Statistical Modelling and Mapping of Health Outcomes in Developing Countries

by

Jessie Jane Khaki

A thesis submitted for the degree of *Doctor of Philosophy*

in the

Faculty of Health and Medicine

Lancaster Medical School

January 2025

Statistical Modelling and Mapping of Health Outcomes in Developing Countries

Jessie Jane Khaki

Lancaster Medical School

A thesis submitted for the degree of *Doctor of Philosophy*.

Abstract

The 2030 Sustainable Development Goals (SDGs) aim at improving the lives of people. To monitor the progress towards achieving the SDGs and effectively improve people's lives, there is a need to efficiently use publicly available data to inform decisions. However, developing countries struggle to track the SDGs due to limited financial resources and technical skills. This thesis explores how health SDG outcomes can be tracked and modelled using publicly available datasets in low- and middle-income countries (LMICs).

In Chapter 3, this thesis investigates how passive surveillance data arising from a typhoid point pattern process in Blantyre, Malawi, can be analysed using environmental and individual-level covariates such as age and gender. Chapter 4 applies multilevel and mixed effects models to publicly available geostatistical demographic and health survey data from Malawi to model and map the double and triple malnutrition burden among mother-child pairs without spatial correlation.

Chapter 5 extends the work carried out in Chapter 4 by applying model-based geostatistics to publicly available geostatistical soil-transmitted helminth survey data from 35 African countries. Chapter 5 also discusses some challenges encountered when using sparse data from LMICs and provides recommendations on ideal data for geospatial predictions. Lastly, Chapter 6 characterises the dengue

outbreak in 77 Nepalese districts between 2006 and 2022. Using district-level areal data and a modified Negative Binomial model, the thesis estimates the timing and duration of 3 outbreak intensity functions within each district.

This thesis demonstrates the use of statistical modelling in tracking health outcomes in developing countries. The thesis additionally discusses the challenges associated with publicly available data in LMICs, such as sparse data, and proposes solutions to these challenges. Finally, the thesis suggests ways in which each aspect of the research can be extended in future studies.

Contents

List of Tables

List of Figures

Abbreviations

AETC	Accident and Emergency Treatment Centre
AIC	Akaike Information Criteria
AOR	Adjusted Odds Ratio
AMR	Antimicrobial Resistance
ANC	Antenatal Care
BMI	Body Mass Index
BYM	Besag-York-Mollié
CDF	Cumulative distribution function
CI	Confidence Interval
DALYS	Disability-Adjusted Life Years
DBM	Double Burden of Malnutrition
DENV	Dengue Virus
DHS	Demographic and Health Survey
EA	Enumeration Area
EDCD	Epidemiology and Diseases Control Division
ESPEN	Expanded Special Project for Elimination of Neglected Tropical Diseases
GPS	Global Positioning System
HAZ	Height/Length for Age Z-score
HB	Hemoglobin
IHME	Institute for Health Metrics and Evaluation
IPP	Inhomogeneous Poisson Process
LMIC	Low and Middle Income Country
LRT	Likelihood Ratio Test
MBG	Model-based Geostatistics
MDA	Mass Drug Administration
MDR	Multidrug-resistant
MICS	Multiple Indicator Cluster Survey
MM	Milimeters
MNHSRC	Malawian National Health Sciences Research Committee
NCD	Non-communicable Diseases
NTDs	Neglected Tropical Diseases
OIF	Outbreak Intensity Function

PC	Preventive Chemotherapy
PCA	Principal Components Analysis
PHIA	Population-based HIV Impact Assessments
QECH	Queen Elizabeth Central Hospital
SD	Standard Deviation
SDG	Sustainable Development Goals
SPA	Service Provision Assessment
SSA	sub-Saharan Africa
S. Typhi	*Salmonella enterica* Typhi
STH	Soil-Transmitted Helminths
STRATAA	Strategic Alliance Against Typhoid Across Africa and Asia
TBM	Triple Burden of Malnutrition
WASH	Water Sanitation and Hygiene
WAZ	Weight for Age Z-score
WHZ	Weight for Height Z-score
WHO	World Health Organization
UN	United Nations
VIF	Variance Inflation Factor

Acknowledgements

I would like to thank the Almighty God for the grace and privilege of bringing me so far in my academic journey.

I am extremely grateful to Dr. Emanuele Giorgi for his patience, supervision, and mentorship during my studies. I am also very grateful to Dr. Marc Henrion and Professor Melita Gordon for their guidance and mentorship, without whose mentorship during my pre-doctoral internship at the Malawi-Liverpool-Wellcome Trust I would not have been able to get the funding and necessary skills for this PhD. Many thanks should also go to Professor Mavuto Mukaka, who has been a constant mentor since my graduate studies at the University of Malawi. I would also like to thank Professor Victor Mwapasa and Professor Kenneth Maleta for their support in my career over the years. To Professor Jo Knight, Dr. Anastacia Ushakova, Professor Peter Diggle, Dr. Claudio Fronterre, Professor Chris Jewell, Dr. Jon Read and Barry, thank you for the interesting chats during coffee. I would also like to appreciate my cohort mates and friends: Rachael, Irene, Olatunji, Alex, Fran, Yawen, Misaki, Amitha, Matt, Rui, Jess, Cian, Peter, Marianne, Noni, Alinane, Richard, Kristina and Nirali; thank you for the chats we had whilst trying to do our work.

Thank you to my wonderful family: my son, husband, parents (George and Joyce Khaki), siblings, and friends for being strong pillars during my studies. Special thanks should go to my sister, Clara Khaki, for always being there whenever LB needed to Facetime his auntie. You were thousands of miles away while I was doing this PhD, but LB knows you as if you were living with us; it really takes a village to raise a child!

Declaration

I declare that this thesis and the work presented in it are my own. I also confirm that no part of this work has been submitted for a degree or any other qualification at this university or any other institution. The word count of this thesis, including appendices, but excluding the bibliography, does not surpass the maximum allowable length of 80,000 words. The estimated word count of the thesis is approximately 29,439.

This thesis is composed of 4 research papers as follows:

Chapter 3

Modelling Salmonella Typhi in high-density urban Blantyre neighbourhood, Malawi, using point pattern methods.

Authors: Khaki, J. J., Thindwa, D., Henrion, M. Y. R., Jere, T., Msuku, H., The STRATAA Consortium, Heyderman, R., Gordon, M. A., Giorgi, E.

Published in: *Scientific Reports.*

Contribution: Conceptualization, methodology, formal analysis, writing the original draft.

Chapter 4

Prevalence and determinants of double and triple burden of malnutrition among mother-child pairs in Malawi: a mapping and multilevel modelling study.

Authors: Khaki, J. J., Macharia, P., Benova, L., Giorgi, E., Seeman, A.

Published in: *Public Health Nutrition.*

Contribution: Conceptualization, methodology, formal analysis, writing the original draft.

Chapter 5

Using ESPEN Data for Evidence-Based Control of Neglected Tropical Diseases in sub-Saharan Africa: a Comprehensive Model-based Geostatistical Analysis of Soil-Transmitted Helminths.

Authors: Khaki, J. J., Minnery, M., Giorgi, E.

Published in: *PLOS Neglected Tropical Diseases.*

Contribution: Methodology, formal analysis, writing the original draft.

Chapter 6

Disentangling Outbreak Patterns of Dengue Fever in Nepal: A District-Level Analysis from 2006 to 2022.

Authors: Khaki JJ, Acharya BK, Pandey BD, Moritaf K, Giorgi E.

In preparation for submission.

Contribution: Methodology, formal analysis, writing the original draft.

For my son and husband. Also, for my late mother, Sarah Naomi Khaki (nee Nkhonjera, 1965-2005). We did it, mum!

Chapter 1

Introduction

1.1 Background

Spatial statistics are a potentially useful complement to routinely used data sources in developing countries. These statistics play a pivotal role in research by providing insights into the geographic distribution of health outcomes, identifying health outcome clusters, and investigating the association between environmental factors and health outcomes [1]. Developing countries experience a high burden of health outcomes such as malnutrition and diseases compared to developed countries. Countries in low and middle income countries often experience a high incidence and prevalence of both infectious diseases, non-infectious diseases and maternal and child health issues. Within these countries, the burden of disease also varies widely by geographic region and in time. Ensuring effective control and monitoring of health outcomes is crucial for reducing the high disease and health outcome burdens in developing countries. Nevertheless, many developing countries still face substantial hurdles in collecting the requisite data to measure progress accurately [2]. The quest for improved healthcare delivery in developing countries necessitates an in-depth understanding of the spatial dimensions that underlie disparities in the incidence and prevalence of health outcomes, but the data along these dimensions is acutely lacking in developing countries [3].

In resource-constrained settings, representative household surveys and routine health surveillance management information systems serve as primary data sources

for monitoring health indicators [2]. However, developing countries still face challenges in implementing these surveys and using publicly available data for several reasons, including limited technical skills and financial resources [2, 4, 5]. Efforts are thus needed to optimize the utilization of publicly available data within low and middle-income countries to effectively facilitate the monitoring of the health-related outcomes, as informed by the United Nations' (UN) Sustainable Development Goals (SDGs).

Another phenomenon that underscores the need for close attention to spatial distribution of health outcomes is urbanization. Urbanization in developing countries is advancing rapidly, thereby exerting significant effects on population health [3, 6, 7]. Given this trend, addressing sub-national disparities across sub-national units such as districts becomes increasingly crucial for effective health planning and resource allocation. For instance, 15 of the 17 health-oriented SDGs include a spatial component to facilitate monitoring of these indicators at a sub-national level. These sub-national geographic units serve as focal points for monitoring progress and as proxies for populations that share similar attributes, such as access to health services or increased exposure to environmental risk factors [3]. Consequently, investigating health disparities across sub-national areas can provide valuable evidence and rationale for directing health initiatives and policies, particularly when inequalities are pronounced.

There are, generally, three types of spatial data: geostatistical data, point process data, and lattice data. In geostatistical data, the study area (A) is a continuous fixed set where observations can be observed anywhere within the study area [8]. However, the observed locations in A are non-stochastic [9]. On the other hand, a spatial point pattern dataset comprises random observed locations of an event [8, 10]. Examples of geostatistical data include the prevalence of malaria among children in villages, while an example of point process data is the geographical coordinates of the households of tuberculosis patients in an area A. In lattice data, also referred to as areal data, the study area A is fixed, and it is partitioned into a finite number of regular or irregular areal units at which the outcomes of interest are aggregated [8, 9, 11, 12]. An example of lattice data is the number of malnourished children within each district in a country A. In contrast,

spatio-temporal data have a time observation in addition to the location data.

The goals of analyzing spatial data can be grouped as follows: explaining the relationships between a health outcome of interest and risk factors; carrying out predictions, given an identical spatial process as observed in the data being analyzed; and improving sampling for spatial surveys [13]. This doctoral thesis focuses on the first two goals and applies them to health outcome data from all three types of spatial data collected in resource-constrained settings. We also discuss some challenges that may arise when using secondary data collected in developing countries and provide recommendations on addressing the challenges in future work. Furthermore, our work discusses the importance of fitting national and sub-national level models due to varying health outcome trends and shared risk factors within the sub-national areas.

In the next chapter, we introduce the 3 main types of spatial data and provide the primary statistical methods used to model them. The current chapter concludes with an outline of the thesis and the main projects to which the methods were applied.

