimated 241 million cases. In addition, 627 thousand deaths occurred worldwide in 2020, with 14 million more cases and 69,000 more deaths in 2020 than in 2019 (WHO, 2021). According to a (WHO, 2021) report, the incidence is increasing, particularly in Ethiopia, in 2019, which might be associated with the interruption of diagnosis and treatment during the pandemic. Like most infectious diseases, malaria risk is identified by spatial and temporal patterns (Kitawa et al., 2022; WHO, 2021), which could partly be associated with environmental, climatic, and human factors. Therefore, these elements are essential to predict spatio-temporal patterns of the incidence, identify hot spots, and enable cost-effective disease monitoring, control, and efficient allocation of limited resources (Colborn et al., 2018a; Kitawa et al., 2022). Social and economic factors in addition to environmental variables, also play a significant role in affecting the spatial distribution and effect of monitoring, as illustrated by (Chirombo et al., 2020).

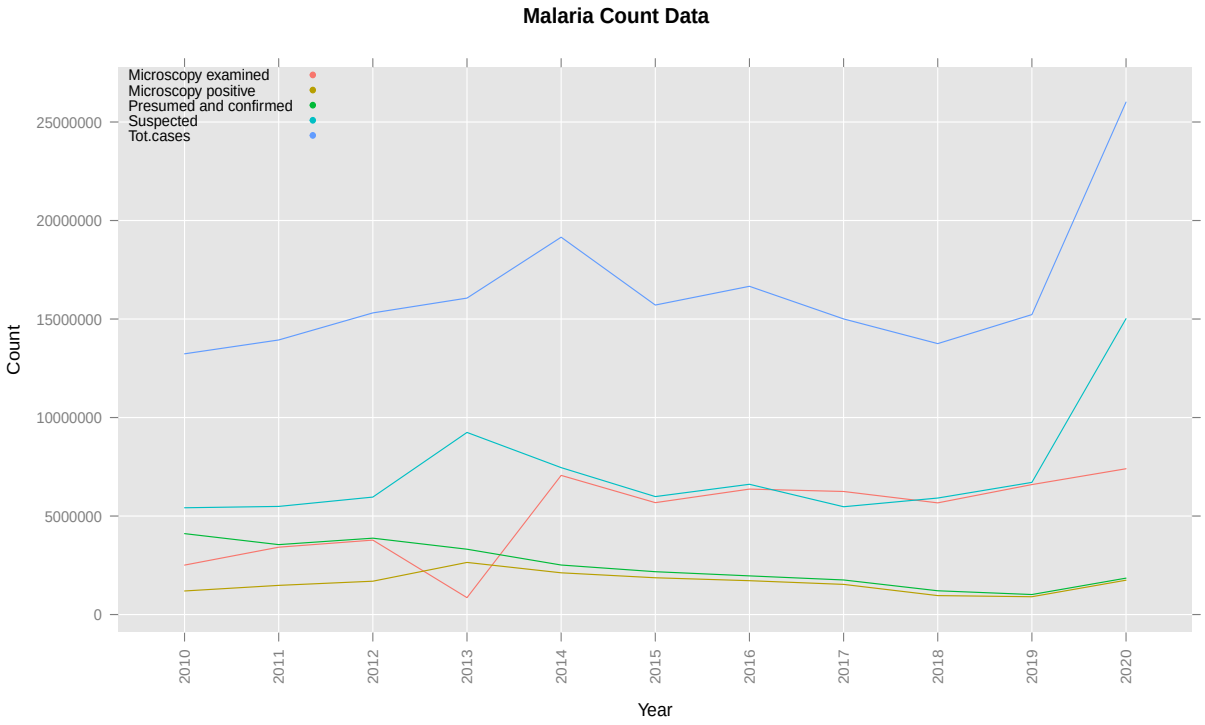

Figure 4.1 Suspected, tested, examined and confirmed counts of malaria in Ethiopia from 2010 to 2020 (WHO, 2021)

Many African countries, including Ethiopia, have previously reported significant success in reducing malaria incidence. Nothing but unfortunate incidents has re-emerged in some countries, such as Ethiopia, for the last two years, as shown in Fig. 1. The distribution of the incidence is rapidly changing, including areas characterized by lower disease risk (Kitawa et al., 2022; Kitawa and Asfaw, 2023). In this way, mapping such incidences using high-resolution risk maps is essentially needed, as it highlights districts with an unending spread that require urgent interventions (Cohen et al., 2017). Much of the current literature focuses on mapping such cases employing one of the *Plasmodium* species without devoting much

attention to other malarial cases. On the other hand, it is understood that interactions between various *Plasmodium* species associated with differences in environmental, climatic, and socioeconomic factors might underlie variation in the spread of the risk (Colborn et al., 2018b). Of these malaria species, *Plasmodium* causes significant morbidity and mortality from which *P. falciparum* and *P. vivax* are dominant in Ethiopia.

Unfortunately, there are no sufficient studies on the joint dynamics of these two dominant malaria species, i.e., *P. falciparum* and *P. vivax* in Ethiopia. The extant literature focuses on *P. falciparum* without reference to the potential effects of others and the corresponding confections while making predictive inferences about the incidence and identification of hot spots in various settings. Sometimes, it is recognized that most endemic regions are co-endemic for some or all of the other *Plasmodium* species (Colborn et al., 2018b). Individually, the spatial distribution of such threats has been extensively investigated (Kitawa et al., 2022; Kitawa and Asfaw, 2023; Midekisa et al., 2012), but the potential joint effect was sufficiently unincorporated in the case of malaria risk mapping by multiple *Plasmodium* species in the region. Thus, identifying malaria risk via joint modelling is important to point out different correlation structures across cases in diverse districts of the region over time. These can be incorporated through modelling such multiple spatially dependent variables by accounting for the correlation among the risks that helps to obtain a more accurate picture of the disease burden in areas.

The importance of joint modelling over univariate study is to estimate associations with the outcomes that are made impossible by independent analyses, illustrate correlations within and among different malaria risks, highlight areas where either of the threats is dominant, and obtain improved predictive inference by borrowing strength across space, time and dis-

ease through modelling their dependency (Quick et al., 2017; Palmí-Perales et al., 2021; Eyre et al., 2020). Furthermore, the results obtained from joint or multivariate analysis sometimes provide better precision for casual noise generated from multiple outcome variables, as it accounts for various sources of dependence (Eyre et al., 2021, 2020). A few studies have investigated the spatio-temporal distributions of malaria in some specific regions and districts of Ethiopia (Seyoum et al., 2017; Nigatu et al., 1992; Leonard et al., 2022; Deress and Girma, 2019). For instance,(Seyoum et al., 2017) recommended that malaria risk by *P. vivax* varies more spatially than that of *P. falciparum* in children. The author argues that the risk of *P. vivax* is less sensitive to environmental changes than that of *P. falciparum*, suggesting the importance of incorporating other factors to distinguish spatial-temporal non-uniformity. Another study suggests that biological makeup may also influence the spatial variation in disease risk (White, 2011). However, there is no sufficient study to explain the joint spatio-temporal distribution of malaria hazards by multiple *Plasmodium* species in the country.

To analyze such multiple outcome variables in a small area (SAE), several modelling alternatives are formulated depending on the data-generating process and computational issue. For example, (Taylor et al., 2015; Eyre et al., 2020) developed a multivariate model whose random effects are generated from a continuous Gaussian process. This approach is appealing because it considers distance while defining the spatial correlation between districts (Benjamin et al., 2018), even though it is computationally demanding. Additionally, incorporating cross-correlations for modelling continuous multivariate spatial processes is another challenge; in particular, their processing requirements frequently necessitate the adoption of less complex options (Lawson, 2018). Therefore, we could be interested in the vector of intensities that might, at least for simplicity, be approximated using a Markov random field

(MRF) specification (Lawson, 2018). Integrated Nested Laplace Approximation (INLA) is one such method that has been developed as a computationally effective alternative to MCMC (Blangiardo et al., 2013; Rue et al., 2009). Several multivariate spatial and spatio-temporal models have been developed, (Gómez-Rubio and Rue, 2018; Gómez-Rubio and Palmí-Perales, 2019; Gómez-Rubio et al., 2019; Palmí-Perales et al., 2021) to mention some in different resource settings using Laplace approximation via INLA. AS fitting techniques and results obtained by MCMC and INLA approaches are nearly identical (Vicente et al., 2020), one can consider either approach through balancing computation issues and precision of the estimates.

From such a different modelling framework, it is worth mentioning to look toward the conditional autoregressive (CAR) model (Besag, 1974; Besag et al., 1991), which was later extended to include its multivariate counterpart, the multivariate CAR (Gelfand and Vounatsou, 2003). Recently, (Quick et al., 2017) developed multivariate space-time CAR, which was later extended by (Lee et al., 2022) to accommodate second-order autocorrelation. In all those modelling perspectives, the spatial autocorrelation derived from the CAR model depends on a simple neighbourhood matrix ($\mathbf{W}$). However, the suitability of ($\mathbf{W}$) for the different data sets is rarely assessed in contrast to the geostatistical model that considers variograms to identify an appropriate spatial correlation (Diggle et al., 1998; Gneiting et al., 2006). Moreover, constructing ($\mathbf{W}$) using a simple border-sharing rule does not always give an appropriate autocorrelation structure for the data under study. This is because spatial autocorrelation is not consistently present throughout the study region; rather, there are pairs of neighbouring areas that exhibit substantial differences between their values.

To such an extent, graph-based nearest neighbour searches are gaining popularity in various

comparable studies (Lin and Zhao, 2019a). This approach outperforms other known simple border-sharing rules and is important in identifying autocorrelation (Enright et al., 2021; Lee et al., 2022). Several alternative approaches have been proposed to better capture spatial correlation in small area estimation (SAE) (Bivand et al., 2008). Among those alternatives, we consider the recently proposed graph-based optimization algorithm to estimate the appropriate neighbourhood matrix developed by (Lee et al., 2021). However, these estimates; (**WE**) and (**WEt**) obtained from graph-based optimization are specific to a single outcome variable, as proposed by (Lee et al., 2021). Some studies on the other hand claimed that the estimates obtained from two or more variables experiencing similar spatial patterns are similar (Jack et al., 2019). Based on these contexts, (**WE**) and (**WEt**) can be estimated by searching an initial graph after covariate effects have been accounted for. Then, the algorithm estimates (**WE**) and (**WEt**) by including or removing the edge following the distribution of the residuals. This method allows the estimated matrix to be static or vary dynamically over time t. Subsequently, we implement this approach to a model proposed by (Lee et al., 2022) using aggregated malaria data in Southern Ethiopia.

In this study, multivariate spatio-temporal models for malaria risk mapping were developed, aimed at the following:

- Quantifying malaria incidences using two dominant malaria species; *P. falciparum* and *P. vivax* in Southern Ethiopia informed by an aggregate count at the district level.
- Identifying the spatial correlation both within and across species and determining the relative distribution of each species in the region.
- To identify areas or districts where one of the incidences is significantly dominant and point out some of the factors associated with it.

The paper is arranged as follows. A description of the data set is specified in Section 2, A Multivariate spatio-temporal model for spatially aggregated data including a novel, graph-based optimization algorithm is described in Sect 3. Section 4 presents results applied to new malaria risks mapping using multiple *Plasmodium* specious in Southern Ethiopia, including model fit and inference in comparison to **(W)**. Sect. 5 discusses the results of the analysis, and Sec 6 concludes the results of the study.

4.2 Malaria data and the predictors

Throughout the paper, we use the reported monthly aggregated malaria count at the district level in Southern Ethiopia which is obtained from the Ethiopian Public Health Institute from August 2013 to May 2019. The data set consists of monthly aggregated malaria counts of genus *P. falciparium* and *P. vivax* from 149 districts in southern Ethiopia. The covariates included in the study are:

1. Population: a yearly population data set of each district was obtained from the regional demographic department and is projected based on the 2007 Ethiopian census data (CSA, 2007). Because it is the most recent census done in Ethiopia, (CSA, 2007) was used to project the population up until this point.

2. Temperature and precipitation - average lowest temperature (^0C), average maximum temperature (^0C), and total precipitation (mm) is obtained from meteorological and climate data supplied at 2.5 minutes or ($\sim 21km^2$) spatial resolution (https://www.worldclim.org/data/monthlywth.html

3. Relative Humidity: average relative humidity was derived from European Centre for Medium-Range Weather Forecasts (ECMWF) from ERA-Interim global atmospheric reanalysis.

4. Enhanced vegetation index (EVI) was obtained from a Moderate-resolution Imaging Spectroradiometer (MODIS) available at 1km2 spatial resolution (lpdaac.usgs.gov/ products/mod13a3v006/).

5. Nighttime light (NTL)- It is obtained from NOAA's National Centers for Environmental Information, Visible Infrared Imaging Radiometer Suite (https:// ngdc.noaa.gov/eog /viirs/index.html) which is available at a resolution of 3 arc (approximately 100m at the equator).

6. Distance from a coastal area (DCA)- is obtained from WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton), at a resolution of 3 arc (approximately 100m at the equator) for the whole 2000-2020 period.

7. Water vapor (kPa): is spatial data obtained from historical climatic data, WorldClim version 2.1 (https://www.worldclim.org/data/worldclim21.html) at the spatial resolutions of 30 seconds (1 km2).

8. Elevation: The elevation data is obtained from "http://www.diva-gis.org/Data" freely available for any country in the world consisting of administrative boundaries, roads, railroads, altitude, land cover, and population density.

All satellite data sets were aggregated to the district level by taking the average of all grid points failing inside each district.

4.3 Statistical Models for Spatially Aggregated Data

To answer our research question, we have considered the following modelling approach for multiple outcome variables.

4.3.1 Spatial autocorrelation

Due to geographical dependence, the data set may occasionally contain some information that is replicated among nearby places. In such occasions, spatial autocorrelation can be used to measure the degree to which a spatial random effect correlates with itself at various locations (Cressie, 1993). To test for the presence and strength of spatial association among areal units, we consider Moran's I, and Geary's C (Moran, 1950) which are global distance-based measures given as:

$$I = \frac{K \sum_k^K \sum_h^K w_{kh}(y_k - \bar{y})(y_h - \bar{y})}{(\sum_k^K \sum_h^K w_{kh})(\sum_k^K (y_k - \bar{y})^2}$$

where w_{kh} are components of a spatial weight matrix and y_k stands for the vector of observations made across K different locations. The Moran's I will be assessed by test statistic (the Moran's I standard deviate) that measures the statistical significance of spatial autocorrelation in model residuals.