1.2 Thesis structure and objectives

Thesis objectives

The primary goal of this thesis was to contribute to the use of publicly available datasets in low and middle income countries by developing and applying spatial statistical methods to contribute to the third SDG of good health and well-being. In this thesis, two distinct approaches were used: the first and fourth studies develop methods for point pattern data and areal-level disease outbreak data, respectively. Both the first and fourth studies use routinely collected health facility data. The second and third studies involve the application of already existent geospatial methods to model and map publicly available health survey data. The main objectives of each study are as follows:

1. To develop spatial and spatio-temporal inhomogeneous Poisson process models for typhoid data that explicitly account for multiple marks (Chapter 3).

2. To investigate the prevalence and determinants of malnutrition of mother-child pairs in Malawi and to assess the geographic distribution of mother-child malnutrition (Chapter 4).

3. To establish the prevalence of soil-transmitted helminths (STH) at the fine-scale and subnational levels; and to classify subnational units to the World Health Organization STH prevalence classes using publicly available STH data from the Expanded Special Project for Elimination of Neglected (ESPEN) tropical diseases database (Chapter 5).

4. To determine the timing and duration of multiple outbreak intensity functions in each Nepalese district using annual health facility data (Chapter 6).

Structure of the thesis

This chapter provides a background to the PhD work. Chapter 2 reviews statistical methods for modelling and mapping three types of spatial data.

Chapter 3 (Paper 1), models data arising from a typhoid spatio-temporal point pattern process from routinely collected health facility data from a peri-urban area in a developing country. We also discuss the challenges arising from such data and propose how the model can be further extended. Chapter 4 (Paper 2) analyzes publicly available demographic and health survey data collected in Malawi. Using geospatial malnutrition data from mother-child pairs in Malawi, we demonstrate how meaningful spatial mapping can be conducted even in the absence of spatial correlation within the dataset.

Similar to Chapter 4, Chapter 5 (Paper 3) focuses on the use of publicly available data to inform the monitoring and control of neglected tropical diseases in low- and middle-income countries. To achieve this goal, we apply model-based geostatistical methods to soil-transmitted helminths prevalence data collected in 35 African countries. The results of the work are presented in the paper, and Shiny applications are developed to showcase the results of the study. In Chapter 6 (Paper 4), we analyze areal-level dengue data collected in Nepal between 2006 and 2022. The dengue analysis aimed to characterize multiple dengue outbreak intensity functions within each Nepalese district by assessing their timing,

duration, and potential associations with environmental climatic covariates.

Chapter 7 presents the conclusions of this doctoral thesis. The chapter also discusses the main contributions of each paper and possible areas for future research.

References

[1] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

[2] L. Zhao, B. Cao, E. Borghi, S. Chatterji, et al. "Data gaps towards health development goals, 47 low-and middle-income countries". In: *Bulletin of the World Health Organization* 100.1 (2022), p. 40.

[3] W. H. Organization. *World Health Statistics 2016 [OP]: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization, 2016.

[4] M. Li, I. Brodsky, and E. Geers. "Barriers to use of health data in low-and-middle countries a review of the literature". In: *MEASURE Evaluation* (2018).

[5] J. Nabyonga-Orem. "Monitoring Sustainable Development Goal 3: how ready are the health information systems in low-income and middle-income countries?" In: *BMJ global health* 2.4 (2017), e000433.

[6] J. Saghir and J. Santoro. "Urbanization in Sub-Saharan Africa". In: *Meeting Challenges by Bridging Stakeholders. Washington, DC, USA: Center for Strategic & International Studies*. JSTOR. 2018.

[7] A. Zerbo, R. C. Delgado, and P. A. González. "Vulnerability and everyday health risks of urban informal settlements in Sub-Saharan Africa". In: *Global Health Journal* 4.2 (2020), pp. 46–50.

[8] P. J. Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

[9] N. Cressie. "Statistics for Spatial Data". In: (1991).

[10] A. Baddeley, E. Rubak, and R. Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

[11] P. Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC, 2019.

[12] Y. Yamagata and H. Seya. *Spatial analysis using big data: Methods and urban applications*. Academic Press, 2019.

[13] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

Chapter 2

Review of statistical methods for modelling and mapping spatial data

In this Chapter, we provide a literature review of the types of spatial data, namely point pattern, geostatistical, and areal data. The Chapter also provides an overview of current modelling approaches for these types of spatial data.

2.1 Point pattern data

A spatial point process is a stochastic process comprising random variables $X_i = X_1, \ldots, X_n$ that have been observed at some locations in a study region A. A spatial point pattern is a realisation of the stochastic point process in $\mathbb{R}^2$ (i.e $A \subset \mathbb{R}^2$) [1, 2]. Although spatial point pattern data are uncommon in developing countries, they may arise from routinely collected data at health facilities, for instance, in a passive surveillance study. In these studies, the participants are asked to pinpoint the exact locations of their households. We assume that the households are the sources of exposure for the health outcomes of the study participants.

A spatial point pattern may contain several types of variables. For example, a point pattern dataset may contain the gender and HIV status of a tuberculosis

patient in addition to the geographical coordinates of their residence. Supplementary information of the points that provide further detail with respect to the individual or the location where a health outcome of interest or event occurred is called a mark [1]. Spatial point patterns with such variables are called marked point patterns. In addition to marks, spatial data may also contain variables referred to as covariates. Covariates are explanatory variables that are observed at all the spatial locations in the study area [1]. The surface amount of rainfall in an area and the elevation of the area are examples of spatial covariates.

2.1.1 First and second order properties

First and second order properties are used to characterise a spatial point process. A spatial point process which does not vary depending on the location or orientation is called a stationary and isotropic process.

Denoting $|dx|$ as the area for a location x and $N(dx)$ as the number of observed events in the area dx, the first order moment can be defined as follows:

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\} \tag{2.1}$$

where $\lambda(x)$, also called the intensity of the point process, is defined as the total number of events per unit area.

Second-order properties define the relationships between events in various sub-regions of the space domain. Letting $|dy|$ denote the area for a location y, the second order intensity function can be mathematically defined as follows.

$$\lambda_2(x, y) = \lim_{|dx|,|dy| \to 0} \left\{ \frac{E[N(dx, dy)]}{|dx||dy|} \right\} \tag{2.2}$$

The first and second-order properties of a point process form the basis for statistical analyses of point process data.

2.1.2 Checking for residual spatial correlation

The first step in analysing spatial point pattern data is to investigate the presence or absence of spatial correlation in the data. This is carried out using an inhomogeneous K-function, which is mathematically defined as [1]

$$\widehat{K}(r) = \frac{1}{D|W|} \sum_i \sum_{h \neq k} \frac{I\left\{||x_k - x_h|| \leq r\right\}}{\hat{\lambda}\left(x_k\right) \hat{\lambda}\left(x_h\right)}. \tag{2.3}$$

where: $D = \frac{1}{|W|} \sum_i 1/\hat{\lambda}\left(x_i\right)$; r is the distance at which the function is evaluated; $\hat{\lambda}(x)$ is the estimated intensity of the model at location x; and $I\left\{||x_k - x_h||\right\}$ is an indicator function that takes the value 1 if the absolute distance between two locations x_k and x_h is less than or equal to r, and 0 otherwise.

The inhomogeneous K-function is fitted under the assumption of complete spatial randomness (or spatial independence) in the data. An envelope is used to assess the assumption that the observed point process is an inhomogeneous Poisson process by following the steps below [2]:

(i) Generate spatial point patterns with the same intensity (λ) as the observed point pattern X in the study region A under the assumption of complete spatial randomness.

(ii) Estimate the K-function for each of the point patterns simulated in step (i).

(iii) Compute a 95% envelope for each K-function computed in step (ii).

The null hypothesis of complete spatial randomness is rejected if any part of the K-function falls outside the envelope simulated in the steps outlined above.

2.1.3 Model formulation

Given n set of events observed in an area A, we can assume that the set of events follows a Poisson distribution with a mean of λ. The point pattern data can, therefore, be modelled using the inhomogeneous Poisson process model, which has the following likelihood function [1]:

$$L(\lambda) = \sum_{i=1}^{n} \log \lambda\left(x_i\right) - \int_A \lambda(x) dx \tag{2.4}$$

In the likelihood function in equation 2.5, the mean of the Poisson process is defined as:

$$\lambda_i(x) = \exp\left(d(x)'\beta + \log(m_i(x))\right) \tag{2.5}$$

In the above equation, β denotes the vector of coefficients associated with a linear combination of spatial covariates, $d(x)$. Finally, $m_i(x)$ is an offset that corresponds to the population of an individual. This model can be fitted using the Spatstat package in R [3].

Some of the limitations of modelling point pattern data include the lack of a statistical model and software to model spatial and spatio-temporal point process pattern data with multiple marks [1]. The inhomogeneous Poisson process model in equation 2.5 with marks provides additional information about the underlying point process of the event being studied. Marks are only observed at the locations of events and are not defined throughout the entire study region, unlike covariates that are typically defined over the entire study domain [1]. Including marks as in the model as a linear covariate would, therefore, require extrapolating their values to non-event locations, which is generally not appropriate or feasible. The first paper of this thesis focuses on the extension of the model 2.5 to include marks as multiple intercepts associated with the health outcome of interest.

2.2 Geostatistical data

Geostatistical data are one of the most common types of spatial data available in developing countries. This is due to several donor-funded studies conducted routinely in these regions at household and health facility levels every 4 to 5 years. Examples of such studies include Population-based HIV Impact Assessments (PHIAs), Service Provision Assessments (SPAs), Multiple Indicator Cluster Surveys (MICS), and Demographic and Health Surveys (DHS), the latter of which are conducted in more than 50 developing countries across sub-Saharan Africa, South and Southeast Asia and other continents. To protect the anonymity of the study participants, household-level surveys only provide geolocations (longitude and latitude) of the study participants at fixed locations, such as clusters or

enumeration areas (usually a collection of households in an urban or rural area) in which the households are based.

Geostatistical data can be categorized into various forms, including count data, prevalence data, and continuous outcome data. Count data, which comprise the number of events (Y_i) that occur at a spatial location x, are modelled using the Poisson distribution. Also, continuous outcomes, such as lead pollution, are analysed using a Gaussian distribution. Prevalence geospatial data, p_i, which represents the number of individuals Y_i with a health outcome out of m_i number of people tested are commonly modelled using the Binomial distribution. The rest of this subsection focuses on models used to model and map Binomial prevalence data since these are the methods that were applied in Paper 4 and Paper 5.

2.2.1 Generalised linear binomial mixed effects model

In Binomial geostatistical data, given that Y_i individuals have a health outcome of interest out of m_i sampled individuals at a location x_i, the first step in modelling the geostatistical data is fitting a Binomial generalised linear mixed model. In the Binomial generalised linear mixed model, where the observations are conditional on mutually independent distributed Gaussian variables, Z_i, the logit linear predictor for prevalence ($p(x_i)$) at a given location is defined as

$$\log\left\{\frac{p\left(\mathbf{x}_i\right)}{1-p\left(\mathbf{x}_i\right)}\right\} = d\left(\mathbf{x}_i\right)\beta + Z_i \tag{2.6}$$

where $d\left(\mathbf{x}_i\right)$ is the vector of spatial explanatory variables and β is a vector of regression coefficients associated with the spatial covariates. In the Binomial generalised linear model in equation 2.6, Z_i has a mean of 0 and variance of τ^2. Then, the residuals from a generalised linear mixed model are investigated for the presence of spatial correlation.

2.2.2 Checking for residual spatial correlation

Residual spatial correlation in geostatistical data can be tested using the empirical variogram based on the random effects Z_i after removing the effect of covariates [4, 5]. A variogram for a spatial process is defined as:

$$\hat{V}(u) = \frac{1}{2|N(u)|} \sum_{(h,k)\in N(u)} \left(\hat{Z}_h - \hat{Z}_k\right)^2 \tag{2.7}$$

where $n(u)$ represents the collection comprising all neighbouring pairs separated by the distance u, and $|n(u)|$ indicates the count of unique pairs within $n(u)$ ($N(u) = \{(h,k) : \|x_h - x_k\| = u\}$).

Figure 2.1 illustrates the theoretical variogram ($V(u)$). In the diagram, an upward trend in the black solid line, as the distance $u = \|x_h - x_k\|$ increases, typically implies the lack of spatial independence in the data. The nugget variance (denoted as τ^2) in the figure denotes the value of the variogram ($\hat{V}(u)$) when the distance u equals zero [4]. The sill in the variogram captures the total variance (τ^2 plus the signal variance, σ^2) [4]. The practical range is the distance u at which the variogram plateaus (i.e. where the correlation function decays to 0.05) [4].

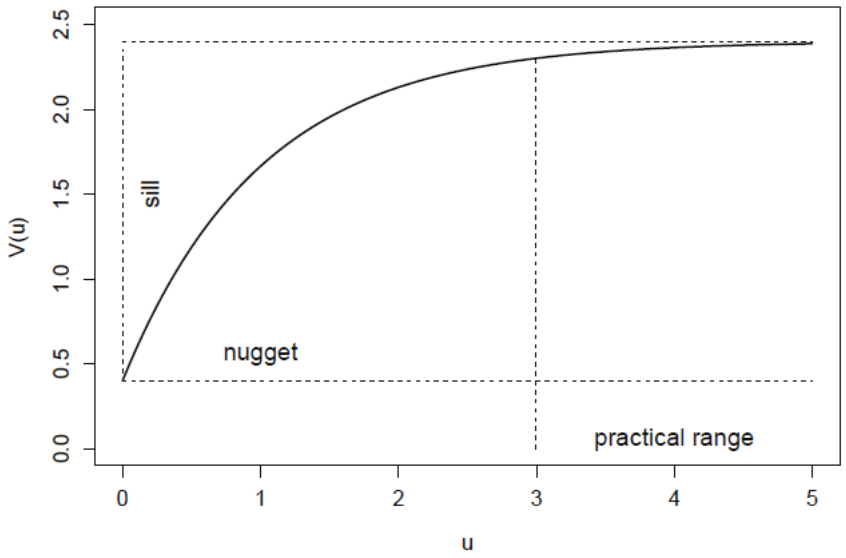

Figure 2.1: A diagrammatic illustration of a variogram (Source: Diggle and Giorgi, 2019 [4])

The following Monte Carlo procedure is used to assess the assumption of spatial independence in the data as follows:

(i) Randomly shuffle the labels of the variable Z_i while keeping the locations x_i constant.

(ii) Calculate the empirical variogram given in equation 2.7 using the Z_i permuted

in the above step.

(iii) Repeat steps (i) and (ii) B times.

(iv) Assuming spatial independence, utilize the B resulting empirical variograms from steps 1 and 2 to compute 95% intervals using a pre-defined distance.

The presence of unexplained residual spatial correlation in the data is established if the observed variogram does not lie within the 95% envelope computed in step 4 above.

2.2.3 Binomial geostatistical model

If the residual spatial correlation is present in the data, a geostatistical model, which is obtained by introducing a spatial Gaussian process, $S(\mathbf{x}_i)$ is fitted. Equation (2.6) can, therefore, be modified as follows:

$$\log\left\{\frac{p(\mathbf{x}_i)}{1-(\mathbf{x}_i)}\right\} = d(\mathbf{x}_i)\beta + S(\mathbf{x}_i) + Z_i \tag{2.8}$$

In the above equation, $S(\mathbf{x}_i)$ is a zero-mean stationary and isotropic Gaussian process with an exponential function with a mean of 0 and a variance σ^2. The covariance function of the process is therefore defined as:

$$\text{Cov}\left\{S(\mathbf{x}_h), S(\mathbf{x}_k)\right\} = \sigma^2 \exp\left\{-u_{hk}/\phi\right\}$$

where u_{hk} denotes any distance between any two areas $\mathbf{x}_h$ and $\mathbf{x}_k$ and ϕ is a scale parameter that determines the rate at which the spatial correlation decays to 0 as the distance u_{hk} increases. The model given in equation 2.8 is fitted using the Laplace approximation in the PrevMap package in R [6].

2.3 Lattice and areal data

As briefly discussed in Chapter 1, areal or lattice data are spatial data that are aggregated at some unit. In lattice data, we observe the realizations $Y(A)$, where A is a geographical unit (village, district, region or country) that forms part of the study region ($A \in \mathbb{R}^2$). Lattice data, consequently, refer to the set of spatial units that are organized in a regular grid, whilst areal data refer to the set of spatial

units that are irregularly shaped [7]. Areal data, which are often defined by natural or man-made boundaries such as administrative boundaries, are common in health research compared to lattice data. Data are aggregated at areal units such as villages, counties or districts to protect the study participants' privacy [7].

In Chapter 6 (Paper 4), we depart from typical spatial analyses of areal count data by focusing on each areal unit over time, treating the data as temporal rather than spatial. This approach is prompted by our finding that there was no strong evidence of residual spatial correlation in the dengue data used in Paper 4. Therefore, we model the temporal count data of reported dengue cases collected over 17 years in 77 Nepalese districts.

2.3.1 Model formulation

Assuming that the observed outcomes in a geographical unit such as a district (Y_t, t=1,2,...,T) are distributed as a Poisson random variable ($Y_t \sim \text{Pois}(\lambda_t)$) where λ_t is the mean number of dengue cases at time t, and is expressed as:

$$\lambda_t = \exp\left(\boldsymbol{d}_t\boldsymbol{\beta}\right) \tag{2.9}$$

where β are the coefficients associated with the spatial covariates in the matrix ($\boldsymbol{d}_t$).

One of the limitations of using the Poisson model, especially on data collected from health facilities or data for low-incidence diseases such as dengue, is overdispersion [8]. Equidispersion, one of the assumptions of the Poisson model, is a concept where we assume that the mean and variance of the data are equal. We use a Negative Binomial model in this study to overcome the equidispersion challenge. The Negative Binomial model extends the Poisson model by explicitly including a dispersion parameter, denoted by α. The Negative Binomial likelihood function is specified as follows [9]:

$$L = \prod_{t=1}^{T} p\left(y_t\right) = \prod_{t=1}^{T} \frac{\Gamma\left(y_t + 1/\alpha\right)}{\Gamma\left(y_t + 1\right)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\lambda_t}\right)^{1/\alpha} \left(\frac{\alpha\lambda_t}{1+\alpha\lambda_t}\right)^{y_t} \tag{2.10}$$

Although the most common analysis method for spatial and spatio-temporal areal level data is the Besag-York-Mollié (BYM) model [10], this thesis focuses on the identification of outbreaks in areal-level data. This thesis, therefore, provides a brief overview of the current modelling approaches for outbreak data for estimating various parameters for disease outbreak data.

A common model for areal-level spatio-temporal data are HHH4 models that decompose the data into endemic and epidemic components [11]. The challenge, however, is that the model requires more granular data at weekly or monthly intervals. Similarly, a HHH4ZI model extends the HHH4 model by explicitly including a zero inflation parameter to account for excess zeroes [12]. However, both the HHH4 and HHH4ZI models require weekly or monthly data to estimate seasonality trends. Both models, also, do not estimate the duration and intensity of the outbreak intensity functions.

Another model for characterizing outbreaks is the two-stage cluster hierarchical model [13]. This model is advantageous over the above models because it helps identify multiple disease clusters. The clusters are, however, only estimated in space and not time. Similarly, Ramadona et. al developed a Bayesian spatio-temporal model to estimate the timing of dengue outbreaks in Indonesia [14]. Their spatio-temporal model employed two adjacency matrices: one based on geographical proximity and the other on human mobility patterns [14]. The study found that the matrix incorporating human mobility patterns outperformed the geographical proximity-based one [14]. The human mobility patterns recommended from their study were, however, not available for our dengue data. Furthermore, both the two-stage cluster hierarchical model and Bayesian spatio-temporal model estimate the outbreaks within a year, and hence require granular data which was not available in our study [13, 14].

In addition to the models employed by Anderson et. al (2014) and Ramadona et. al (2023), previous approaches, such as the one used by Guzman et al., require data to be aggregated into spatio-temporal blocks (e.g., at 1 week, 3 weeks, and 5 weeks) to investigate outbreaks within those blocks [13–15]. In contrast, our model

sought to contribute to the outbreak data modelling knowledge base by providing insights into the duration and size of each outbreak. Importantly, our model does not require aggregation of data into blocks, and it also offers a further understanding of dengue outbreak dynamics by estimating both the scale parameter of each outbreak intensity function (OIF) and approximating the contribution of each outbreak to the overall dengue epidemic.

This thesis extends the Negative Binomial model in equation 2.10 to include the estimation of multiple dengue outbreaks in Nepal. Specifically, our model allows for estimating the timing and duration of multiple dengue outbreaks in each Nepalese district. We also propose how the model can be extended in future research, especially in cases where more than three outbreaks are suspected to have occurred in an area. Our proposed model is particularly useful in cases where granular data at daily, weekly, or monthly intervals is unavailable, as existing modelling approaches typically require such high-resolution temporal data.

References

[1] A. Baddeley, E. Rubak, and R. Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

[2] P. Moraga. *Spatial Statistics for Data Science: Theory and Practice with R*. CRC Press, 2023.

[3] A. Baddeley and R. Turner. "Spatstat: an R package for analyzing spatial point patterns". In: *Journal of statistical software* 12 (2005), pp. 1–42.

[4] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

[5] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

[6] E. Giorgi and P. J. Diggle. "PrevMap: an R package for prevalence mapping". In: *Journal of Statistical Software* 78 (2017), pp. 1–29.

[7] A. F. Zuur, E. N. Ieno, and A. A. Saveliev. "Spatial, temporal and spatial-temporal ecological data analysis with R-INLA". In: *Highland Statistics Ltd* 1 (2017).

[8] Y. Chen, T. Liu, X. Yu, Q. Zeng, et al. "An ensemble forecast system for tracking dynamics of dengue outbreaks and its validation in China". In: *PLoS computational biology* 18.6 (2022), e1010218.

[9] M. Zwilling. "Negative binomial regression". In: *The Mathematica Journal* 15 (2013).

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

[11] S. Meyer, L. Held, and M. Höhle. "hhh4: Endemic-epidemic modeling of areal count time series". In: *J Stat Softw* 1 (2016), pp. 1–55.

[12] J. Lu and S. Meyer. "A zero-inflated endemic–epidemic model with an application to measles time series in Germany". In: *Biometrical Journal* 65.8 (2023), p. 2100408.