Correlation- we computed the correlation coefficient to assess the strength of the link between two *Plasmodium* species. We first fit separate GLM models for *P. falciparum* and *P. vivax* by including covariates. Then, compute the correlation between two *Plasmodium* species at different time points to visualize the strength of the association between two incidents.

4.3.2 Spatio-temporal Multivariate CAR Model (MSTCAR)

Suppose Y_{itj} is set of aggregated malaria count associated with $i^{th}; [i = 1, \ldots, K]$ districts at time $[t; t = 1, \ldots, T]$ in months from August 2013 to may 2019 for the j^{th} outcome, $[j = 1, 2]$ representing malaria type by genus *P. falciparum* and *P. vivax*. We then model this outcome variable as conditionally independent Poisson distributions given as:

$$Y_{itj} \sim Poisson(m_{it} * \lambda_{itj})$$

$$log(\lambda_{itj}) = d_{it}^{\intercal}\beta_j + S_{itj} \tag{4.1}$$

Where, λ_{itj} is the true unknown incidence of malaria in districts i and time t of type j, S_{itj} is the spatio-temporal variation. d_{it} is a vector of spatio-temporal referenced explanatory variables with associated regression coefficient β for each outcome and m_{it} represents the overall population in each district at time t, which was used as an offset.

4.3.3 Lagged influence of climatic factors on the occurrence of malaria

The relationship between the occurrence of malaria and other environmental factors is nuanced, and different research has produced varied results. One of the discrepancies may be related to the exclusion of the delayed effect (Rotejanaprasert et al., 2021). To help public health authorities plan and allocate resources for the elimination of malaria more effectively, including lagged effect may increase the understanding of the relationship and predict changes in incidence (Gasparrini, 2014). The additional lag dimension of an exposure-incidence relationship, which describes the effect's time series, is then needed to model the link with environmental exposures. To account for the delayed impacts as a result, we fur-

ther extended the Eq.4.1 as:

$$log(\lambda_{itj}) = d_{it}^{\intercal}\beta_j + \sum_{l=1}^{4} d_{it-l}^{\intercal *}\beta_j^* + S_{itj} \tag{4.2}$$

Where: $l = 1, 2, 3, 4$ is the lagged time point considered for the vector of each lagged explanatory variable d_{it}^* with β_j^* is a regression coefficient associated with each lag covariate for j^{th} outcome. We consider lagged time till 4 months as the effect significantly declines beyond some time point.

Now, the question is how we model the random effect $[S_{itj}]$ in Eq. 4.2. Among various alternatives for modelling multivariate spatio-temporal random effect for aggregated count data, we consider a modelling approach by (Lee et al., 2022). We consider these approaches following extensive literature and exploratory analysis. This approach represents $[S_{ijt}]$ with a single set of random effects which is modelled as the joint outcome variable in space and time. These random effects, therefore, induce the spatial (auto) correlations in time, space, and between outcomes (Lee et al., 2018). The entire set of random effects is given as; $[\mathbf{S} = (S_1, S_2, \ldots, S_T)]$, where $[S_t = (S_{1t}, S_{2t}, \ldots, S_{it})]$ stands for the collection of $K \times J$ random effects at time t, while $[S_{it} = (S_{it1}, S_{it2}, \ldots, S_{itj})]$ indicates the subset of these effects for all J outcomes at the i^{th} district. Following the modelling approach by (Lee et al., 2022), $\mathbf{S}$ assumed to follow zero-mean multivariate Gaussian Markov random field gives as:

$$\mathbf{S} \sim \mathcal{N}\left(\mathbf{0}, \left[\mathbf{D}(\alpha) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right)$$

where $\otimes$ denotes a Kronecker product, $\mathbf{D}(\alpha) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}$ is the precision matrix from which; $\mathbf{D}(\alpha_{T\times T})$, $\mathbf{Q}(\mathbf{W}, \rho)_{K\times K}$ and $\mathbf{\Sigma}_{J\times J}$ are induced to control the temporal, the spatial and capture between outcome correlations respectively. The model is defined in terms of its precision matrix, rather than its covariance matrix which can in turn be described by three components.

- Between outcome correlation; $\mathbf{\Sigma}$ assigned conjugate Inverse-Wishart prior distribution; $\mathbf{\Sigma} \sim$ Inverse-Wishart $(d, \mathbf{\Omega})$. One can specify hyperparameters at $(df = \nu + J - 1, \mathbf{\Omega} = 2vdiag(1/A_1, ..., 1/A_J))$ with $diag(1/A_1, ..., 1/A_J)$ denotes a diagonal matrix with diagonal entries $(A_1, ..., A_J)$ represent the scale parameter for each outcome at ν degrees of freedom for the prior distribution, in case alternative prior distribution is used (Huang and Wand, 2013).

- Temporal autocorrelation; is modeled using first- or second-order autoregressive processes.

  1. First-order autoregressive process; the joint prior distribution $f(\mathbf{S})$ can be decomposed as:

$$f(\mathbf{S}) = f(S_1) \prod_{t=2}^{T} f(S_t|S_{t-1})$$

$$= N\left(\mathbf{S}_1|\mathbf{0}, \left[\mathbf{D}(\alpha) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right) \prod_{t=2}^{T} N\left(S_t|\alpha S_{t-1}, \left[\mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right)$$

  2. The second-order autoregressive process, the joint prior distribution $f(\mathbf{S})$ can be decomposed as:

$$f(\mathbf{S}) = f(S_1) f(S_2) \prod_{t=3}^{T} f\left(S_t|S_{t-1}, S_{t-2}\right)$$

$$= N\left(\mathbf{S}_1|\mathbf{0}, \left[\mathbf{D}(\alpha) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right) N\left(\mathbf{S}_2|\mathbf{0}, \left[\mathbf{D}(\alpha) \otimes \mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right)$$

$$\times \prod_{t=3}^{T} N\left(S_t|\alpha_1 S_{t-1} + \alpha_2 S_{t-2}, \left[\mathbf{Q}(\mathbf{W}, \rho) \otimes \mathbf{\Sigma}^{-1}\right]^{-1}\right)$$

- (S_{itj}, S_{isj}) are conditionally independent if; $s \notin (t-1, t, t+1)$ in AR(1) and $s \notin (t-2, t-1, t, t+1, t+2)$ in case of AR(2).

- Spatial autocorrelation; is modelled by a conditional autoregressive (CAR) prior after specifying $K \times K$ adjacency matrix $\mathbf{W} = w_{kh}$ that measures the degree of geographical proximity between each pair of districts (k,h). Then, $w_{kh} = 1$ if regions (k, h) have a common border and $w_{kh} = 0$ otherwise. These standard border-sharing approaches, may not always provide an appropriate autocorrelation structure, making it unlikely that defining **W** based on this rule will produce a valid autocorrelation structure. As an alternative, there are likely pairs of close-by areal units that exhibit notable differences in their data values. One of the best strategies in such circumstances is the estimation of **W** based on the graph-based optimization proposed by (Lee et al., 2021).

4.3.4 Graph-based optimisation for estimating WE and WEt

Before estimating **WE** and **WEt**, we compute the residual structure of the data set. The residual variations at time t after the effects of the covariates have been brought into account can be captured from Eq.4.2 as:

$$\hat{S_{itj}} = ln(Y_{itj}) - \left(d_{it}^{\intercal} \hat{\beta}_j + \sum_{l=1}^{4} d_{it-l}^{\intercal} \hat{\beta}_j^* \right) \tag{4.3}$$

Here, $\hat{\beta}_j$ and $\hat{\beta}_j^*$ from the conventional Poisson regression model without random effects can be estimated using the maximum likelihood method. Now, for S_itj in Eq. 4.3, we consider two potential implementations of our graph-based optimization technique.

4.3.4.1 Case 1: Static WE

A single **WE** can be estimated for all periods if the remaining spatial surfaces are stable across time. The residual spatial surface by averaging over T time intervals given as:

$$\tilde{S}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \hat{S}_{itj}, \quad i = 1, \ldots, K \tag{4.4}$$

4.3.4.2 Case II: Temporally varying WEt

In contrast, if the residual spatial surface changes significantly over time, a suitable neighbourhood structure will likewise change. In such occasions, it is good to apply the graph-based optimization algorithm on the residuals $(S_{1tj}, \ldots, S_{Ktj})$ from Eq.4.4 for each period t, resulting in a distinct matrix WEt for each t. Multiple realizations of the residual spatial surface are necessary to estimate **WEt** properly due to the dependence of residuals at various time lags. These residuals from Eq. 4.4 can be estimated using a $2q+1$ times moving average given as:

$$\tilde{S}_{ij} = \frac{1}{2q+1} \sum_{r=t-q}^{r=t+q} \hat{S}_{irj} \tag{4.5}$$

For instance, q=1 and t=2, $\tilde{S}_i = \frac{1}{3}\sum_{r=1}^{3} \hat{S}_{ir}$ and for t=T-1, $\tilde{S}_i = \frac{1}{3}\sum_{r=T-2}^{T} \hat{S}_{ir}$. Finally, we estimate **WE** using a set of spatial residuals $(\hat{S} = \hat{S}_{1j}, \ldots, \hat{S}_{Kj})$ computed from all the data, whereas **WEt** are estimated separately for each period t using a set of spatial residuals $(\hat{S} = \hat{S}_{1j}, \ldots, \hat{S}_{Kj})$ that is computed separately for each month in our case.

Then, **WE** and **WEt** will be estimated from a baseline neighbourhood matrix (**W**) based on the residual structure of the data from Eq. 4.4 and Eq. 4.5 using graph-based optimization algorithm proposed by (Lee et al., 2021; Enright et al., 2021). The algorithm works as follows:

Given $\tilde{S}$, we calculate $WE_{kh} = (0, 1)$ if $w_{kh} = 1$, but WE_{kh} remains set at zero if $w_{kh} = 0$. Furthermore, we assume that every district (vertex) in the graph must maintain at least one edge, which corresponds to the requirement $\sum_{k=1}^{K} WE_{kh} > 0$ for every k.

Suppose f$(Z, \tilde{S})$ denotes the values of the objective function $J(\tilde{S})$ corresponding to (WZ) for any generic graph (Z), the adjacency matrix corresponding to the sub-graph Z of G. Our

aim is finding a subgraph of G that maximizes f($\tilde{G}, \tilde{S}$) and has a minimum degree through a heuristic search approach.

- First, creates a graph G that corresponds to the original matrix **W** using input $\tilde{S}$.
- Determines a collection of edges occurring with initial vertex (v) that should be deleted using the original graph G.
- After deleting these edges, we create a new graph called $\tilde{G}$ and go on to the next vertex.
- The process repeats until all possible vertices have been traversed.

Even though estimating neighbourhood matrix **WE** or **WEt** is developed for univariate analysis, the estimated matrices for two or more variables with comparable spatial patterns should be similar as proposed by (Jack et al., 2019). We estimated residuals in Eq. 4.4 and 4.5 by averaging it for j^{th} outcome. Then, we model the spatial autocorrelation via the CAR prior proposed by (Leroux et al., 2000) with spatial precision matrix:

$$\mathbf{Q}(\mathbf{WE}, \rho) = \rho(diag[\mathbf{WE1}] - \mathbf{WE}) + (1 - \rho)\mathbf{I} \tag{4.6}$$

Here, **(1, I)** represents a $K \times 1$ vector of once and the $K \times K$ identity matrix, whereas $diag[\mathbf{WE1}]$ represents a diagonal matrix with diagonal elements (**WE** $\times$ 1). Parameter ρ measures overall spatial dependency, with a value of 0 corresponding to spatial independence. We specify a non-informative uniform before the unit interval to offer equal prior weight for all values of ρ and let the data take centre stage. For time-varying cases, we replace **WE** in Eq. 4.6 by **WEt**. Finally, the model is fitted in a Bayesian setting using the Markov chain Monte Carlo method. We assign weakly independent Gaussian prior for $\beta_j \sim N(0, 100, 000)$ to allow the data to play the dominant role, the prior degrees of freedom for the Inverse-Wishart follow marginally weakly-informative uniform on the interval $[-1, 1]$, assigned to

each correlation parameter. A non-informative uniform prior on the unit interval for ρ i.e. $\rho \sim Uniform(0, 1)$, to give equal prior weight for all allowable values of ρ. We use the R package CARBayesST (Lee et al., 2018) for fitting the models.

4.4 Results

4.4.1 Exploratory Analysis

An initial exploratory analysis provides invaluable insights into the distribution of the incidences. Consequently, we focus on the assessment of the relationship between incidences and covariates as well as test the residual spatio-temporal correlation. To assess the presence of spatial correlation, a separate Poisson generalized linear model was fitted for each outcome by including covariates. Then, Moran's I statistics were computed from the residuals of each model from August 2013 to May 2019. The results indicate that almost all Moran's I statistics values for every month are significant at the 5% level. For instance, the value of Moran's I statistics for each outcome in August 2013 is 0.214 and 0.116 with a corresponding p-value less than 5% indicating a signal of spatial correlation. The pairwise correlation between the residuals is 0.484, indicating, the presence of a significant correlation between two malarial cases that need to be modelled. The distribution of the incidence as shown in Fig. 4.2 varies with space and time. One can also observe more extraordinary cases of *P. falciparum* than *P. vivax* in the region.