[13] C. Anderson, D. Lee, and N. Dean. "Identifying clusters in Bayesian disease mapping". In: *Biostatistics* 15.3 (2014), pp. 457–469.

[14] A. L. Ramadona, Y. Tozan, J. Wallin, L. Lazuardi, et al. "Predicting the dengue cluster outbreak dynamics in Yogyakarta, Indonesia: a modelling study". In: *The Lancet Regional Health-Southeast Asia* 15 (2023).

[15] L. M. G. Rincón. "Statistical Methods for Campylobacter Outbreak Detection using Genomics and Epidemiological Data". PhD Thesis. University of Warwick, 2020.

Chapter 3

Paper 1: Modelling *Salmonella* Typhi in high-density urban Blantyre neighbourhood, Malawi, using point pattern methods

Jessie J. Khaki [1,2,3], James E. Meiring[4], Deus Thindwa[5], Marc Y. R. Henrion[2], Tikhala M. Jere[2], Harrison Msuku[2], The STRATAA Consortium, Robert S. Heyderman[6], Melita A. Gordon[2,7], Emanuele Giorgi[1].

[1] Lancaster Medical School, Lancaster University, Lancaster, United Kingdom.

[2] Malawi-Liverpool-Wellcome Trust Programme, Blantyre, Malawi.

[3] School of Global and Public Health, Kamuzu University of Health Sciences, Blantyre, Malawi.

[4] Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, United Kingdom.

[5] Department of Epidemiology of Microbial Diseases, Yale University, New Haven, United States of America.

[6] Division of Immunity and Infection, University College London, London, United Kingdom.

[7] Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, United Kingdom.

Summary

Salmonella Typhi is a human-restricted pathogen that is transmitted by the fecal-oral route and causative organism of typhoid fever. Using health facility data from 2016 to 2020, this study focuses on modelling the spatial variation in typhoid risk in the Ndirande township in Blantyre. To pursue this objective, we developed a marked inhomogeneous Poisson process model that allows us to incorporate both individual-level and environmental risk factors.

The results from our analysis indicate that typhoid cases are spatially clustered, with the incidence decreasing by 54% for a unit increase in the water, sanitation, and hygiene (WASH) score. Typhoid intensity was also higher in children aged below 18 years than in adults. However, our results did not show evidence of a strong temporal variation in typhoid incidence. We also discuss the inferential benefits of using point pattern models to characterise the spatial variation in typhoid risk and outline possible extensions of the proposed modelling framework.

Keywords: spatial point patterns; inhomogeneous Poisson model; typhoid; mapping; incidence.

3.1 Introduction

Salmonella enterica serovars Typhi (*S.* Typhi) is a human-restricted pathogen transmitted by faeco-oral route and the causative organism of typhoid fever. *S. Typhi* is estimated to cause more than 10.9 million cases each year, with about 116,000 of the cases resulting in death [1, 2]. Whilst the global incidence of typhoid is estimated at 293 cases per 100,000 person-years, the highest burden of typhoid is reported to be in resource-constrained settings, particularly in sub-Saharan Africa and South Asia [2, 3]. A meta-analysis in 2017 estimated a typhoid incidence of 149 cases per 100,000 person-years in southern sub-Saharan Africa, whilst South Asia was estimated to have a typhoid incidence of 204 cases per 100,000 person-years [2].

Typhoid is primarily transmitted when a healthy person comes into contact with stool-contaminated food or water [3–5]. Inadequate access to clean water and sanitation are thus two of the main risk factors associated with typhoid [6]. One study has indeed shown that, in Malawi, typhoid risk is highly affected by the type of water that a household uses for cooking and cleaning [5]. Elevation also plays an important role in the risk of typhoid infection. A study in Kenya showed that individuals, particularly children, living in low-elevation areas were twice more likely to contract typhoid than people living at higher elevations [4]. This can be explained by the accumulation of faecal waste in low-elevation areas due to the downstream flow of contaminated water [4]. Recent studies [3, 7] have also reported that rainy seasons are associated with an increased risk of typhoid, suggesting that the occurrence of typhoid follows a seasonal pattern with variations dependent on the climatic and environmental conditions of the region. On the other hand, heavy-intensity rainfall is shown to have a negative association with typhoid incidence as the high-intensity rainfall may wash away faecal substances [7].

The risk of typhoid also varies across different groups of age and gender. Several studies have shown that the burden of typhoid is highest among children between 5 and 19 years, an age group typically identified as school-going children [1]. A study in Blantyre, Malawi, showed that the highest typhoid-attributable risk

percentage among the children in the study arose from spending a day in a daycare or school [5]. This result is in agreement with the results from another study where the incidence of typhoid was highest among children aged 5 to 9 years, followed by those aged between 2 and 4 years [8]. Evidence of the effect of gender on typhoid is, on the other hand, contradictory. While other studies have shown that both occurrences of typhoid and mortality due to typhoid are higher among males [1], others have reported a higher occurrence of typhoid among females [9].

Typhoid is monitored using passive or enhanced surveillance methods depending on a country's level of endemicity and public health objectives. The World Health Organisation (WHO) recommends that endemic countries such as Malawi should have, as a minimum, laboratory and facility-based surveillance [10]. The surveillance can be carried out through passive reporting of results from the laboratory, the establishment of a surveillance system, or active review of laboratory records to find patients whose results meet the criteria for a confirmed typhoid case [10]. The WHO, additionally, recommends surveillance through population-based studies to estimate the population-based incidence of a country and generate information for programmatic interventions [10]. In this study, we used data collected from a passive surveillance study in Malawi [6, 11, 12].

In Malawi since 1998, blood cultures have been routinely collected from febrile patients at Queen Elizabeth Central Hospital (QECH) in Blantyre [13]. A study showed that an average of 14 cases per year were recorded between 1998 and 2010 at QECH [14]. The same study also reported a rapid increase in typhoid cases starting from 2011, with a peak observed in 2014 at 782 cases [14]. The outbreak of typhoid in both Malawi and other African countries is due to a multidrug-resistant (MDR) typhoid strain to ampicillin, chloramphenicol, and cotrimoxazole that originated in Asia [14, 15]. The escalating issue of antimicrobial resistance (AMR) is a threat to global health as current drug AMR trends may hinder efforts to control typhoid through antibiotic treatment and lead to an increase in the risk of typhoid worldwide [16].

Understanding the spatial variation in the risk of typhoid can help to identify

disease hotspots and develop more targeted control interventions. Spatial and spatio-temporal statistics can thus play a critical role by utilising information across time and space and making the best use of data from constrained resource settings. Among previous typhoid research, some studies have used a quasi-Poisson generalised linear model and an over-dispersed Poisson generalised linear model to assess the relationships between typhoid and climatic variables, such as temperature and rainfall [7, 17]. Another previous study in Blantyre, Malawi, used geostatistical methods to model and map the inhomogeneous distribution of typhoid genomic data [18]. Similarly, a study from Ghana has shown that typhoid incidence at the district level exhibits spatial and temporal patterns and modelled that using negative binomial autoregressive moving average model [19]. Another study in Uganda used a spatial scan statistic for incidence to identify hotposts and a standard Poisson model with no overdispersion to investigate spatio-temporal trends of typhoid [20]. One of the main drawbacks of spatial scan statistics is the inability to correctly identify non-circular or irregularly shaped clusters [21]. Our work builds on the current literature by developing a spatially explicit statistical model for point pattern process typhoid data.

The focus of this paper is to develop a spatial point pattern model to assess the effect of environmental and individual risk factors on typhoid fever, using health facility data. To the best of our knowledge, this is the first study that uses spatial point pattern models for the analysis of geo-located typhoid cases. This work, therefore, extends prior research on geostatistical modelling of typhoid data in Blantyre, Malawi, by modelling geo-located households using both individual-level and spatial covariates in the modelling [18]. The specific objectives of the study were as follows:

- to investigate the association between spatial and temporal covariates with the occurrence of typhoid in Ndirande township after adjusting for individual-level markers, namely age and gender; and
- to investigate spatial and temporal trends of typhoid in Ndirande township.

3.2 Methods

3.2.1 Study site

The study was conducted in Ndirande township in Blantyre city in Malawi between October 2016 and February 2020. Ndirande, which had a population of about 100,000 people in 2018, spans an area of approximately 6.7 km^2 and is serviced by one government health clinic [6]. Blantyre city, which is in the southern part of Malawi, lies 35° east of Greenwich Meridian and 15° 42" south of the Equator. Blantyre city was selected for the study because of the well-known high burden of typhoid fever and the research capacity to carry out complex studies [6].

Malawi has two main climate seasons: the rainy and dry seasons. The rainy season can be further distinguished between the early rain (November to February) and the late rain (March to April) seasons [7]. Similarly, the dry season can also be distinguished into the cool dry (May to August) and the hot dry (September to October) seasons [7]. A recent study protocol reported that the number of typhoid cases per month in Ndirande township in Blantyre district in Malawi increased in the months of December through February, which corresponds to the rainy season in Malawi [6]. Ndirande exhibits a variation in elevation, ranging from 970 to 1,200 meters, with a median elevation of 1,118 meters. Total precipitation also varied from 819 millimeters (mm) to 1,602 mm from 2016 to 2019. The variation in total precipitation across Ndirande was, however, minimal with the maximum difference being 209 mm each year. In this study, we included season as a temporal covariate in our modelling.

3.2.2 Data

3.2.2.1 Passive surveillance study of the STRAATA project

The Strategic Typhoid Alliance across Africa and Asia (STRATAA) study was carried out in Bangladesh, Nepal and Malawi with the aim of measuring the burden of typhoid in these three sites [6]. In Malawi, the STRATAA study was carried out by the Malawi-Wellcome-Liverpool Clinical Research Programme at the government-run Ndirande health clinic, which is the largest clinic in Ndirande

township. In this paper, our focus is on the passive surveillance sub-study of the STRAATA project.

In the passive surveillance study, patients presenting with a history of fever for at least 2 days or a patient presenting with a temperature of at least 38.0°C at the Ndirande health clinic were approached with the intention of enrolling them into the study [6, 12]. Passive surveillance was, additionally, performed at Queen Elizabeth Central Hospital (QECH) for patients from Ndirande who presented to the Accident and Emergency Treatment Centre (AETC) or were admitted to the wards [12]. A blood culture was collected from the patients who consented to be enrolled in the study. A total of 161 typhoid cases were recorded at Ndirande health clinic in a passive surveillance study between October 2016 and February 2020. The gender and age of the study participants were collected as part of the routine data collected in the study. However, 1 case did not have a date of collection and was therefore excluded from the analysis. Handheld Global Positioning Systems (GPS) devices were used to collect the locations (latitude and longitude) of the households of the typhoid cases.

Two marks, namely the gender (male or female) and age in years of a typhoid case were included in our model. Age was categorised into 3 levels (0 to 5, 6 to 17 and 18+ years) given previous studies on the association between typhoid and several age groups [5, 8].

3.2.2.2 Population data

The STRATAA study also carried out household and individual-level population censuses in 2018. The population census, which enumerated 102,242 individuals, was used as an offset in the model.

3.2.3 Spatial covariates

Covariates selection was informed by previous research on the associations between typhoid and environmental covariates [4, 5, 17, 22, 23]. For this study, we restricted our attention to those covariates that are available at a spatial resolution of 100 m^2 for Ndirande. Hence, our spatial covariates are: distance to Ndirande

health clinic in meters, elevation (in meters) and a Water, Sanitation and Hygiene (WASH) score.