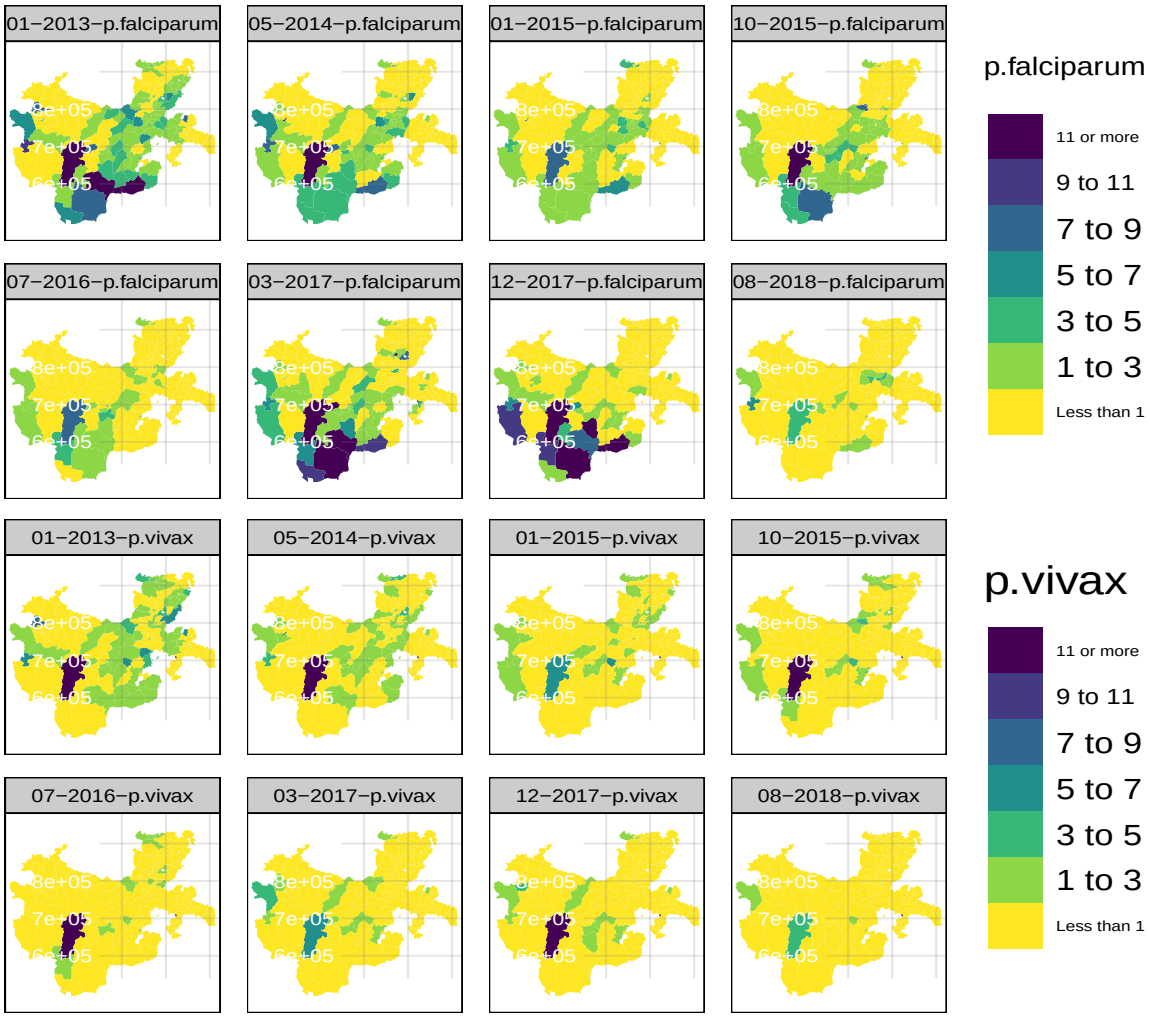

Figure 4.2 Maps indicating the observed incidence of *P. falciparum* and *P. vivax* for some selected months from August 2013 to May 2019 in Southern Ethiopia per 1000 population

The map in Fig. 4.2 indicates the distribution of malaria in the region for which higher incidences were observed in the southwestern region. Furthermore, more elevated incidences of *P. falciparum* were observed in the districts located in the southwest like; Konso, Hammer, Bero, Bako Gazer, and others, and the more increased incidences of *P. vivax* were observed in

southwestern and central districts like; Bako Gazer, Surma, and others. The decreasing trend of incidence with time was observed from August 2013 to May 2019 in Southern Ethiopia as shown in Fig. 4.3.

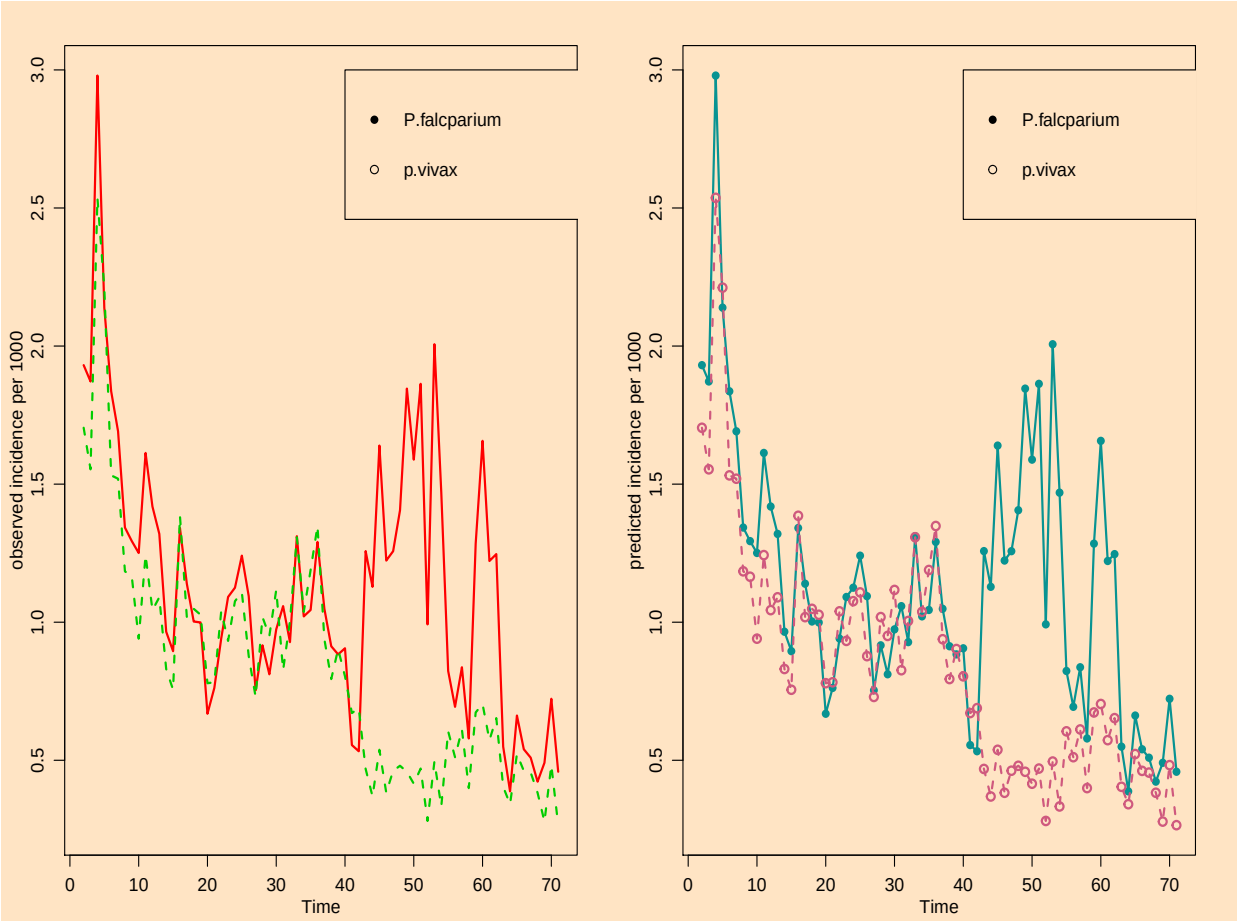

Figure 4.3 Scatter plots showing the temporal trends of the observed incidence; panel (A) and predicted incidence panel (B) obtained from the final model MSTCAR (AR(2) model with WE)

The red and green colour in panel (A) indicates observed incidence by *P. falciparum* and *P. vivax* respectively. Panel (B) describes the predicted incidence with deep-sky-blue represent-

ing the prediction by *P. falciparum* per 1000 population in Southern Ethiopia. Additionally, smoothed lines were used in the case of *P. falciparum*, and dotted lines were considered in cases of *P. vivax*. The correlations between the incidences for each month range from 0.1 to 0.6, suggesting the presence of a weak to strong correlation between the incidents. This is further supported by the right panel of Fig. 4.3 which demonstrates that while the temporal trends in the observed incidence occasionally fluctuate, similar patterns were generally seen throughout the time. From the plot, it is clear that the temporal patterns for each case are comparatively similar, beginning to decline from August 2013 to March 2015, and then appearing to be entirely in a stable state until July 2016. From July 2016 to January 2017, both incidents began to fall, and from January to December 2017, *P. vivax* incidence continued to decline while *P. falciparum* incidence increases, showing a different pattern between the two incidences throughout this time. Then, *P. falciparum* shows a decreasing trend whereas *P. vivax* appears to show some fluctuation. Finally, toward the beginning of 2019, there is an indication that the incidence starts to increase which coincides with the WHO report as shown in Fig. 4.1.

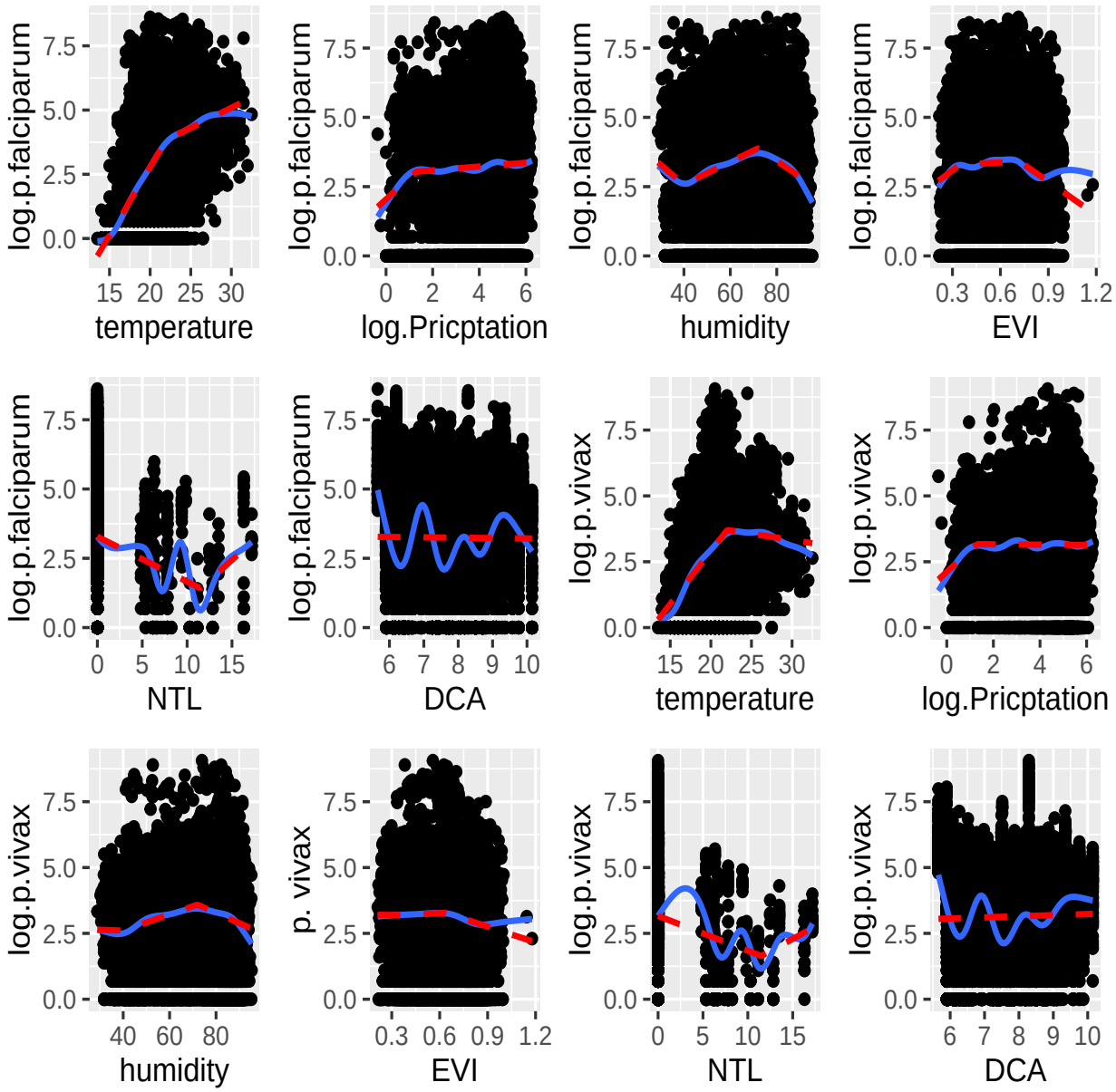

Figure 4.4 Scatter plot of *log.P. falcparium* and *log.P. vivax* against temperature, log-precipitation, EVI, Humidity, NTL, and DCA. The solid blue line shows natural splines and the dashed red line shows linear splines

Over here, we considered regression splines among several alternatives to draw a smoothing curve that helps to better visualize the nature of any remaining non-linear relationships between response and covariates as shown in Fig. 4.4. From the resulting curves, we select the knots for all covariates experiencing a noisy relationship with incidence by inspecting graphs for which an increasing trend is followed by a decreasing one, or vice versa (Giorgi et al., 2021). In our case, we plot the observed incidence in log scale against each of the covariates as shown in Fig. 4.4. We have taken a logarithm in the case of precipitation to influence a more linear relationship with the incidence.

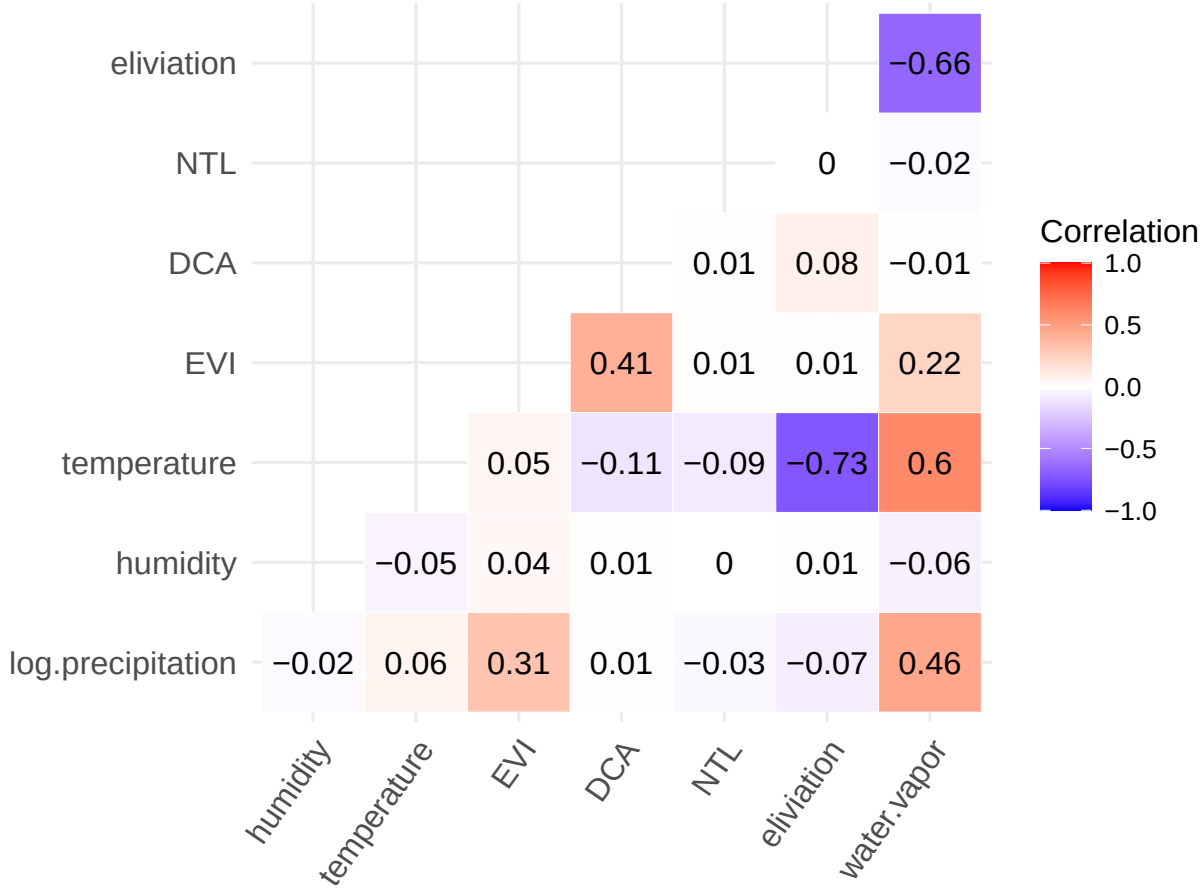

Figure 4.5 Correlation matrix between temperature, log-precipitation, humidity, EVI, DCA, NTL, elevation, and water vapor. Some of the variables having strong correlations were removed from the final model.

The resulting gracious curve from regression splines was implemented to select a point where the regression line changes from increasing to decreasing and vice versa. Moreover, we have considered the correlation matrix as shown in Fig. 4.5 to perceive how the covariates are

related to each other and then, withdrew some of the covariates like; elevation and water-vapor having significant correlations with others using AIC.

4.4.2 Estimating Static and time-varying waiting matrix from malaria risk mapping in Southern Ethiopia

To estimate the residual spatial structure, we first fit the GLM model Eq.4.1 without random effects. Using estimated spatial residuals from Eq.4.4 and Eq.4.5, we estimate static **WE** and time-varying **WEt** neighbourhood matrices. In the case of **WEt**, we use a three-month moving average to predict $\tilde{S_{rt}}$ that allow **WEt** to vary spatially over time. The temporally varying neighborhood structures is of the form $(WE_2, \ldots, WE_{69})$, for which $(WE_1 = WE_2$ and $WE_{69} = WE_{70})$. The residual spatial surfaces S_t are subsequently generated in our case as $S_t = S + S_t^*$ due to the static boundary during the study period in Southern Ethiopia. Despite similarities of the boundary, the random effects surfaces are assumed to be similar but not identical across time intervals. Thus, we incorporate a common spatial surface **S** for all periods and time-specific deviations with lower variance (S_t^*) to undertake any difference in each time point.

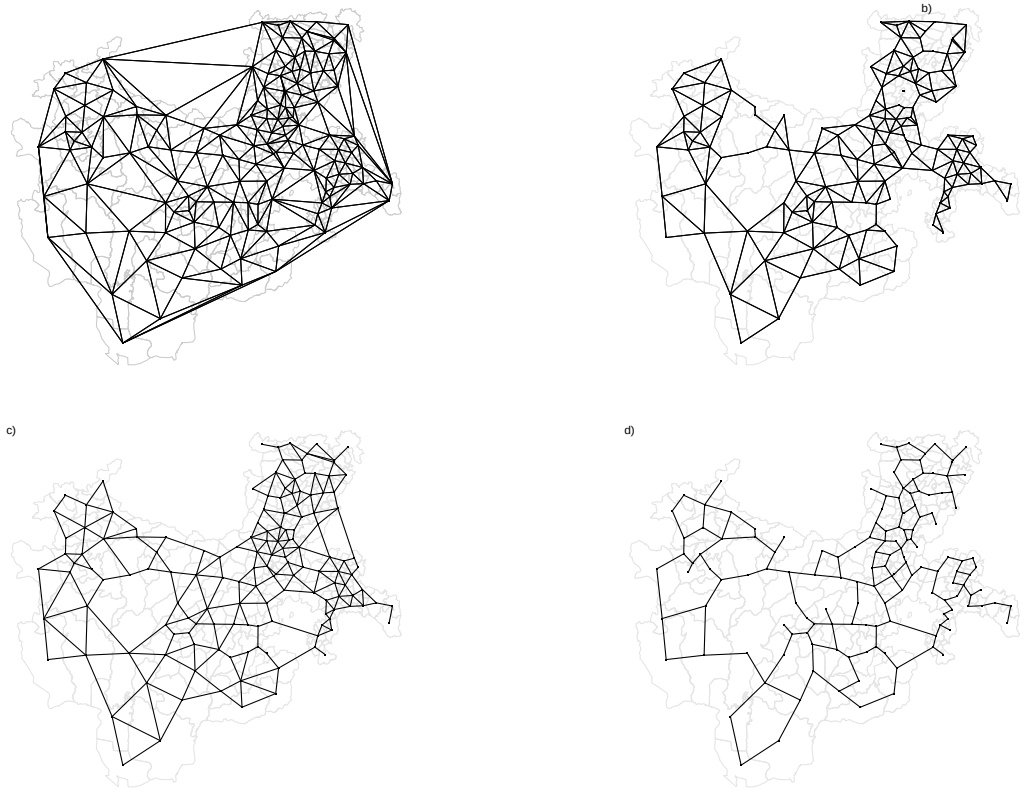

Figure 4.6 Neighbourhood identification with graph-based optimization; first map (a) shows neighbourhoods using the Simple border-sharing rule, 2^2&3^{rd} (b, c) neighborhoods using some other approaches and 4^{th} (d) shows graph-based Optimization methods with fewer edges

As boundaries are introduced into the spatial surface by the mean of $S = \mu$, if two adjacent regions (u, v) have the same mean, ($i.e.\mu_u = \mu_v$), then there won't be any boundary between them, but ($\mu_u \neq \mu_v$) can result in distinct random effects for the two areas Fig. 4.6. Finally, two distinct model specifications AR(1) and AR(2) are now fitted to the data using the

neighbourhood matrices with simple border sharing (**W**), static (**WE**), and time-varying (**WEt**) rule.

4.4.3 Model Fitting

Ultimately, we fit 4 models i.e. (1) first and second-order temporal autoregressive structures (AR(1) and AR(2)) and (2) the neighbourhood matrix defined by (**W**) and (**WE**) with different nearest proximity rules. Then, separate 136 models were fitted using time-varying **WEt** for (AR(1) and AR(2)). We fit these models with varying spatio-temporal correlations to indicate the sensitivity of the results to each model choice. The model with a temporal first-order autoregressive process is a model proposed by (Quick et al., 2017), while the second-order autoregressive process model is presented by (Lee et al., 2022). But a new graph-based optimization algorithm of estimating appropriate neighbourhood structure that we have considered here is proposed by (Lee et al., 2021). Inference for each of the 4 models is based on 5000 MCMC samples generated by collecting 60,000 samples with burn-in at 20,000 then thinned by 8 to greatly reduce their autocorrelation. Convergence was assessed using trace plots and Gelman-Rubin diagnostic.

The spatio-temporal correlations within and between outcomes indicate there is a significant spatial variety among the distribution of incidence by both *Plasmodium* specious as the posterior medians of $[\Sigma_{11}, \Sigma_{22}]$ are different for both models, with larger variation occurring in *P. falciparum* in some districts, see Table. 4.1. The variation between outcome computed as $[\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}]$ are around 0.7 for all models. The overall fit of each model to the data is compared via DIC, p.d, WAIC, p.w, and LMPL as shown in Table. 4.1. The result shows that the estimate obtained by static neighbourhood matrices (**WE**) with AR(2) results in better model fit as compared to simple border-sharing specifications. As the log marginal

Table 4.1 Summary of overall fit of the model via the DIC, model complexity via the effective number of independent parameters (p.d), and predictive ability via the log marginal predictive likelihood (LMPL)

	AR[1]		AR[2]	
Quantity	W	WE	W	WE
DIC	135708.03	135540.09	135622.61	135342.88
p.d	13788.77	13266.89	13205.18	13153.15
WAIC	134707.73	135453.20	133935.13	133754.91
p.w	9451.83	9855.505	9244.23	9204.08
LMPL	-72706.75	-73500.093	-73393.16	-74321.54
Σ_{11}	1.014 (0.967, 1.062)	0.657(0.631, 0.688)	1.140(1.092,1.190)	0.686(0.663,0.711)
Σ_{22}	0.763(0.726, 0.803)	0.731(0.639,0.857)	0.877(0.839,0.915)	0.517 (0.497, 0.537)
Σ_{12}	0.642 (0.641,0.642)	0.503 (0.504, 0.503)	0.735(0.734,0.735)	0.448(0.447, 0.448)
$\Sigma_{12}/\sqrt{\Sigma_{11} * \Sigma_{22}}$	0.730 (0.696, 0.765)	0.727(0.655, 0.794)	0.735 (0.705, 0.767)	0.752 (0.725, 0.779)
ρ_S	0.643(0.607, 0.683)	0.978 (0.948, 0.990)	0.641(0.611, 0.674)	0.488 (0.459, 0.516)
α	0.949 (0.944, 0.954)	0.943(0.937, 0.950)		
α_1			0.712 (0.692, 0.733)	0.709 (0.688,0.730)
α_2			0.243 (0.223, 0.263)	0.246 (0.225, 0.267)

predictive likelihood (LMPL) is higher in Table 4.1 with less effective numbers of independent parameters as assessed by p.d., employing the estimated **WE** will likewise yield superior predictive performance as opposed to utilizing **W**. Using **WE** and **WEt** minimizes the amount of variation between S_it and its spatially weighted mean, increasing precision. Unlike W, these methods do not include edges between pairs of geographically close districts that show significant differences in their residuals. The levels of spatial dependence estimated by AR(1) in each case are high in comparison to AR(2) due to the presence of temporal dependence while considering the model using AR(1) which is further undertaken using AR(2). Also, in the AR(1) and AR(2) models, the respective 95% credible intervals for α and α_1, α_2 are not close to zero which represents temporal independence.

Table.4.2 provides measures such as DIC, p.d. LMPL, and others using time-varying **WEt** for a randomly chosen 12-month period. The table demonstrates that employing **WEt** results in a better model fit when compared to **W** and **WE**. As was shown in Table.4.1 and Table.4.2, using either the static or the time-varying estimated neighbourhood matrices results in a better model fit than using the ordinary border sharing specification. The log marginal predictive likelihood (LMPL) in Tables.4.1 and 4.2 summarizes how much better the predictive performance is when using the estimated neighbourhood matrices **WE** or **WEt** than when using **W**.

Definition of notations:

In the first value of the index, 1 represents *P. falciparum* and 2 represents *P. vivax*. Whereas, *log.p*=log.precipitation, H=humidity, NTL=night time light, EVI=enhanced temperature index, l2=lag 2 and l4=lag 4. Table.4.3 presents estimated malaria risks with associated 95% credible intervals for each covariate, where each relative risk is related to the realistic rise in each covariate. The results of the analysis indicate that precipitation is significantly

Table 4.2 Model fitting parameters in cases of time-varying waiting matrix (**WEt**)

Quantity	10/2013	02/2014	10/2014	03/2015	10/2015	02/2016
DIC	135104.58	135207.81	134962.45	134999.85	134935.27	135538.08
p.d	13090.55	13079.50	13051.33	13197.05	13010.04	12945.79
WAIC	132708.57	133039.25	132467.77	132410.81	132457.51	133895.81
p.w	9143.82	9293.67	9049.61	9082.06	9031.32	9568.96
LMPL	-74769.64	-74983.56	-74494.99	-74675.55	-74472.33	-75438.51
Time	10/2016	04/2017	10/2017	02/2018	11/2018	02/2019
DIC	135242.70	135067.99	135080.72	134971.24	135139.53	134914.38
p.d	12995.73	13036.26	13191.74	13139.99	13103.86	13115.62
WAIC	133196.97	132636.84	132559.53	132428.31	132680.74	132312.66
p.w	9320.12	9083.48	9127.20	9075.72	9112.93	9016.43
LMPL	-75055.44	-74621.62	-74709.52	-74594.65	-74656.17	-74511.73

Table 4.3 Parameter estimates of the models and their 95% CI based on the MST-CAR [AR1] and MSTCAR [AR2]