The distance to the health clinic raster was derived by calculating the Euclidean distances from each location within Ndirande township to the health clinic. The elevation raster file was downloaded from the WorldPop website [24]. The raster was cropped to a 100 m^2 Ndirande grid.

A water, sanitation, and hygiene (WASH) survey was carried out in 14,136 households in Ndirande township in 2018 as part of the STRATAA study. The WASH variables were self-reported in the questionnaire. A WASH score was derived using principal components analysis (PCA), and a linear geostatistical model was used to interpolate the WASH score over the grid. Further details on the spatial covariates, including how the WASH score was derived, are supplied in the supplementary material A.

3.2.4 Modelling of reported typhoid fever cases using point-pattern models

We develop an inhomogeneous spatial marked point process that allows us to incorporate both spatial covariates and individual-level covariates as marks [25]. Let i denote the subscript for gender, with $i = 1$ corresponding to "male" and $i = 2$ to "female". We then use j to denote the subscript that identifies a specific age group, $j = 1$ representing individuals between 0 and 5 years, $j = 2$ between 6 and 17 years, and $j = 3$ for those above 17 years. Our outcome variable corresponds to the locations of the reported diagnosed cases x that fall in A, representing the area encompassed by the boundaries of Ndirande township. It, therefore, follows that n_{ij} corresponds to the number of typhoid cases in a specific age-gender combination. By setting age and gender as marks, we model the cases reported within each age-gender subgroup as independent inhomogeneous Poisson processes. More specifically, we model the intensity of the subgroup for gender i and age j as $\lambda_{ij}(x) = \exp\left(\alpha_i + \gamma_j + d(x)'\beta + \log m_{ij}(x)\right)$. In the above equation, we use α_i to account for gender effect and γ_j to account for differences across age groups. The vector $d(x)$ denotes a linear combination of spatial covariates: distance, measured in meters, to Ndirande health clinic (β_1); elevation, in meters

(β_2); and the WASH score (β_3). Finally, $m_{ij}(x)$ is an offset corresponding to the population for an individual with gender i and age j at location x.

We denote the vector of unknown parameters with θ, which consists of intercepts quantifying the gender effects (α_i, for $i = 1, 2$) and age effects (γ_j, for $j = 1, 2, 3$) and the regression coefficients β. The likelihood function for θ is then given by

$$L(\theta) = \sum_{i=1}^{2} \sum_{j=1}^{3} L_{ij}(\theta) \tag{3.1}$$

where

$$L_{ij}(\theta) = \sum_{k=1}^{n_{ij}} \log \lambda_{ij}(x_k) - \int_A \lambda_{ij}(x)\, dx \tag{3.2}$$

We use a quadrature procedure to approximate the integral in equation 3.2 based on a 100 m by 100 m regular grid of the study area denoted as A [26]. To obtain confidence intervals for the parameters θ, we use parametric bootstrap [27] based on the following iterative steps.

1. Simulate N= 10,000 samples from the fitted point process model with mean:

$$\lambda_{ij}(x) = \exp\left(\alpha_i + \gamma_j + d(x)'\beta + \log m_{ij}(x)\right) \tag{3.3}$$

2. Fit the model to the N bootstrap realisations simulated in step (1).
3. Store parameter estimates from each of the fitted models.
4. Use the percentile method to get a 95% confidence interval from the estimates stored in step (3).

We fitted both a spatial model (3.2) and spatio-temporal model (equation A.3 in Appendix A) to our data. We tested for temporal trends in the data by comparing the purely spatial model and model with temporal covariates using a likelihood ratio test under the null hypothesis that the spatial model should be used to fit the data.

We computed predicted incidence rates for each combination of marks (age and gender) while adjusting for the spatial covariates and population as defined in the intensity equation above $\left(\lambda_{ij}(x) = \exp\left(\alpha_i + \gamma_j + d(x)'\beta + \log m_{ij}(x)\right)\right)$. In addition to plotting the age and gender predicted incidence rates on the 100m by

100m regular grid, we also estimated the area-wide incidence for Ndirande, defined as

$$\frac{\int_A \lambda_{ij}(x)dx}{\int_A m_{ij}(x)dx}. \tag{3.4}$$

The integrals in equation 3.4 were approximated using a regular grid with a spatial resolution of 100m by 100m.

3.2.4.1 Model validation

To validate the compatibility of the spatial point pattern model presented in the previous section with the data, we develop a simulation procedure based on the K-function, which is expressed as [28]

$$\widehat{K}(r) = \frac{1}{D|W|} \sum_h \sum_{h \neq k} \frac{I\left\{||x_k - x_h|| \leq r\right\}}{\hat{\lambda}\left(x_k\right)\hat{\lambda}\left(x_h\right)}. \tag{3.5}$$

where: $D = \frac{1}{|W|}\sum_h 1/\hat{\lambda}\left(x_h\right)$; r is the distance at which the function is evaluated; $\hat{\lambda}(x)$ is the estimated intensity from the model at location x; and $I\left\{||x_k - x_h||\right\}$ is an indicator function that takes the value 1 if the absolute distance between any two locations x_k and x_h is less or equal to r, and 0 otherwise.

We then validate our model using the following bootstrap procedure.

1. By plugging in the maximum likelihood estimate for θ, simulate a data set based on the inhomogenous marked point process defined in the previous section.
2. Compute the inhomogeneous K-function defined in equation 3.5 for the simulated data set in the previous step.
3. Repeat steps (i) and (ii) 10,000 times.
4. For a set of predefined distances r compute the 95% confidence intervals using the 10,000 functions obtained from the previous steps.

On completion of the last step, we then conclude that the data do not show evidence against the fitted model if the K-function computed on the original data falls within the 95% envelope for each of the age-gender combinations.

3.2.5 Ethics consideration

The Oxford Tropical Research Ethics Committee (reference number 39-15) and the Malawian National Health Sciences Research Committee (reference number 15/5/1599) gave the approval to conduct the STRATAA study (trial number ISRCTN 12131979) in Malawi [6]. At the household level, the head of the household provided written informed consent for household surveys on behalf of the entire household. In the other components of the STRATAA study, an informed consent form was signed by study participants aged at least 18 years. On the other hand, informed consent forms were signed by parents or guardians of children less than 18 years old. Assent was, additionally, sought from children aged between 11 and 17 years. We confirm that the methods performed in this study were conducted in accordance with appropriate regulations and guidelines. Furthermore, we confirm that the study complies with the Declaration of Helsinki.

3.3 Results

A total of 161 typhoid cases were recorded at Ndirande Health clinic between October 2016 and February 2020. Out of these, only 1 case did not have complete information on age, gender and the date of sample collection. The analysis presented is thus based on the 160 typhoid cases with no missing data. A total of 43% (n=69) of the study participants were aged between 6 and 17 years. The median age of the study participants was 11 years (interquartile range: 6 to 21 years). Further, 52% (n=83) of the sample were females. Figure 3.1 shows the distribution of typhoid cases by gender in Ndirande. Table 3.1 further summarises the characteristics of the sample.

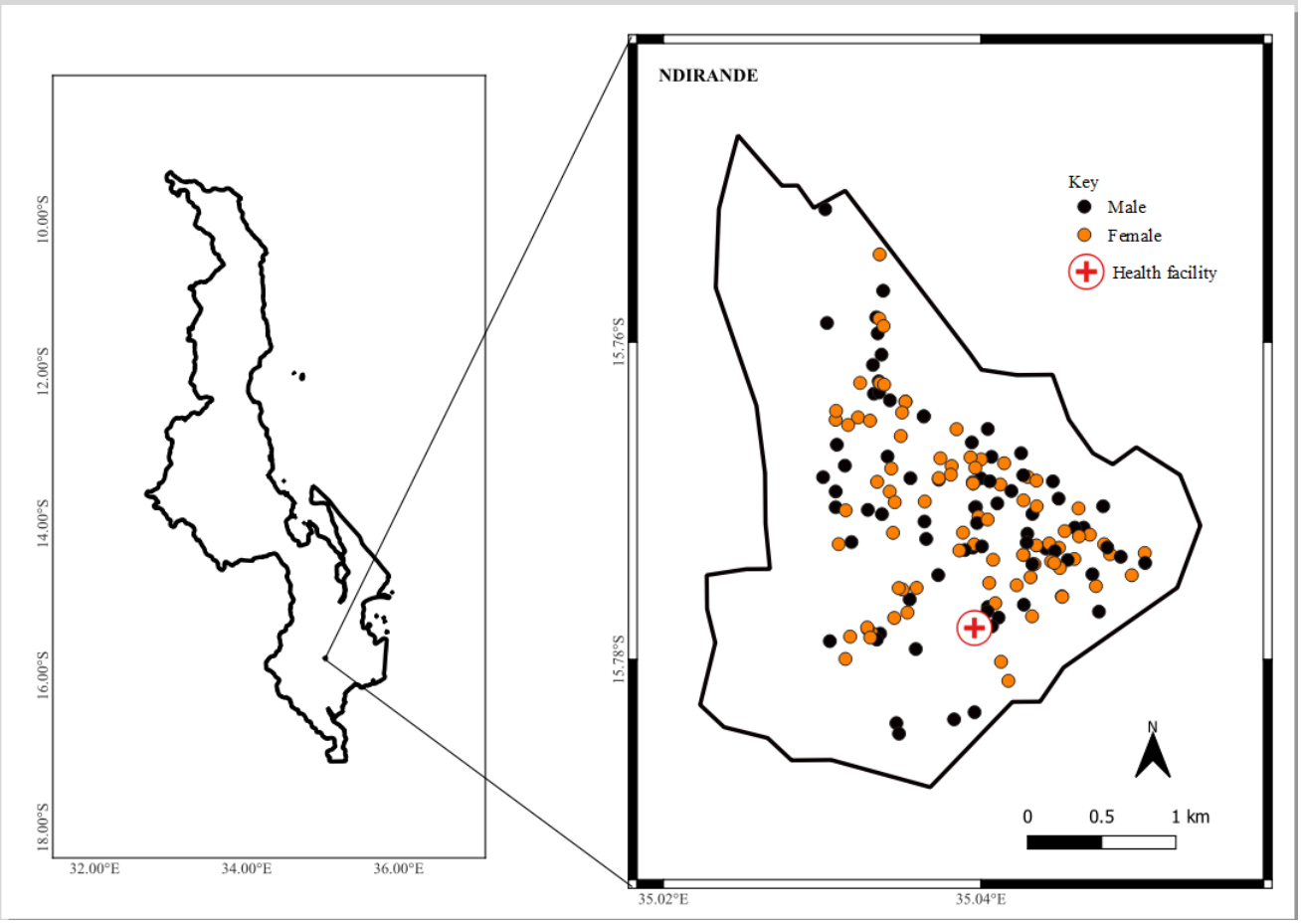

Figure 3.1: Locations of 160 typhoid cases and Ndirande health clinic from October 2016 to January 2020. The shaded area represents the study region.

Table 3.1: Distribution of the study participants.

Variable	**Total (n)**	**Percentage (%)**
Age (median, IQR)	11 years (6 to 21 years)	
Age (years)		
0-5	32	20%
6-17	69	43%
18+	49	37%
Gender		
Male	77	48%
Female	83	52%

Figure 3.2 illustrates the typhoid cases recorded per season from October 2016 to February 2020. This plot does not show any discernible temporal pattern.