	MSTCAR [AR1]		MSTCAR [AR2]	
Parameter	*W*	*WE*	*W*	*WE*
$\beta_{1.0}$	-2.95 (-3.33,-2.62)	-3.90 (-5.72,-2.91)	-9.94 (-10.23,-9.36)	-7.92 (-8.38,-7.42)
$\beta_{1.1} log.precipitation$	0.187 (0.072, 0.263)	0.369 (0.251, 0.474)	0.281(0.215, 0.337)	0.401 (0.351, 0.441)
$\beta_{1.2} I((log.p - 1.3) * (log.p > 1.3))$	-0.112 (-0.057,-0.105)	-0.217(-0.106,-0.051)	-0.115(-0.172,-0.064)	-0.112 (-0.261,-0.059)
$\beta_{1.3} log.precipitation_{l2}$	0.000 (0.000, 0.001)	-0.002 (-0.003,-0.002)	0.000(0.000,0.001)	0.001(0.000, 0.001)
$\beta_{1.4} log.precipitation_{l4}$	0.000 (0.000, 0.001)	0.000 (-0.001,0.000)	0.000(0.000,0.001)	0.000(0.000,0.001)
$\beta_{1.5} Temperature$	0.042 (0.030, 0.059)	0.128 (0.115, 0.138)	0.072 (0.059, 0.085)	0.041 (0.032, 0.049)
$\beta_{1.6} Temperature_{l2}$	0.084 (0.065, 0.104)	0.092(0.081, 0.106)	0.121 (0.098,0.134)	0.121(0.104, 0.139)
$\beta_{1.7} Temperature_{l4}$	0.165 (0.132, 0.187)	0.235 (0.187, 0.288)	0.227 (0.188, 0.257)	0.189 (0.170, 0.209)
$\beta_{1.8} Humidity$	-0.070 (-0.072,-0.066)	-0.061 (-0.065,-0.052)	0.005 (0.002, 0.008)	-0.024 (-0.026,-0.023)
$\beta_{1.9} I((H - 40) * (H > 40))$	0.072 (0.068, 0.075)	0.020 (0.015, 0.027)	0.000 (-0.003,0.004)	0.069 (0.050, 0.088)
$\beta_{1.10} I((H - 72) * (H > 72))$	-0.013 (-0.018,-0.007)	-0.014 (-0.021,-0.007)	-0.008 (-0.013,-0.002)	-0.009 (-0.013,-0.006)
$\beta_{1.11} Humidity_{l2}$	0.002 (0.000, 0.004)	0.002(-0.004,0.007)	0.001(-0.003,0.004)	0.004(0.002,0.006)
$\beta_{1.12} Humidity_{l4}$	0.003(0.001,0.005)	0.009(0.004,0.015)	0.007(0.004,0.008)	0.006(0.003,0.008)
$\beta_{1.13} EVI$	-0.187(0.418, -0.116)	-0.508 (0.746, -0.289)	-0.133 (-0.69,- 0.08)	-0.26 (-0.486, -0.005)
$\beta_{1.14} EVI_{l2}$	0.144 (0.000, 0.291)	-0.028 (-0.200,0.152)	0.367(0.222,0.535)	0.339(0.159,0.504)
$\beta_{1.15} EVI_{l4}$	-0.061(-0.214,0.057)	-0.042 (-0.242,0.155)	-0.002(-0.169,0.173)	0.081(-0.159,0.271)
$\beta_{1.16} NTL$	-0.043 (-0.084,0.005)	-0.024 (-0.060,0.010)	-0.033(-0.086,-0.007)	-0.045(-0.085,-0.009)
$\beta_{1.17} I((NTL - 11) * (NTL > 11)$	0.107 (-0.076,0.250)	0.022(-0.107,0.155)	0.086(-0.007,0.214)	0.096(-0.046,0.181)
$\beta_{1.18} DCA$	-0.001 (-0.001,-0.001)	-0.002(-0.002,-0.001)	0.000 (0.000,0.000)	0.000 (0.000, 0.001)
$\beta_{2.0}$	-10.52 (-11.70,-9.11)	-3.20(-5.32, -0.88)	-14.52 (-15.05,-13.94)	-11.05(-12.32,-9.85)
$\beta_{2.1} log.precipitation$	0.722=(0.59,0.89)	0.78(0.71, 0.84)	0.75(0.65,0.87)	0.47(0.35,0.58)
$\beta_{2.2} I((log.p - 1.3) * (log.p > 1.3))$	-0.80 (-0.99,-0.65)	-1.50 (-1.81,-1.26)	-0.85 (-0.97.-0.75)	-0.213 (-0.342, -0.093)
$\beta_{2.3} log.precipitation_{l2}$	0.000 (0.000, 0.001)	-0.005 (-0.006,-0.003)	0.000 (0.000, 0.001)	0.000 (0.000, 0.001)
$\beta_{2.4} log.precipitation_{l4}$	0.000 (0.000,0.000)	-0.001 (-0.002,-0.001)	0.000 (0.000,0.001)	0.000 (0.000, 0.000)
$\beta_{2.5} Temperature$	0.025(0.017, 0.034)	0.127 (0.096, 0.158)	0.044(0.023,0.065)	0.013(0.000, 0.023)
$\beta_{2.6} Temperature_{l2}$	0.073 (0.059, 0.087)	0.022 (-0.009,0.051)	0.087(0.077,0.097)	0.070(0.054,0.089)
$\beta_{2.7} Temperature_{l4}$	0.135 (0.104, 0.155)	0.104 (0.078, 0.122)	0.183(0.163,0.198)	0.141(0.124,0.166)
$\beta_{2.8} Humidity$	0.120(0.107, 0.131)	0.035(0.019, 0.052)	0.148(0.140, 0.155)	0.097 (0.087, 0.106)
$\beta_{2.9} I((H - 40) * (H > 40))$	-0.12(-0.14,-0.11)	-0.14 (-0.18,-0.12)	-0.15 (-0.15,-0.13)	-0.09(-0.11,-0.08)
$\beta_{2.10} I((H - 72) * (H > 72))$	-0.002(-0.006,0.005)	-0.03 (-0.03,-0.02)	0.001(-0.005,0.008)	-0.002 (-0.006,0.001)
$\beta_{2.11} Humidity_{l2}$	0.004(0.002, 0.007)	-0.01(-0.02,-0.006)	0.004(0.001,0.007)	0.005 (0.003, 0.006)
$\beta_{2.12} Humidity_{l4}$	0.004(0.002,0.005)	0.042 (0.031, 0.049)	0.008(0.005,0.010)	0.006(0.004, 0.008)
$\beta_{2.13} EVI$	0.019(-0.135,0.153)	-0.401 (-0.663,-0.150)	0.196(0.047,0.499)	0.133(-0.032,0.323)
$\beta_{2.14} EVI_{l2}$	-0.111 (-0.256,0.002)	-0.388 (-0.617,-0.202)	0.117(0.019,0.230)	0.026(-0.124, 0.168)
$\beta_{2.15} EVI_{l4}$	0.042 (-0.157,0.204)	0.010(-0.274, 0.238)	0.058(-0.080,0.228)	0.103(-0.040,0.275)
$\beta_{2.16} NTL$	-0.018 (-0.060,0.021)	-0.008 (-0.059,0.044)	-0.006 (-0.052,0.026)	-0.005(-0.047,0.024)
$\beta_{2.17} I((NTL - 11) * (NTL > 11))$	0.04(-0.10, 0.17)	0.000(-0.18,0.16)	0.02(-0.11,0.14)	-0.003(-0.09,0.13)
$\beta_{2.18} DCA$	0.000(-0.001,0.001)	0.001(0.000,0.002)	0.001 (0.000,0.001)	-0.002 (-0.004,-0.001)

associated with malaria risk in the region; in particular, precipitation below $1.3mm$ on a log scale is significantly associated with an estimated between 0.2 and 0.5 increase in malaria risk depending on the model and precipitation above $1.3mm$ is associated with a 0.112 decrease in malaria risk. In general, a unit increase in precipitation is significantly associated with a $0.401 - 0.112 = 0.289$ increase in malaria risk of the genus *P. falciparum* in the region. On the other hand, precipitation on a lag of 2 and 4 months indicates no effect on malaria risk of the genus *P. falciparum*. Similarly, 1^oc increase in temperatures is significantly associated with a 0.03 to 0.3 increase in the malaria risk of the genus *P. falciparum*. Also, temperatures in 2 and 4-month time lag are positively associated with a 0.06 to 0.3 increase in the malaria risk of the genus *P. falciparum*. In the case of humidity, an increase in humidity below 40% was significantly associated with a 2.4% decrease in *P. falciparum* risk, $0.069 - 0.024 = 4.5\%$ increase in malaria risk in humidity between 40 to 70% and $0.069 - 0.024 - 0.009 = 3.6\%$ increase in malaria risk for humidity above 70%. Also, a unit increase in humidity at a lag of 2 and 4 months respectively results in 0.4% and 0.6% increase in malaria risk of the genus *P. falciparum*. This indicates an increase in the malaria risk for every unit increase in humidity. An increase in EVI was significantly associated with a 0.263% decrease in *P. falciparum* risk. Additionally, EVI at a time lag of 2 months is also significantly associated with a 0.339 (0.159, 0.504) rise in *P. falciparum* risk but not significant on other lags. On the other hand, an increase in NTL below 11.5% is significantly associated with a -0.045(-0.085,-0.009) decrease in *P. falciparum* risk. Finally, an increase in distance from the coastal area is not significantly associated with *P. falciparum* risk.

In the case of *P. vivax*, an increase in precipitation below 1.3mm in the log scale is significantly associated with a 0.467 increase in malaria risk of the genus *P. vivax*, whereas the increase in precipitation above 1.3mm in the log scale is significantly associated with a

$0.467 - 0.213 = 0.254$ decrease in *P. vivax* risk. Contrarily, the lag effect of precipitation at 2 and 4 months lag does not have a significant effect on *P. vivax* risk. Additionally, an increase in temperature is significantly associated with a 1.3%, 7%, and 14% increase in *P. vivax* risk at 0, 2, and 4-month lag respectively. An increase in humidity above 40% was significantly associated with a 9.7% increase in *P. vivax* risk and an increase of humidity between 40% to 70% is significantly associated with $0.097 - 0.094 = 0.3\%$ increase in *P. vivax* risk in the region, holding all other factors constant. Also, humidity at a time lag of 2 and 4 months is significantly associated with 0.5% and 0.6% increase in malaria risk of genus *P. vivax* respectively. EVI and NTL are not significant with *P. vivax* risk at different time lags whereas a km increase in DCA is significantly associated with a 0.2% decline in malaria risk of genus *P. vivax*.

The results also revealed that using (**WE**) and (**WEt**) results in better covariate effect estimation compared with using (**W**) in the majority of covariates considered with narrow credible intervals. We couldn't present the result by (**WEt**) due to the number of models considered and the similarity of the results compared with (**WE**). Additionally, the widths of the 95% credible intervals are somehow narrower for the model using (**WE**) and (**WEt**) in comparison to the models using (**W**).

4.4.4 Spatio-temporal trend in Southern Ethiopia

The prediction and exceedance maps presented below are based on the AR(2) CAR model with waiting matrix **(WE)**, which is the best model as it was shown in Table. 4.1 and Table. 4.2 with smaller DIC and other metrics. And, the model helps to effectively capture both the temporal and spatial correlations in the data set.

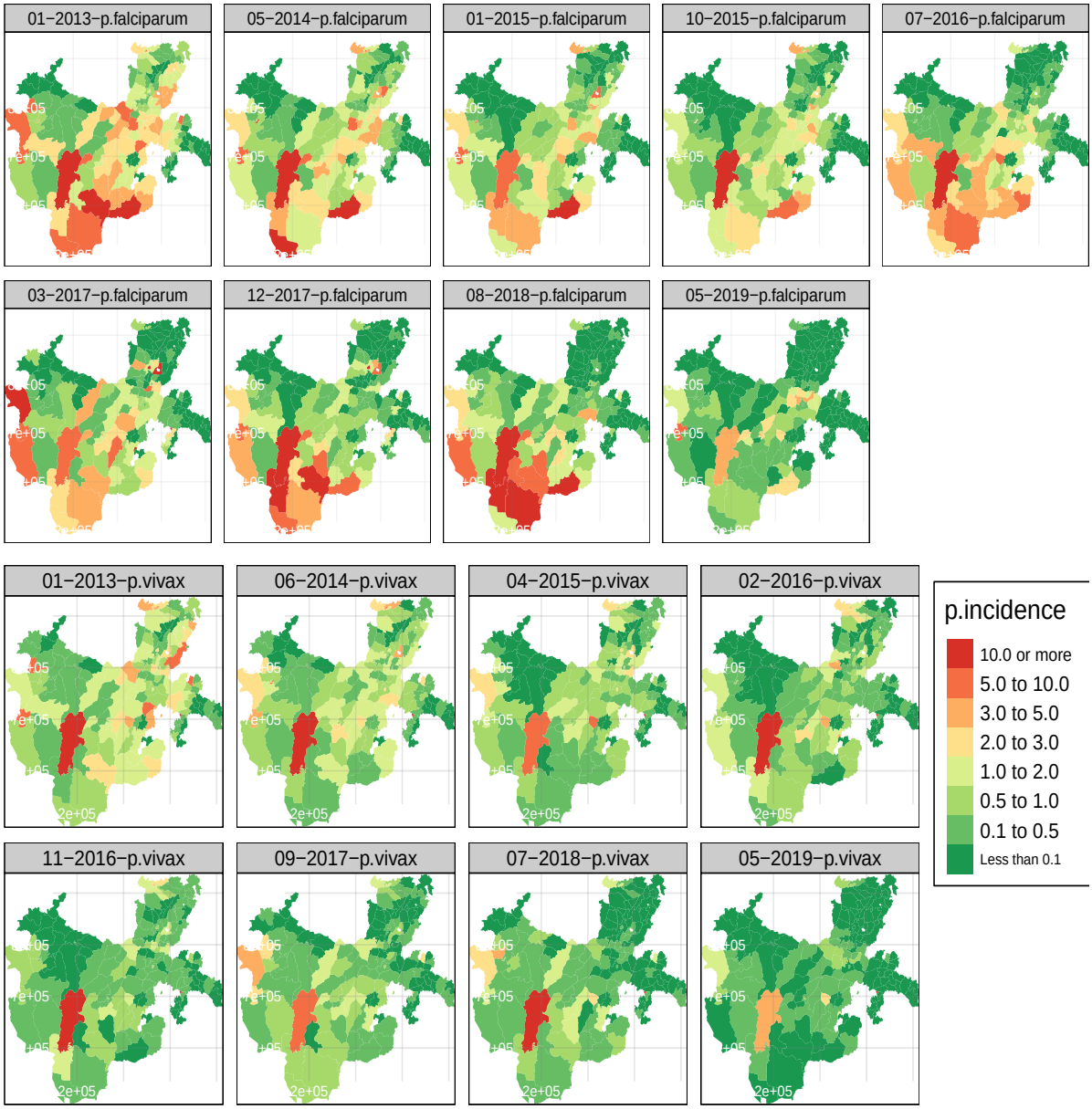

Figure 4.7 Multivariate spatio-temporal prediction maps of malaria risk for selected months from August 2013 to May 2019 of two *Plasmodium* specious i.e. *P. falciparum* and *P. vivax* per 1000 population in Southern Ethiopia

The predictions by AR(2) using **(WE)** exhibit much less noise than the observed incidences as shown in the left panel of Fig. 4.3 which might be due to the Spatio-temporal smoothing applied by the model. The trends in the predicted incidence for *P. falciparum* have a steeper ascent and descent compared to the *P. vivax*. The Spatio-temporal trend as indicated by both species in Fig. 4.7 highlighted spatial and temporal variation of the incidences. The map shows that most districts have relatively low incidences of *P. vivax* throughout the study period except similarities were observed from late 2013 to early 2015 as shown in Fig. 4.7 and Fig. 4.3. The map also shows the *P. vivax* incidences do not show a pronounced spatial variation which might be due to the lower incidence relative to the total population at risk.