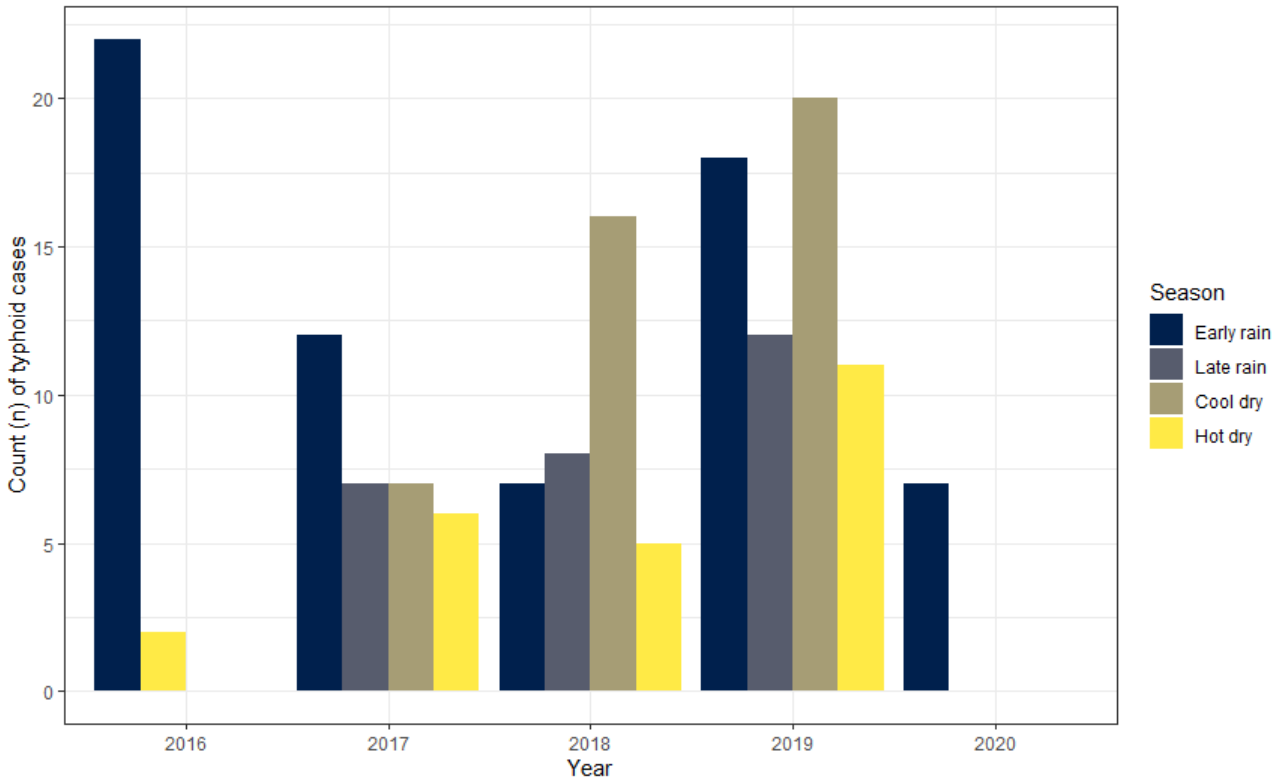

Figure 3.2: Observed typhoid cases per season from October 2016 to February 2020.

Our study results show that a 50 meters increase in the distance away from the health clinic decreased the estimated incidence rate of typhoid by 1% (100 * {1 - exponent of coefficient (coef): -0.01}, 95% confidence interval (CI): -0.03, 0.01). Further, a 50 meters increase in the elevation decreased the estimated incidence rate of typhoid by 9% (coef: -0.10, 95% CI: -0.42, 0.12). With further regard to the spatial covariates, a one-unit increase in the WASH score was associated with a decrease in the incidence rate of typhoid of 54% (coef: -0.78, 95% CI: -1.34, -0.45). We find that only the WASH score shows a significant effect at the 5% conventional confidence level. However, all the point estimates of the regression component align with the expected direction, as informed by our understanding of typhoid fever epidemiology.

Predicted relative intensities were computed and plotted for each combination of marks (age and gender) while adjusting for the spatial covariates and population. Figure 3.3 shows the average predicted reported incidence for males and females of any age at any point in time in the study per 100,000 population. As can be seen

Table 3.2: Maximum likelihood estimates and 95% confidence intervals (CI) for the parameters of the model specified in (3.3).

Variable	**Estimate**	**95% CI**
Age (years)		
0-5	-3.119	(-5.147, -0.193)
6-17	-3.162	(-5.189, -0.230)
18+	-3.906	(-5.929, - 0.973)
Gender		
Male	-5.140	(-8.177, -0.746)
Female	-5.047	(-8.087, -0.652)
Spatial covariates		
Distance to health facility × 50 meters	-0.010	(-0.027, 0.008)
Elevation × 50 meters	-0.098	(-0.420, 0.123)
WASH score	-0.782	(-1.338, -0.449)

in Figure 3.3, the areas with the highest typhoid risk were the central and southeast areas of Ndirande. The highest predicted reported incidence overall was in females (400 typhoid cases per 100,000 population) and males (365 typhoid cases per 100,000 population) aged between 0 and 5 years. This finding concurs with the model coefficients reported in Table 3.2. When comparing the adjusted predicted reported incidences within each gender, the 0-5 age group had the highest predicted relative intensity for both males and females per 100,000 population per month, as shown in Table 3.3.

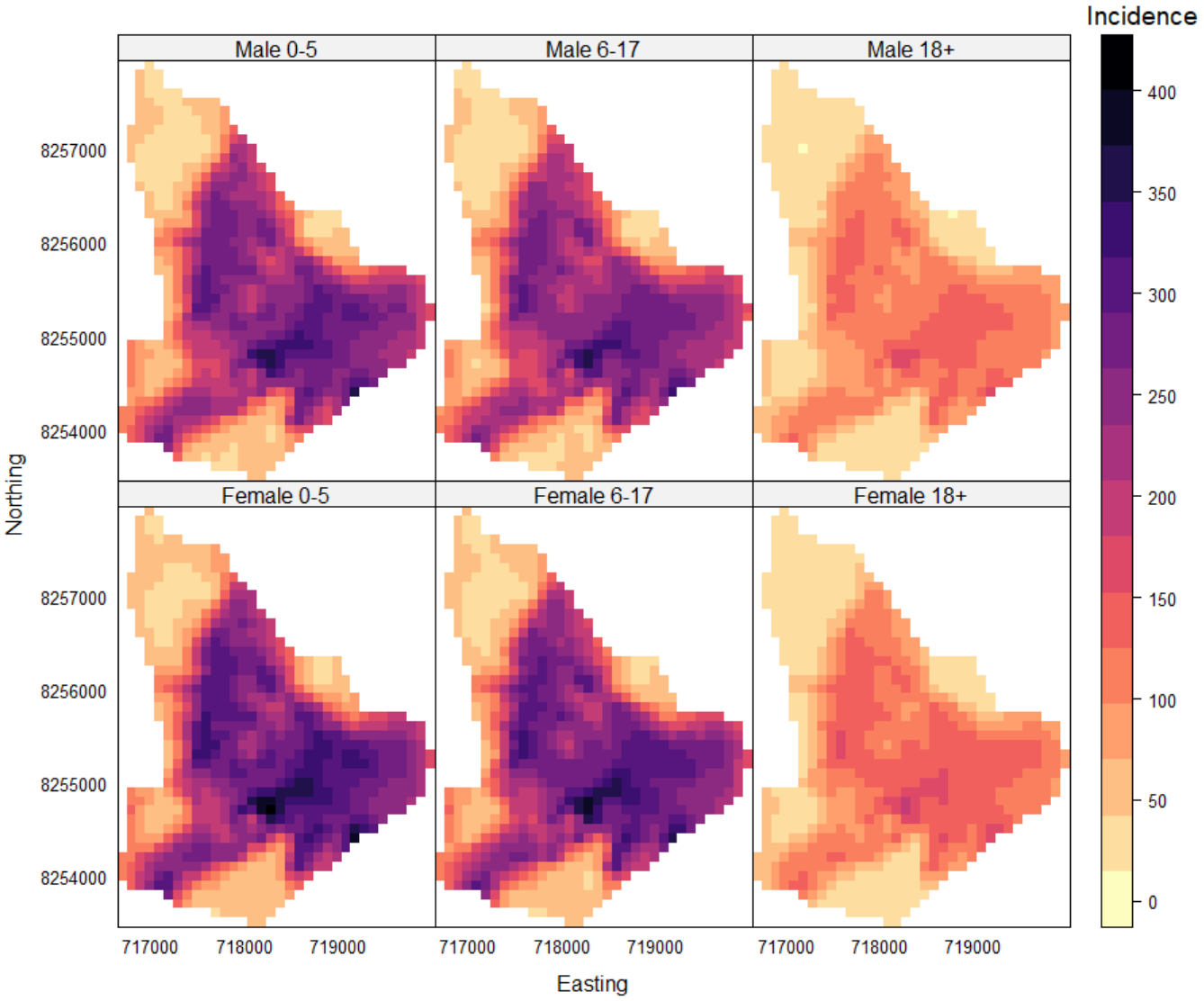

Figure 3.3: Predicted incidence of typhoid by gender and age per 100,000 population. The rows represent the gender of a typhoid case, whilst the columns represent the age group of the case.

We fitted an inhomogeneous K-function to validate our spatial point pattern model. The model validation plots for the final model are attached in the supplementary material A. Overall, the figures show that the K-functions from the observed data mostly fell within the simulated envelope for most of the distances. This suggests that our model was a good fit for the data.

Table 3.3: Predicted incidence and 95% confidence intervals (CI) per 100,000 population for Ndirande; for the definition of the predictive target in equation 3.4.

Group	**Number**	**Incidence rate**	**95% CI**
Male 0-5	14	222	(219, 224)
Male 6-17	36	216	(215, 216)
Male 18+	27	104	(103, 105)
Female 0-5	18	240	(238, 242)
Female 6-17	33	237	(236, 237)
Female 18+	32	114	(113, 115)

3.4 Discussion

In this study, we have shown how spatial point pattern methods can be used to analyze reported cases of typhoid fever in health facilities. Our approach based on a multiple-marked inhomogeneous Poisson process model allowed us to estimate typhoid incidence at the household level while adjusting for both spatial and individual-level risk factors.

Several modelling challenges were encountered in the analysis. First, the small number of reported cases across time and space makes it more challenging to model the relationships between risk factors and overall incidence patterns. In this context, the interpretation of the regression relationships should not only be guided by statistical summaries, such as *p-values*, but prior knowledge about the disease context should also be used to inform the selection of covariates. For this reason, we decided to retain variables that were not statistically significant, namely distance to a health facility, and elevation, to generate the spatial predictions for typhoid fever incidence. Our general guiding principle is that a variable should be retained in the final model, regardless of its statistical significance, 1) if there is an established body of evidence on the importance of the variable to model the health outcome of interest, and 2) if the point estimate is in accordance with the expected direction of the relationship based on that prior knowledge. In the case of the three variables considered, it has been established in previous research that these three variables are important risk factors for typhoid

[4, 5, 22, 23] and both of the aforementioned criteria are met.