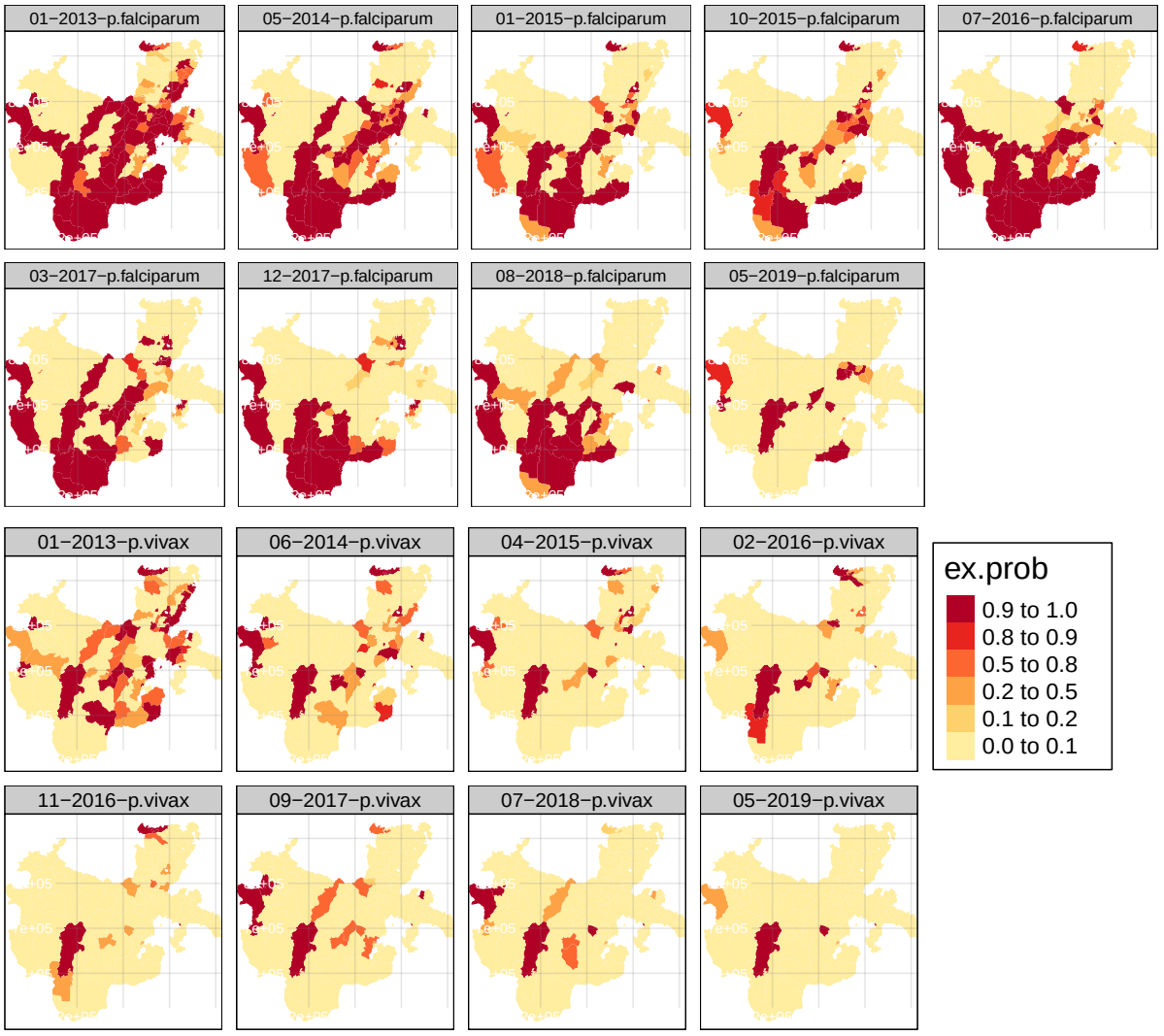

Figure 4.8 Exceedance probability (ex. prob) maps of malaria incidence by both *P. falciparum* and *P. vivax* incidence using malaria risk mapping per 1000 population in Southern Ethiopia.

Figure. 4.8 indicates the districts where the incidences are higher. Here we consider a simple threshold of 1 and plot districts whose predicted incidences are above 1 out of 100 population at risk or predicted incidence ≥ 0.01. From the map, districts around the southwest; there is

a higher risk of *P. falciparum* whereas more cases of *P. vivax* were observed in the districts found in the central region. Also, those districts are identified as a hot spot by the Ethiopian Public Health Institute (FMHE, 2020).

4.5 Discussion

This study considers graph-based optimization methods proposed by (Lee et al., 2021) for multivariate spatio-temporal data in southern Ethiopia. The method aids in improving future prediction by taking time-varying influences into account and better quantifying the impact of various factors on historical spatio-temporal patterns by accounting for delayed changes in climate. Subsequently, we incorporated the estimated **(WE)** and **(WEt)** to the MSTCAR model proposed by (Lee et al., 2022) assuming that disease risk by both *plasmodium* species follows a similar pattern across space and time supported by the exploratory analysis in the left panel of Fig. 4.3 and some similar studies (Zhou et al., 2005; Jack et al., 2019). The results indicated that estimating the neighbourhood matrix using a graph-based optimization algorithm provides improved estimation and uncertainty quantification in comparison to using a simple border-sharing approach since the model has a smaller DIC value with fewer effective parameters, which coincides with a study by (Lee et al., 2021; Goepp and van de Kassteele, 2022). This might be notably due to the reduction of the edge effect in the approach compared to the simple border-sharing rule. The approach is equally dominant in the identification of geographically adjacent boundaries with very different data values or characteristics (Lee et al., 2021; Bivand et al., 2008).

We noticed the importance of multivariate modelling over univariate analysis, as it helps to obtain a more coherent picture of the distribution of the overall malaria risk in areas by both *Plasmodium* species simultaneously with some of the corresponding disease-specific or general factors among the cases. These studies revealed strong clustering of *P. falciparum* over *P. vivax* in various districts of the region over time, which coincides with a study by (Tessema et al., 2020) that showed marked temporal clustering of *P. falciparum* in the low-

transmission setting of the country. This might also be appropriate to lower parasite densities in *p.vivax* compared to *P. falciparum*, resulting in low detectability (Hofmann et al., 2017; Tessema et al., 2020). This additionally tells why many scholars focus on estimating the relative burden of *P. falciparum* over *P. vivax*.

From the spatio-temporal prediction maps shown in Fig. 4.7 and the exceedance probability maps in Fig. Fig. 4.8, the distributions of malaria risk by either species vary with space and time. For instance, the Soro districts of the Gurage Zone, Analemo in the Hadiya Zone, Loka Abaya in the Sidama Zone, Yirgachefe and Dilla Zuria in the Gedeo Zone, Sodo Zuriya and DFamot Woyida in the Wolayta Zone, Selamago, southern Ari, Hamer, Chena, Mirab Abaya, Bero, Yemi, and Sheka are identified as high-risk areas for both *P. falciparum* and *P. vivax* species located in the western and central parts of the region, which is consistent with some previous works (Abeku et al., 2004; Midekisa et al., 2012; Rodo et al., 2021).

In some districts, more elevated incidences by both *plasmodium* species are noticed with homogeneous spatio-temporal patterns (Hundessa et al., 2017; Bi et al., 2013), but high spatial heterogeneity is observed in many areas in some selected times and abundances where one of the rates is dominant and others are not. Regarding temporal dynamics, the time when a higher incidence of *P. falciparum* is observed to differ from that of *P. vivax* in most districts. Mostly, more elevated incidences of *P. falciparum* was observed from September to November then from January to March and from June to August following a heavy and average rain season that coincides with (WHO, 2021; Kitawa et al., 2022). This is sometimes associated with their response to various environmental and climatic factors (Kitawa et al., 2022).

From the results of the analysis, malaria threat by *P. falciparum* and *P. vivax* is positively as-

sociated with precipitation, temperature, and relative humidity. More specifically, increases in mm of precipitation on the log scale are associated with a 0.289 (0.090, 0.382) increase in malaria hazard by the genus *P. falciparum* whereas a unit increase in precipitation is significantly associated with a 0.254 (0.009, 0.489) increase in *P. vivax* risk. However, a nonlinear relationship was observed between malaria risk and precipitation that coincides with a study by (Kitawa et al., 2022; Giorgi et al., 2021). On the other hand, the delayed precipitation effect did not have a significant effect on malaria risk by either species on a 2 and 4-month time lag. However, at early lags of 1 month, precipitation is significantly associated with malaria risk which was initially removed due to a high correlation that coincides with a study by (Li et al., 2013).

In the case of temperature, a 1^0C increase in temperature is associated with a 0.041 (0.032, 0.049) and 0.013 (0.00, 0.023) increase in *P. falciparum* and *P. vivax* risk respectively. Similarly, the delayed temperature effect at a 2 and 4-month time lag is significantly associated with malaria risk by either species which coincides with the study by (Midekisa et al., 2012; Rotejanaprasert et al., 2021). According to several studies, greater temperatures encourage the development of mosquito parasites, and it is well known that malaria transmission in Ethiopia often occurs during the rainy season when temperatures are typically lower. As a result, the transmission may take place at longer lags and the vector's life span in the area may be extended. (Rotejanaprasert et al., 2021).

An increase in humidity on the other hand is significantly associated with a 0.036 (0.05, 0.02) and 0.023 (0.007, 0.062) increase in *P. falciparum* and *P. vivax* risk respectively holding all other factors constant. The risk initially decreases as an increase in humidity below 40% and then rises for humidity between 40% and 70% for *P. falciparum* risk. Whereas it initially

increases in cases of *P. vivax* below 40% of humidity and then declines for the humidity between 40% to 70%. Also, the delay effect of humidity at a time lag of 2 and 4 months will increase malaria risk by 4% to 6% respectively on both *Plasmodium* species. This suggests that, under an optimal lag period, there is a positive correlation between humidity and malaria risk (Li et al., 2013). The increase in incidence by *P. falciparum* is negatively associated with EVI, DCA, and NTL, i.e., As the EVI increases, the incidence of *P. falciparum* decreases by 0.263 (0.005, 0.49) holding other factors constant. This coincides with the recent study by (McMahon et al., 2021) in northern Ethiopia associating the highest malaria rates with low vegetation cover but contradicts the study by (Giorgi et al., 2021) in malaria risk mapping in Tanzania. On the other hand, EVI at a time lag of 2 months is positively associated with malaria risk of genus *P. falciparum*. Regarding NTL and DCA, an increase in NTL results in a -0.119 (-0.001, -0.199) decline in malaria risk by *P. falciparum*. This is sometimes associated with urbanization and population size as NTL increases; urbanization and population density increase. Thus, as urbanization increases infrastructure and quality of life improve, leading to a lower incidence of malaria, as suggested by (Giorgi et al., 2021).

Finally, an increase in distance from the coastal area results in a decreased risk of malaria by either *Plasmodium* species. This result was supported by different studies (Kitawa et al., 2022; Rodo et al., 2021; Bhatt et al., 2015), but *P. vivax* is less sensitive to different environmental factors than *P. falciparum* (Seyoum et al., 2017) as some of the covariates like EVI and NTL are not significant with *P. vivax* risk. The delay effect of temperature and humidity is positively associated with malaria risks by both *Plasmodium* species and the effect is greater as they form favorable conditions for mosquito breeding. The delayed effect is even greater as compared non-delayed effect. However, the relationship with some of the

covariates, such as precipitation, temperature, and NTL, is nonlinear, as shown by the spline curve in Fig. 4.4, which differs at different times. In the case of *P. vivax*, more elevated incidences were observed from June to August, followed by January to March and September to November, with a decreasing trend. This could be the case because the distribution of the incidence is not stable, as the malaria report indicated (WHO, 2018), which also coincides with the climatic conditions varying from zone to zone as well as from district to district in the region. By adopting the MSTCAR model to the malaria data, we detected substantial geographic variation in malaria risk by the genus *P. falciparum* and *P. vivax* in southern Ethiopia. The spatiotemporal variation of malaria incidence by *P. falciparum* and *P. vivax* has not been sufficiently incorporated previously due in part to a lack of methods equipped for multivariate spatio-temporal modelling, and in this way, we have pointed out how the spatio-temporal variation in the incidence varies across space and time in the setting.

Graph-based approaches have received considerable attention recently and are rarely used in some monitoring similar situations. It starts from one or several random seeds (vertices) and explores the neighbourhoods of visited vertices. As demonstrated in diverse contexts, the approach (Lin and Zhao, 2019b; Prokhorenkova and Shekhovtsov, 2020) indicates considerably superior performance over another simple neighbourhood approach. However, a more detailed emphasis is required to obtain more robust methods for multivariate spatio-temporal data in the future. Thus, employing multivariate methods while estimating the neighbourhood matrix using graph-based optimization helps to capture the spatial correlation better in comparison to the commonly used border-sharing rule. Such results are achieved by viewing the areal units as the vertices of a graph and the neighbour as the set of edges evidenced by malaria risk mapping in Southern Ethiopia between August 2013 and May 2019. One way of extending this approach would be to develop a more relevant estimation method for

the neighbourhood matrix for multivariate spatial and spatio-temporal data that possibly account for the boundary change and the presence of excess zeros.