Based on the effects of these risk factors, the southeast zone of Ndirande was found to show the highest typhoid incidence rate. This area of Ndirande is characterized by a high population density which could contribute to poor sanitary facilities as indicated by the WASH poor facilities. Our incidence map provides a more granular distribution of typhoid compared to previous work [18]. The finding on typhoid incidence decreasing with good WASH facilities is in line with the findings from another study carried out in the Blantyre district in Malawi [5]. The result of an increase in the elevation being associated with a decrease in the incidence of typhoid is also consistent with results from previous studies [4, 23]. The maximum distance observed between the health center and the study area was recorded as 3.1 km. Our results further showed that an increase in the distance to the Ndirande health clinic was associated with a decrease in the reported incidence of typhoid. This suggests that people living far away may be more reluctant to go to the clinic unless they are seriously ill [22]. It is important to note, however, a potential limitation of these findings. The GPS coordinates used in this study were collected at the household level, and thus may not reflect the true locations of the exposure to typhoid.

In addition to the spatial (environmental) risk factors, the age of an individual is found to play an important role in the variation of typhoid risk. Our study findings indicate a higher occurrence of typhoid among children after adjusting for the spatial covariates. This result is consistent with previous studies that also reported a higher typhoid incidence among children compared to adults [2, 14, 18, 29]. The estimated typhoid intensities for the 3 age groups in this study are, however, lower than the adjusted typhoid incidences recently reported in Blantyre in Malawi because we did not adjust the incidence in our study by a number of factors such as blood culture sensitivity and healthcare-seeking probability [12, 30]. In contrast to previous studies, we did not find any statistically significant difference in the estimated incidence between females and males [31, 32].

Another important limitation of this study is the under-reporting arising from

passive surveillance data collected from individuals who visit a health facility [1, 33, 34]. To account for the under-reporting, our model can be extended in future work using a thinned inhomogeneous Poisson process model, whereby the intensity of the Poisson process is scaled by the probability of visiting the health centre [28]. However, one of the challenges of this approach is that some covariates may affect both typhoid fever risk and the probability of visiting a clinic, making the estimation of regression relationships more problematic. This issue has also been reported in ecology, where similar methods have been used in citizen science data [35]. Future research should focus on a better understanding of the factors and mechanisms that drive the likelihood of attending health facilities, to better parameterise the probability of going to the hospital and overcome the identifiability issues in the estimation.

The proposed modelling approach in this study may be applied to the analysis of reported cases from passive surveillance data for other diseases. One of the strengths of the illustrated modelling approach is its flexibility in being adapted to any other environmentally driven diseases through the selection of suitable covariates. Through the application of this approach, we have further demonstrated that, for example, typhoid occurrence is higher among children and in areas with households with poor WASH facilities. Optimal typhoid control initiatives could focus on this age group and on improving WASH facilities in households.

Data availability

The data that support the findings of this study are available from the chief investigator, Professor Andrew Pollard, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the corresponding author upon reasonable request and with permission of the chief investigator (andrew.pollard@paediatrics.ox.ac.uk). The code used to run the models in this study can be accessed on *Github*.

References

[1] J. D. Stanaway, R. C. Reiner, B. F. Blacker, E. M. Goldberg, et al. "The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet Infectious Diseases* 19.4 (2019), pp. 369–381.

[2] M. Antillón, J. L. Warren, F. W. Crawford, D. M. Weinberger, et al. "The burden of typhoid fever in low-and middle-income countries: a meta-regression approach". In: *PLoS neglected tropical diseases* 11.2 (2017), e0005376.

[3] N. J. Saad, V. D. Lynch, M. Antillón, C. Yang, et al. "Seasonal dynamics of typhoid and paratyphoid fever". In: *Scientific reports* 8.1 (2018), pp. 1–9.

[4] A. Akullian, E. Ng'eno, A. I. Matheson, L. Cosmas, et al. "Environmental transmission of typhoid fever in an urban slum". In: *PLoS neglected tropical diseases* 9.12 (2015), e0004212.

[5] J. S. Gauld, F. Olgemoeller, R. Nkhata, C. Li, et al. "Domestic river water use and risk of typhoid fever: results from a case-control study in Blantyre, Malawi". In: *Clinical Infectious Diseases* 70.7 (2020), pp. 1278–1284.

[6] T. C. Darton, J. E. Meiring, S. Tonks, M. A. Khan, et al. "The STRATAA study protocol: a programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using prospective population census, passive surveillance, serological studies and healthcare utilisation surveys". In: *BMJ open* 7.6 (2017), e016283.

[7] D. Thindwa, M. G. Chipeta, M. Y. Henrion, and M. A. Gordon. "Distinct climate influences on the risk of typhoid compared to invasive non-typhoid Salmonella disease in Blantyre, Malawi". In: *Scientific reports* 9.1 (2019), pp. 1–11.

[8] R. F. Breiman, L. Cosmas, H. Njuguna, A. Audi, et al. "Population-based incidence of typhoid fever in an urban informal settlement and a rural area in Kenya: implications for typhoid vaccine use in Africa". In: *PloS one* 7.1 (2012), e29119.

[9] A. Fusheini and S. K. Gyawu. "Prevalence of typhoid and paratyphoid fever in the hohoe municipality of the Volta region, Ghana: a five-year retrospective trend analysis". In: *Annals of global health* 86.1 (2020).

[10] W. H. Organization et al. "Typhoid and other invasive salmonellosis". In: *Vaccine-preventable diseases surveillance standards. Geneva, Switzerland* (2018), pp. 1–13.

[11] J. E. Meiring, M. B. Laurens, P. Patel, P. Patel, et al. "Typhoid Vaccine Acceleration Consortium Malawi: a phase III, randomized, double-blind, controlled trial of the clinical efficacy of typhoid conjugate vaccine among children in Blantyre, Malawi". In: *Clinical Infectious Diseases* 68.Supplement_2 (2019), S50–S58.

[12] J. E. Meiring, M. Shakya, F. Khanam, M. Voysey, et al. "Burden of enteric fever at three urban sites in Africa and Asia: a multicentre population-based study". In: *The Lancet Global Health* 9.12 (2021), e1688–e1696.

[13] P. Musicha, J. E. Cornick, N. Bar-Zeev, N. French, et al. "Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998–2016): a surveillance study". In: *The Lancet infectious diseases* 17.10 (2017), pp. 1042–1052.

[14] N. A. Feasey, K. Gaskell, V. Wong, C. Msefula, et al. "Rapid emergence of multidrug resistant, H58-lineage Salmonella typhi in Blantyre, Malawi". In: *PLoS neglected tropical diseases* 9.4 (2015), e0003748.

[15] V. E. Pitzer, N. A. Feasey, C. Msefula, J. Mallewa, et al. "Mathematical modeling to assess the drivers of the recent emergence of typhoid fever in Blantyre, Malawi". In: *Clinical Infectious Diseases* 61.suppl_4 (2015), S251–S258.

[16] A. J. Browne, B. H. Kashef Hamadani, E. A. Kumaran, P. Rao, et al. "Drug-resistant enteric fever worldwide, 1990 to 2018: a systematic review and meta-analysis". In: *BMC medicine* 18.1 (2020), pp. 1–22.

[17] J. S. Gauld, S. Bilima, P. J. Diggle, N. A. Feasey, et al. "Rainfall Anomalies and Typhoid Fever in Blantyre, Malawi". In: *Epidemiology & Infection* (2022), pp. 1–22.

[18] J. S. Gauld, F. Olgemoeller, E. Heinz, R. Nkhata, et al. "Spatial and genomic data to characterize endemic typhoid transmission". In: *Clinical Infectious Diseases* 74.11 (2022), pp. 1993–2000.

[19] F. B. Osei, A. Stein, and S. D. Nyadanu. "Spatial and temporal heterogeneities of district-level typhoid morbidities in Ghana: A requisite insight for informed public health response". In: *Plos one* 13.11 (2018), e0208006.

[20] K. Ismail, G. Maiga, D. Ssebuggwawo, P. Nabende, et al. "Spatio-temporal trends and distribution patterns of typhoid disease in Uganda from 2012 to 2017". In: *Geospatial health* 15.2 (2020).

[21] T. Tango. "Spatial scan statistics can be dangerous". In: *Statistical Methods in Medical Research* 30.1 (2021), pp. 75–86.

[22] M. I. Khan, R. Ochiai, S. Soofi, L. VON-SEIDLEIN, et al. "Risk factors associated with typhoid fever in children aged 2–16 years in Karachi, Pakistan". In: *Epidemiology & Infection* 140.4 (2012), pp. 665–672.

[23] S. Baker, K. E. Holt, A. C. Clements, A. Karkey, et al. "Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission". In: *Open biology* 1.2 (2011), p. 110008.

[24] A. J. Tatem. "WorldPop, open data for spatial demography". In: *Scientific data* 4.1 (2017), pp. 1–4.

[25] P. J. Diggle, I. Kaimi, and R. Abellana. "Partial-likelihood analysis of spatio-temporal point-process data". In: *Biometrics* 66.2 (2010), pp. 347–354.

[26] M. Berman and T. R. Turner. "Approximating point process likelihoods with GLIM". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41.1 (1992), pp. 31–38.

[27] B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[28] A. Baddeley, E. Rubak, and R. Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

[29] J. A. Crump, S. P. Luby, and E. D. Mintz. "The global burden of typhoid fever". In: *Bulletin of the World Health Organization* 82 (2004), pp. 346–353.

[30] M. T. Phillips, J. E. Meiring, M. Voysey, J. L. Warren, et al. "A Bayesian approach for estimating typhoid fever incidence from large-scale facility-based passive surveillance data". In: *Statistics in medicine* 40.26 (2021), pp. 5853–5870.

[31] A. A. Sattar, M. S. J. H. Chowdhury, M. A. Yusuf, S. Jesmin, et al. "Age and gender difference of typhoid Fever among paediatric patients attended at a tertiary care hospital in Bangladesh". In: *Bangladesh Journal of Infectious Diseases* 3.2 (2016), pp. 36–39.

[32] A. M. Dewan, R. Corner, M. Hashizume, and E. T. Ongee. "Typhoid fever and its association with environmental factors in the Dhaka metropolitan area of Bangladesh: a spatial and time-series approach". In: *PLoS neglected tropical diseases* 7.1 (2013), e1998.

[33] W. H. Organization et al. *A toolkit for national dengue burden estimation*. Tech. rep. World Health Organization, 2018.

[34] X. Li, H. H. Chang, Q. Cheng, P. A. Collender, et al. "A spatial hierarchical model for integrating and bias-correcting data from passive and active disease surveillance systems". In: *Spatial and Spatio-temporal Epidemiology* 35 (2020), p. 100341.

[35] R. B. Dissanayake, E. Giorgi, M. Stevenson, R. Allavena, et al. "Estimating koala density from incidental koala sightings in South-East Queensland, Australia (1997–2013), using a self-exciting spatio-temporal point process model". In: *Ecology and Evolution* 11.20 (2021), pp. 13805–13814.

Chapter 4

Paper 2: Prevalence and determinants of double and triple burden of malnutrition among mother-child pairs in Malawi: a mapping and multilevel modelling study

Jessie J. Khaki [1,2,3], Peter Macharia[1,4,5], Lenka Benova[5], Emanuele Giorgi[1], Aline Semaan[5].

[1] Lancaster Medical School, Lancaster University, Lancaster, United Kingdom.