Other possible options are developing a multivariate geostatistical model for such aggregated count data, as suggested by (Eyre et al., 2020), that allows us to borrow information across multiple incidences. One can also look forward to more informative risk factors that are not captured in these studies. Furthermore, the availability of a more recent malaria data set would have provided more latest information about the incidence in the region. Additionally, there was a lack of some significant socioeconomic predictors that would have improved the prediction. Other drawbacks include under-reporting, a standard problem in surveillance data in a low-resource setting, and a shortage of a more recent data set about the population because we have used the population data set projected based on (CSA, 2007) that is carried out every 10 years.

4.6 Conclusion

The study delivers a significant contribution to the present body of knowledge for describing the geographical variation of malaria risk by multiple *Plasmodium* species at the district level. By using the surveillance data set aggregated monthly in the region, we have (1) predicted malaria risk by multiple *Plasmodium* species in the region and identified districts where either species are dominant, (2) identified time-varying risk factors associated with malaria risk in the region and (3) incorporated the delayed climatic factors on malaria incidence in the region. From the results of the analysis, our prediction maps provide invaluable insight into the distribution of the incidences and priority areas that necessitate direct urgent interventions. Furthermore, using time-varying neighbourhood specification that takes

into account the distribution of residuals in comparison to a simple border-sharing rule is important.

List of abbreviations

GLM: Generalized Linear Model; CAR: Conditional Autoregressive model; DIC: Deviance Information Criterion; MCMC: Markov chain Monte Carlo; EVI: Enhanced Vegetation Index, NTL: Night time light, MSTCAR: Multivariate spatio-temporal conditional Autoregressive model, p.d: effective number of independent parameters, LMPL: Log marginal predictive likelihood; WAIC: Widely Applicable Information Criterion or Watanabe–Akaike information criterion.

Declarations:

Ethics Approval and consent to participate

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the College of Natural and Computational Science, Hawassa University (protocol code: RERC/030/12, and date of approval: July 20, 2022). The ethical review board of Hawassa University College of Natural and Computational Science waived participant informed consent since the study was conducted using district-level monthly surveillance data.

Consent for publication

Not applicable

Availability of data and materials

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data-sharing policy of the Ethiopian Public Health Institute. Temperature and precipitation data sets were extracted freely from (worldclim.org/data/monthlywth.html and worldclim21.html), Relative Humidity was derived from ERA-Interim global atmospheric reanalysis, Enhanced vegetation index (EVI) was obtained from Moderate-resolution Imaging Spectroradiometer (MODIS) (lpdaac.usgs.gov/products/mod13a3v006/) and Nighttime light (NTL) obtained from NOAAs National Centers for Environmental Information,(https://ngdc.noaa.gov/eog/ viirs/index.html).

Competing interests

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

YS conceived the study idea, designed the study, carried out the statistical analyses, interpreted the results, and drafted the manuscript. ZG participated in the design of the study, reviewed the manuscript and provided technical support for interpreting results and reviewed it for intellectual content. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to the Ethiopian Public Health Institute worker for facilitating data collection and to Hawassa University for creating an opportunity to participate in the research work. Also, we extend our gratitude to Professor Arnoldo Frigessi, for his valuable comments.

4.7 References

Abeku, T., Vlas, S.D., Borsboom, G., Tadege, A., Gebreyesus, Y., Gebreyohannes, H., Alamirew, D., Seifu, A., Nagelkerke, N., Habbema, J., 2004. Effects of meteorological factors on epidemic malaria in ethiopia: a statistical modelling approach based on theoretical reasoning. Parasitology 128, 585–593. URL: https://researchonline.lshtm.ac.uk/id/eprint/14632/.

Benjamin, M.T., Andrade-Pacheco, R., JWS., H., 2018. Continuous inference for aggregated point process data. Royal Statistical Society 181, 1125–1150. doi:10.1111/1467-9876.00113.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological) 36, 192–225.

Besag, J., York, J., Mollie, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43, 1–20. doi:10.1007/BF00116466.

Bhatt, S., Weiss, D., Cameron, E., et.al, 2015. The effect of malaria control on plasmodium falciparum in africa between 2000 and 2015. Nature 526, 207–211. doi:10.1038/nature15535.

Bi, Y., Yu, W., Hu, W., Lin, H., Guo, Y., Zhou, X.N., Tong, S., 2013. Impact of climate variability on plasmodium vivax and plasmodium falciparum malaria in yunnan province, china. Parasites & vectors 6, 1–12.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., Pebesma, E.J., 2008. Applied spatial data analysis with R. volume 747248717. Springer.

Blangiardo, M., Cameletti, M., Baio, G., Rue, H., 2013. Spatial and spatio-temporal models with r-inla. Spatial and spatio-temporal epidemiology 4, 33–49.

Chirombo, J., Ceccato, P., Lowe, R., Terlouw, D.J., Thomson, M.C., Gumbo, A., Diggle, P.J., Read, J.M., 2020. Childhood malaria case incidence in malawi between 2004 and 2017: spatio-temporal modelling of climate and non-climate factors. Malaria journal 19, 1–13.

Cohen, J.M., Le Menach, A., Pothin, E., Eisele, T.P., Gething, P.W., Eckhoff, P.A., Moonen, B., Schapira, A., Smith, D.L., 2017. Mapping multiple components of malaria risk for improved targeting of elimination interventions. Malaria Journal 16, 1–12.

Colborn, K.L., Giorgi, E., Monaghan, A.J., Gudo, E., Candrinho, B., Marrufo, T.J., Colborn, J.M., 2018a. Spatio-temporal modelling of weekly malaria incidence in children under 5 for early epidemic detection in mozambique. Scientific reports 8, 1–9. doi:10.1038/s41598-018-27537-4.

Colborn, K.L., Mueller, I., Speed, T.P., 2018b. Joint modeling of mixed plasmodium species infections using a bivariate poisson lognormal model. The American journal of tropical medicine and hygiene 98, 71.

Cressie, N., 1993. Statistics for spatial data, revised edition wiley. New York, NY.[Google Scholar] .

CSA, 2007. Central statistical authority, 2007 population and housing census of ethiopia. country level. Addis Ababa, Ethiopia .

Deress, T., Girma, M., 2019. Plasmodium falciparum and plasmodium vivax prevalence in ethiopia: a systematic review and meta-analysis. doi:10.1155/2019/7065064.

Diggle, P., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. Applied Statistics 47, 299–350. doi:10.1111/1467-9876.00113.

Enright, J., Lee, D., Meeks, K., Pettersson, W., Sylvester, J., 2021. The complexity of finding optimal subgraphs to represent spatial correlation, in: Combinatorial Optimization and Applications: 15th International Conference, COCOA 2021, Tianjin, China, December 17–19, 2021, Proceedings, Springer. pp. 152–166.

Eyre, M.T., Carvalho-Pereira, T.S., Souza, F.N., Khalil, H., Hacker, K.P., Serrano, S., Taylor, J.P., Reis, M.G., Ko, A.I., Begon, M., et al., 2020. A multivariate geostatistical framework for combining multiple indices of abundance for disease vectors and reservoirs: a case study of rattiness in a low-income urban brazilian community. Journal of the Royal Society Interface 17, 20200398.

Eyre, M.T., Souza, F.N., Carvalho-Pereira, T.S., Nery, N., de Oliveira, D., Cruz, J.S., Sacramento, G.A., Khalil, H., Wunder, E.A., Hacker, K.P., et al., 2021. Linking rattiness, geography and environmental degradation to spillover leptospira infections in marginalised urban settings: an eco-epidemiological community-based cohort study in brazil. medRxiv .

FMHE, 2020. Ethiopia malaria elimination strategic plan by federal ministry of health ethiopia: 2021-2025 .

Gasparrini, A., 2014. Modeling exposure–lag–response associations with distributed lag non-linear models. Statistics in medicine 33, 881–899.

Gelfand, A.E., Vounatsou, P., 2003. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4, 11–15.

Giorgi, E., Fronterre, C., Macharia, P.M., Snow, V.A.A.R.W., Diggle, P., 2021. Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict. J. R. Soc. Interface 18. doi:10.1098/rsif.2021.0104.

Gneiting, T., Genton, M., Guttorp, P., 2006. Geostatistical space-time models, stationarity, separability and full symmetry. American Statistical Association 97, 590–600. doi:10.1201/9781420011050.ch4.

Goepp, V., van de Kassteele, J., 2022. Graph-based spatial segmentation of health-related areal data. Computational Statistics and Data Analysis .

Gómez-Rubio, V., Palmí-Perales, F., 2019. Multivariate posterior inference for spatial models with the integrated nested laplace approximation. Journal of the Royal Statistical Society: Series C (Applied Statistics) 68, 199–215.

Gómez-Rubio, V., Palmi-Perales, F., López-Abente, G., Ramis-Prieto, R., Fernández-Navarro, P., 2019. Bayesian joint spatio-temporal analysis of multiple diseases. SORT-Statistics and Operations Research Transactions , 51–74.

Gómez-Rubio, V., Rue, H., 2018. Markov chain monte carlo with the integrated nested laplace approximation. Statistics and Computing 28, 1033–1051.

Hofmann, N.E., Karl, S., Wampfler, R., Kiniboro, B., Teliki, A., Iga, J., Waltmann, A., Betuela, I., Felger, I., Robinson, L.J., et al., 2017. The complex relationship of exposure to new plasmodium infections and incidence of clinical malaria in papua new guinea. Elife 6, e23708.

Huang, A., Wand, M.P., 2013. Simple marginally noninformative prior distributions for covariance matrices .

Hundessa, S., Williams, G., Li, S., Guo, J., Zhang, W., Guo, Y., 2017. The weekly associations between climatic factors and plasmodium vivax and plasmodium falciparum malaria in china, 2005–2014. Transactions of the Royal Society of Tropical Medicine and Hygiene 111, 211–219.

Jack, E., Lee, D., Dean, N., 2019. Estimating the changing nature of scotland's health inequalities by using a multivariate spatiotemporal model. Journal of the Royal Statistical Society: Series A (Statistics in Society) 182, 1061–1080. doi:10.1111/rssa.12447.

Kitawa, Y., Asfaw, Z., 2023. Space-time modeling of monthly malaria incidence for seasonal associated drivers and early epidemic detection in southern ethiopia .

Kitawa, Y., Jonson, O., Asfaw, Z., Giorgi, E., 2022. A comparison of spatio-temporal models for malaria risk mapping in southern ethiopia. Spatial statistics 50.

Lawson, A.B., 2018. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC press.

Lee, D., Meeks, K., Pettersson, W., 2021. Improved inference for areal unit count data using graph-based optimisation. Statistics and Computing 31, 1–17.

Lee, D., Robertson, C., Marques, D., 2022. Quantifying the small-area spatio-temporal dynamics of the covid-19 pandemic in scotland during a period with limited testing capacity. Spatial statistics 49, 100508.

Lee, D., Rushworth, A., Napier, G., 2018. Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the CARBayesST package. Journal of Statistical Software 84, 1–39–350. doi:10.18637/jss.v084.i09.

Leonard, C.M., Mohammed, H., Tadesse, M., McCaffery, J.N., Nace, D., Halsey, E.S., Girma, S., Assefa, A., Hwang, J., Rogier, E., 2022. Missed plasmodium falciparum and plasmodium vivax mixed infections in ethiopia threaten malaria elimination. The American Journal of Tropical Medicine and Hygiene 106, 667.

Leroux, B.G., Lei, X., Breslow, N., 2000. Statistical models in epidemiology, the environment, and clinical trials, chapter estimation of disease rates in small areas: A new mixed model for spatial dependence. Springer-Verlag, New York , 179–191.

Li, T., Yang, Z., Wang, M., 2013. Temperature, relative humidity and sunshine may be the effective predictors for occurrence of malaria in guangzhou, southern china, 2006–2012. Parasites & vectors 6, 1–4.

Lin, P.C., Zhao, W.L., 2019a. Graph based nearest neighbor search: Promises and failures. arXiv preprint arXiv:1904.02077 .

Lin, P.C., Zhao, W.L., 2019b. Graph based nearest neighbor search: Promises and failures. arXiv preprint arXiv:1904.02077 .

McMahon, A., Mihretie, A., Ahmed, A.A., Lake, M., Awoke, W., Wimberly, M.C., 2021. Remote sensing of environmental risk factors for malaria in different geographic contexts. International journal of health geographics 20, 1–15.

Midekisa, A., Senay, G., Henebry, G., Semuniguse, P., Wimberly, Michael, C., 2012. Remote sensing-based time series models for malaria early warning in the highlands of ethiopia. Malar J 11. doi:10.1186/1475-2875-11-165.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Nigatu, W., Abebe, M., Dejene, A., 1992. Plasmodium vivax and p. falciparum epidemiology in gambella, south-west ethiopia. Tropical medicine and parasitology: official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ) 43, 181–185.

Palmí-Perales, F., Gómez-Rubio, V., López-Abente, G., Ramis, R., Sanz-Anquela, J.M., Fernández-Navarro, P., 2021. Approximate bayesian inference for multivariate point pattern analysis in disease mapping. Biometrical Journal 63, 632–649.

Palmí-Perales, F., Gómez-Rubio, V., Martinez-Beneito, M.A., 2021. Bayesian multivariate spatial models for lattice data with inla. Journal of Statistical Software 98, 1–29. URL: https://www.jstatsoft.org/index.php/jss/article/view/v098i02, doi:10.18637/jss.v098.i02.

Prokhorenkova, L., Shekhovtsov, A., 2020. Graph-based nearest neighbor search: From practice to theory, in: International Conference on Machine Learning, PMLR. pp. 7803–7813.

Quick, H., Waller, L.A., Casper, M., 2017. Multivariate spatiotemporal modeling of age-specific stroke mortality. The Annals of Applied Statistics 11, 2165– 2177. doi:10.1214/17-AOAS1068.

Rodo, X., Martinez, P., Siraj, A., Pascual, M., 2021. Malaria trends in ethiopian highlands track the 2000 'slowdown' in global warming. Nat Commun 12. doi:10.1038/s41467-021-21815-y.

Rotejanaprasert, C., Ekapirat, N., Sudathip, P., Maude, R.J., 2021. Bayesian spatio-temporal distributed lag modeling for delayed climatic effects on sparse malaria incidence data. BMC Medical Research Methodology 21, 1–15.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71, 319–392.

Seyoum, D., Yewhalaw, D., Duchateau, L., Brandt, P., Rosas-Aguirre, A., Speybroeck, N., 2017. Household level spatio-temporal analysis of plasmodium falciparum and plasmodium vivax malaria in ethiopia. Parasites & vectors 10, 1–11.

Taylor, B.M., Davies, T.M., Rowlingson, B.S., Diggle, P.J., 2015. Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. Journal of Statistical Software 63, 1–48.

Tessema, S.K., Belachew, M., Koepfli, C., Lanke, K., Huwe, T., Chali, W., Shumie, G., Mekuria, E.F., Drakeley, C., Gadisa, E., et al., 2020. Spatial and genetic clustering of plasmodium falciparum and plasmodium vivax infections in a low-transmission area of ethiopia. Scientific reports 10, 1–10.

Vicente, G., Goicoa, T., Ugarte, M., 2020. Bayesian inference in multivariate spatio-temporal areal models using inla: analysis of gender-based violence in small areas. Stochastic Environmental Research and Risk Assessment 34, 1421–1440.

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

White, N.J., 2011. Determinants of relapse periodicity in plasmodium vivax malaria. Malaria journal 10, 1–36.

WHO, 2018. Malaria surveillance, monitoring & evaluation: a reference manual.

WHO, 2021. World malaria report 2021 .

Zhou, G., Sirichaisinthop, J., Sattabongkot, J., Jones, J., Bjørnstad, O.N., Yan, G., Cui, L., 2005. Spatio-temporal distribution of plasmodium falciparum and p. vivax malaria in thailand. The American journal of tropical medicine and hygiene 72, 256–262.

CHAPTER 5. CONCLUSION and Future Works

While each paper includes a discussion, we also provide an in-depth analysis of each paper in this chapter, as well as some future perspectives on the issues addressed and suggestions for how to enhance and broaden the applicability of the approach by drawing on more incidences in more diverse parts of the country.

5.1 Conclusion and Way-forward on Paper-I

An essential component of the efficient implementation of control strategies is understanding the temporal dynamics of malaria transmission. Here, we developed a malaria early warning system that helps to indicate elevated risks and compare temporal dynamics and seasonality patterns across districts in Southern Ethiopia. The result indicates significant regional variation in the incidence of malaria caused by $p.faciparum$, ranging from highly seasonal and climatic-associated dynamics to less seasonal and climatic-independent dynamics.

Also, the degree and the timing of seasonal peaks vary significantly between districts. This shows however that, the various dynamics may be grouped into many "dynamical archetypes," each of which is characterized by distinctive temporal characteristics. From the analysis, a variety of environmental factors, such as precipitation, temperature, humidity, EVI, NTL, DCA, and seasonality of $p.falciparum$ incidence, give an appropriate insight to more easily recognize and comprehend patterns of seasonal variation of the risk.

To assist malaria control and elimination, further development and implementation of district-

level malaria forecasting is important. This is because, regardless of the geographical location, similar dynamics were frequently observed in a manner independent of the geographical setting. However, timely and precise data are crucial for developing district-level early warning systems, which surveillance data does not always provide. Although the quality and completeness of accessible surveillance data is a problem, these challenges can be addressed by employing model-based geostatistics techniques (Diggle and Giorgi, 2019).

To model such spatiotemporally referenced data from low-resource settings, geostatistical methods offer a practical and statistically sound approach (Diggle and Giorgi, 2016). The ability to estimate risk at health decision units and to account for uncertainty characteristics is one of these methods' main benefits. However, by taking into account spatiotemporal correlation, model-based geostatistical techniques enable the extrapolation of incompletely observed data in space and time to generate characteristics of risk at administrative units sufficient to make policy-relevant decisions.

For relatively rare diseases, when judgments between maintaining mass control and targeted eradication are needed at the subnational level, modelling uncertainty by exceedance probability has a larger utility. However, it is occasionally possible for decision-makers to consistently misinterpret the degree of uncertainty underlying predictions. i.e., they are unable to assess the likelihood that a health region would fall into a risk category that might indicate a change in a strategy. This is due to recorded parasite incidence becoming increasingly diverse as transmission falls, with most districts reporting figures near zero. To deal with this uncertainty, absolute incidence numbers become hard to understand as a measure of transmission intensity. As a way to quantify the uncertainty of incidence regarding policy-relevant intervention limits, non-exceedance probabilities might be used instead, providing

suitable metrics to make decisions on shifting from sustained malaria control to malaria elimination strategies. Additionally, the non-exceedance probability could be utilized to assess uncertainty concerning a predetermined threshold to identify areas that need immediate attention.

5.2 Conclusion and Way-forward on paper II

Based on the EWS model considered for placing districts based on temporal trends, we understand that incorporating covariates when assessing space-time variation of illness risk is important. Even though the EWS developed for each district assists in clustering districts based on temporal trends, a more toiled approach that incorporates a geographical and temporal heterogeneity of the disease may be important.

The CAR model, spatiotemporal geostatistical model, and spatial discrete approximation to log Gaussian Cox process model (SDALGCP) are three modelling approaches considered to include a spatiotemporal heterogeneity of disease risk here. We have discovered that 1) it is critical to include spatial correlation in detecting spatiotemporal variation in illness risk. 2) While all models perform relatively well in detecting spatiotemporal heterogeneity in disease risk, spatial continuous models perform better with lower RMSE in detecting spatiotemporal heterogeneity in malaria risk. Among the continuous models examined here, the spatiotemporal discrete approximation to log Gaussian Cox processes is significantly superior, with lower RMSE and higher coverage. This might be due to the limitations of representing districts by a single-point, "centroid" in geostatistics and some over-smoothing issues in the CAR model. STDALGCP, on the other hand, assumed disease risk to be spatially continuous and discretized this risk in each area based on population density.

Here, disease count data that have been spatially aggregated is utilized for spatial prediction of disease risk at any desired geographic scale using the SPSDALGCP model, an extension of the SDALGCP model provided by (Johnson et al., 2019). Regardless of data type, we observed the LGCP framework to be an effective statistical paradigm for modelling aggregated disease count data as the risk of disease fluctuates along a geographical continuum. We argue that STSDALGCP offers a computationally effective solution while keeping the spatially continuous nature of disease risk when computing constraints make fitting an LGCP impractical.

There is also the possibility of extension to a multivariate version. This occurs when more than one outcome is measured at each geographic unit. For example, imagine we want to analyze data from a malaria count for several *Plasmodium* species with j=2. Multivariate modelling is useful for investigating the relationship between disease risk in each district and the association between data across units. In the case of multiple outcomes, STSDALGCP may be extended by considering y_ijt disease cases count for i^{th} districts, disease j at time t; let d_{ijt}^T be a vector of explanatory factors for j illness with corresponding coefficient β_j; and let S_{ijt} be a Gaussian process. Then, y_{ijt} may be modelled as a spatially aggregated multivariate log-Gaussian Cox process.

Another thing we've discovered is that, when compared to other models, the CAR model delivers coverage probability closer to 99%. This might be due to the waiting matrix's over-smoothing, implemented throughout the neighbourhood. Because the CAR model is more computationally efficient than others, considering different neighbourhood approaches other than defining through just border sharing rules might be beneficial (Bivand et al., 2008). In this situation, we know that the disease risk is not continuous throughout the region and is

not discrete as it was defined by the government, thus there are tendencies and spaces where disease risk discontinues. Thus, one can also look at a model developed by Orozco-Acosta et al. (2023, 2021) for scaling the risk in some specific range for which disease risk shows some common trend.

5.3 Conclusion and Way-forward on paper III

As we have discussed in Chapter 2, integrating spatial correlation in disease risk mapping is critical. However, most spatial relationship in a small area is approximated by a collection of random effects, a part of a Gaussian Markov random field, with a conditional autoregressive prior distribution. This model's spatial autocorrelation using binary neighbourhood matrix **W**, in which two random effects are supposed to be spatially correlated if they share a common boundary and are otherwise conditionally independent. However, boundary sharing between districts does not necessarily ensure the presence of spatial correlation. Some of the adjacent districts in Ethiopia have various features in terms of incidence distribution, which may be related to the climatic conditions of those area units. In contrast to geostatistical modelling, where variogram analysis is commonly employed to establish an acceptable spatial autocorrelation structure for the data, the suitability of **W** for the data at hand is rarely investigated.

We considered the graph-based optimization technique **W** for identifying the neighbourhood matrix, which follows some of the neighbourhood definition methods offered by Bivand et al. (2008) and some of the current approaches proposed by (Lee et al., 2021). This method better captures spatial correlation structure by seeing areal units as graph vertices and neighbourhood relations as a collection of edges. Furthermore, the strategy reduces the

over-smoothing caused by borrowing information from every surrounding district, which can occasionally restrict the identification of high-risk locations. Another advantage of the graph-based optimization technique over the basic border-sharing rule is the estimation of either a static or a temporally variable neighbourhood matrix.

Following this modelling approach, we have identified spatiotemporal varying risk factors and areas with elevated risks of malaria. Also, we have further pointed out the delayed climatic effect on the distribution of malaria as malaria risks do not only depend on the current climatic conditions. This is because mosquitoes mostly increase in the early dry season after heavy rainy fall. This might be due to; malaria larvae being sub-optimal during the rainy season when flooding occurs but growing progressively as the dry season progresses.

Furthermore, one can look forward to more risk factors, as malaria is not only dependent on climatic factors but also a more toiled approach for taking zero inflation into account. Also, considering a continuous model as proposed in the summary of paper two is one alternative as the problem of zero inflation can also be undertaken using unstructured random effect using a model-based approach (Diggle and Giorgi, 2019). Another alternative to look forward is local discontinuities in the spatial pattern which is not usually modeled. A binary neighbourhood matrix that is built based on a border-sharing condition requires spatial correlation across geographically adjacent regions and is mostly used to represent the auto-correlation (Yin et al., 2022). Enforcing such a connection, however, could mask any irregularities in the disease risk surface, making it more difficult to identify clusters of places with greater or lower risks than their nearby neighbours (Yin et al., 2022). Therefore, leaving out these risk discontinuities causes the risk maps to be too smoothed and hides distinct hot/cold spot

regions. However, risk discontinuities may be modelled by allowing the clusters to be either stable or to change over time.

5.4 Some of the Works in progress

Despite some of the limitations, we are currently working to identify clusters or zones where the Maria risks have similar spatial patterns over time rather than only looking at districts. i.e. Based on the recently developed "divide-and-conquer" procedure, several local CAR models may be fitted once. (Orozco-Acosta et al., 2021). Similarly, (Santafé et al., 2021) also developed scalable models that help to identify clusters through risk discontinuities. This approach significantly reduces computing time while producing credible risk estimations. Also, we are looking to forecast incidence by using spatiotemporal correlation as was initially suggested by (Wang et al., 2018) and further improved (Orozco-Acosta et al., 2023) for short-term forecasting. This approach allows non-stationarity in each subdomain and also space-time interaction effects in the model.

Initially, we have data set on climatic and environmental variables from 40 stations in southern Ethiopia across the desired period, most of which are in town areas. As a result of the limitations of station data, we used satellite data for the covariates instead. Even though satellite data provide promising measurements, accuracy is sometimes questioned. Here, we first anticipate the connection between satellite and station data in a specific area, and then: 1) reformulate the spatiotemporal geostatistical model in Eq. 1.2 to handle such limitations in satellite data and or generate new covariates using Model-based geostatistics. Then, using new variables for each district, we will predict malaria risk in southern Ethiopia to that effect, as station data may still be preferred over satellite data for evaluating climatic conditions

(Mendelsohn et al., 2007). Linear geostatistics might be used more effectively and precisely to describe such forms of incompletely observed data in space and time (Diggle and Giorgi, 2019).

5.4.1 Limitations

The paper analyzed and compared malaria incidence using aggregated count data in the region, there are still some problems:

- We have only considered climatic factors as a covariate due to the absence of other sociodemographic factors that may have improved our predictions
- We have considered population data projected based on the 2007 census, but the recent census may have provided a better picture if available.
- Computational tools that we have in our hands are not sufficient to further explore the topic via simulations and more detailed analysis.

5.5 References

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., Pebesma, E.J., 2008. Applied spatial data analysis with R. volume 747248717. Springer.

Diggle, P., Giorgi, E., 2016. Model-based geostatistics for prevalence mapping in low-resource settings. American Statistical Association 111, 1096–1120. doi:10.1080/01621459.2015.1123158.

Diggle, P., Giorgi, E., 2019. Model-based Geostatistics for Global Public Health: Methods and Applications. Chapman & amp; Hall/CRC Interdisciplinary Statistics, Chapman and Hall/CRC Press.

Johnson, O., Giorgi, E., Diggle, P., 2019. A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. Statistics in Medicine 38, 4871–4887. doi:10.1002/sim.8339.

Lee, D., Meeks, K., Pettersson, W., 2021. Improved inference for areal unit count data using graph-based optimisation. Statistics and Computing 31, 1–17.

Mendelsohn, R., Kurukulasuriya, P., Basist, A., Kogan, F., Williams, C., 2007. Climate analysis with satellite versus weather station data. Climatic Change 81, 71–83.

Orozco-Acosta, E., Adin, A., Ugarte, M.D., 2021. Scalable bayesian modelling for smoothing disease risks in large spatial data sets using inla. Spatial Statistics 41, 100496.

Orozco-Acosta, E., Riebler, A., Adin, A., Ugarte, M., 2023. A scalable approach for short-term disease forecasting in high spatial resolution areal data. arXiv preprint arXiv:2303.16549 .

Santafé, G., Adin, A., Lee, D., Ugarte, M.D., 2021. Dealing with risk discontinuities to estimate cancer mortality risks when the number of small areas is large. Statistical Methods in Medical Research 30, 6–21.

Wang, X., Yue, Y.R., Faraway, J.J., 2018. Bayesian regression modeling with INLA. CRC Press.

Yin, X., Napier, G., Anderson, C., Lee, D., 2022. Spatio-temporal disease risk estimation using clustering-based adjacency modelling. Statistical Methods in Medical Research 31, 1184–1203.