[2] Malawi-Liverpool-Wellcome Trust Programme, Blantyre, Malawi.

[3] School of Global and Public Health, Kamuzu University of Health Sciences, Blantyre, Malawi.

[4] KEMRI Wellcome Trust, Kenya.

[5] Department of Public Health, Institute of Tropical Medicine, Belgium.

Summary

This study aimed to establish the prevalence of double burden of malnutrition (DBM) and triple burden of malnutrition (TBM) among mother-child pairs in Malawi and explore their geographical distribution and associated multilevel factors.

This was a cross-sectional study that used secondary data from the 2015-16 Malawi Demographic and Health Survey. We used a mixed effects binomial model to identify multilevel factors associated with DBM and TBM. Georeferenced covariates were used to map the predicted prevalence of DBM and TBM. The study was carried out in all of the 28 districts in Malawi and the sample comprised mother-child pairs with mothers aged 15 to 49 years and children aged below 59 months (n=4,618 pairs) for DBM and between 6 and 59 months (n=4,209 pairs) for TBM.

Approximately 5.5% [95% confidence interval (CI): 4.7%, 6.4%] of mother-child pairs had DBM and 3.1% [95% CI: 2.5%, 4.0%] had TBM. The subnational-level prevalence of DBM and TBM was highest in cities. The adjusted odds of DBM were threefold higher [Adjusted Odds Ratio, AOR: 2.8, 95% CI: 1.1, 7.3] with a higher proportion of wealthy households in a community. The adjusted odds of TBM were 60% lower [AOR: 0.4; 95% CI: 0.2, 0.8] among pairs where the women had some education compared to women with no education.

Although the prevalence of DBM and TBM is currently low in Malawi, it is more prevalent in pairs with women with no education and in relatively wealthier communities. Targeted interventions should address both maternal overnutrition and child undernutrition in cities and these demographics.

Keywords: double burden; triple burden; malnutrition; mapping; multilevel models; mother-child pairs, DHS.

4.1 Introduction

Malnutrition, including micronutrient deficiencies, excesses, or imbalances, causes serious and long-lasting adverse outcomes at both individual and community levels. Nearly half (45%) of all deaths among under-five children globally are due to nutrition-related factors, and the burden is higher in low- and middle-income countries (LMICs) [1]. Since 2000, there has been substantial progress in reducing the burden of undernutrition among under-five children [2]. The global prevalence of stunting, which is the most common type of child malnutrition, has reduced from 32.6% in 2000 to 22.2% in 2017 [2]. However, a recent World Health Organization (WHO) report cited that populations in 88% of 141 countries experienced multiple types of malnutrition, including child stunting and overweight in women [2]. The report further highlighted that 41 (29%) of the countries, 30 of which are in Africa, experience a high burden of child stunting ($\geq$ 20%) and overweight and obesity in adult women ($\geq$ 35%).

The coexistence of undernutrition (such as stunting, wasting, or micronutrient deficiency) and overnutrition (obesity or overweight) is defined as double burden of malnutrition (DBM) [3]. The WHO states that DBM can occur at the individual level (e.g., coexistence of overnutrition with mineral or vitamin deficiencies in one individual), household level (e.g., nutritional anaemia in a child and overnutrition in another member of the household), and population level (e.g., the existence of a burden of undernutrition and overnutrition in the same community such as a village, district or country) [3]. DBM among mother-child pairs is defined as the coexistence of undernutrition (wasting, stunting or underweight) in the child and overnutrition (overweight or obesity) in the mother [4]. Furthermore, a mother-child pair can also have a child with overnutrition and undernutrition in the mother. Triple burden of malnutrition (TBM) refers to the coexistence of micronutrient deficiencies and undernutrition in children and maternal overnutrition [4]. Similarly, TBM among mother-child pairs can include overnutrition in the child and the coexistence of undernutrition and micronutrient deficiencies in the mother. Childhood malnutrition is associated with multiple adverse outcomes such as delayed cognitive development and mortality [5]. Likewise, overnutrition in adults is associated with an increased risk of acquiring

non-communicable diseases such as high blood pressure and diabetes; and poor pregnancy outcomes among women [6]. DBM and TBM are, therefore, increasingly being recognized as public health threats because of the risks they pose to both the mother and child and the underlying complexity resulting in the co-existence of different types of malnutrition.

The burden of DBM and TBM varies between countries. Although the global mother-child pair prevalence of DBM and TBM is unknown, the prevalence of household-level DBM ranges between 3% and 35% across 126 LMICs, with the highest prevalence reported in Southern Africa, South America, and Asia [7, 8]. For Southern Africa, a study carried out in 2021 in 23 countries of the region estimated the prevalence of household-level DBM and TBM to be 8% and 5%, respectively [9].

Malawi has been experiencing DBM. The most recent 2015-16 Malawi Demographic and Health Survey (MDHS) reported that the prevalence of stunting (37%), underweight (12%), overweight (5%) and wasting (3%) among under-five children declined compared to the 2010 MDHS (stunting = 47%, underweight = 13%, overweight = 8%, and wasting = 4%) [10, 11]. However, the prevalence of overweight among women of reproductive age increased from 17% in 2010 to 21% in 2015-16. Further, the prevalence of anaemia among under-five children decreased from 73% in 2004 to 63% in 2010, and remained at that level in 2015-16, suggesting a stall in the reduction. Although there has been extensive research in Malawi on factors associated with various forms of malnutrition among children and overnutrition among women of reproductive age, research looking at the co-occurrence of undernutrition, micro-nutrient deficiencies among children and overnutrition among their mothers in Malawi is scarce [12–14].

There is limited research on geographical disparities in DBM and TBM among mother-child pairs in LMICs. Tarekegn et al (2022) used the Anselin Local Moran's I test to identify hotspots for DBM and TBM among mother-child pairs in Ethiopia [15]. In Kenya, Kasomo et al (2021) used a Bayesian geoadditive regression model to identify factors associated with DBM among women and to

identify areas with a high burden of DBM [16]. Both studies showed that the distribution of DBM and TBM varies across space. In addition to the spatial studies, other previous research used multilevel analyses to determine factors associated with TBM [17].

The objectives of the current study were to (1) estimate the prevalence of DBM and TBM among mother-child pairs; (2) examine the variability in prevalence of DBM and TBM geographically; and (3) identify individual, household, and community level factors associated with DBM and TBM among mother-child pairs, in Malawi.

4.2 Methods

4.2.1 Data

The most recent 2015-16 MDHS data were used for this study [10]. DHSs are nationally representative cross-sectional household surveys that are carried out in LMICs for tracking health and demographic indicators [18]. Respondents for the 2015-16 MDHS were sampled using a two-stage process which was guided by a sampling frame generated from the 2008 Malawi Population and Housing Census [10]. Details on the MDHS sampling methodology can be found in other reports [10]. In summary, the 2015-16 MDHS sampled 850 Enumeration Areas (EAs) in all the 28 districts of Malawi in the first sampling stage. In the second stage, 33 households in each rural cluster and 30 households in each urban cluster were selected. Enumerators used a global positioning system (GPS) to identify the central point of each EA and to collect coordinates (longitude and latitude). The DHS programme displaces the coordinates by up to 2 km in urban areas and up to 5 km in 99% of the rural areas, and by 10 km in the remaining 1% of EAs in rural areas for anonymization purposes [19]. Coordinate displacements ensure that points remain within the same administrative boundaries of the EA [19].

4.2.2 Study population

The population of interest in the study were living together in a household mother-child pairs where the child was less than 60 months (5 years) old at the

time of survey. We used the women's and children's recode datasets which contain information on women of reproductive age (15-49 years), their most recent birth, and their children. Anthropometry and biomarker data were collected from members of a sub-sample of the interviewed households. Height and weight measurements were taken from women aged between 15 and 49 years and children aged between 0 and 59 months in the eligible households [10]. To measure hemoglobin levels for determining anemia status, blood specimens were collected from a sub-sample of households eligible for anthropometry data collection.

The analysis sample for DBM included women of reproductive age (15-49 years) with all their under-five children residing in the same household. The analysis of TBM was restricted to a subset of women of reproductive age and their children aged between 6 and 59 months, because the DHS collected hemoglobin levels for this child age group only [10]. Pregnant women at the time of the survey and women who gave birth in the two months before the survey were excluded from the sample to avoid their pregnancy weight biasing their body mass index (BMI) [10, 20]. We also excluded women and children whose anthropometric measurements were not recorded [20]. Furthermore, we excluded children whose dates of birth were missing or unknown and also excluded children with anthropometric measurements outside of plausible ranges as defined by the Guide to DHS statistics [20]. The sample inclusion flow chart is provided in supplementary information B.

4.2.3 Outcome and independent variables definition

This study had two outcome variables: double burden of malnutrition (DBM) and triple burden malnutrition (TBM) among mother-child pairs. The operational definitions for DBM and TBM were adapted from previous studies and are presented in Table 4.1, along with definitions of malnutrition indicators such as stunting and wasting available in the MDHS [4, 15].

Table 4.1: Outcome variable definition as adapted from previous DBM and TBM studies [4, 15, 21].

Indicator	Unit of analysis	Definition
Wasting (Weight for Height Z-score - WHZ)	Children 0-59 months	1 = wasted (WHZ < -2 standard deviations -SD), 0 = not wasted
Stunting (Height/Length for Age Z-score-HAZ)	Children 0-59 months	1 = stunted (HAZ < -2 SD), 0 = not stunted
Underweight (Weight for Age Z-score - WAZ)	Children 0-59 months	1 = underweight (WAZ < -2SD), 0 = not underweight
Child undernutrition	Children 0-59 months	1 = Wasted, stunted or underweight, 0 = no undernutrition
Child overnutrition	Children 0-59 months	1 = overweight (WAZ > 2SD), 0 = not overweight
Child anemia	Children 6-59 months	1 = anemic (hemoglobin (HB) level <11.0 g/dl), 0 = not anemic
Maternal overnutrition (overweight or obesity) based on body mass index (BMI)	Women 15-49 years	1 = Overweight (BMI 25.0–29.9 weight (kg)/height(m2)) or obese (BMI 30/kg/m2), 0 = normal/underweight BMI
Maternal underweight	Women 15-49 years	1 = Mildly thin (BMI 17.0-18.4 kg/m2) or moderately or severely thin (BMI <17.0 kg/m2),

Continued on next page

Table 4.1 – continued from previous page

Indicator	Unit of analysis	Definition
		0 = normal/overweight or obese BMI
Maternal short height	Women 15-49 years	1 = Height <145cm, 0 = normal height
Maternal undernutrition	Women 15-49 years	1 = Maternal underweight or short height, 0 = normal
Maternal anemia	Women 15-49 years	1 = anemic (HB level <12.0 g/dl), 0 = not anemic
Double burden of malnutrition (DBM)	Women 15-49 years and their children 0-59 months from their most recent birth, including multiple births	1 = Mother's (maternal) overweight or obesity AND child undernutrition (wasting = 1 or stunting =1 or underweight = 1) OR Maternal undernutrition AND child overnutrition, 0 = no DBM
Triple burden of malnutrition (TBM)	Women 15-49 years and their children 6-59 months from their most recent birth, including multiple births	1 = Maternal overweight or obesity, and undernourished and anemic child (child anemia =1) OR child overweight AND maternal undernutrition and maternal anemia (maternal anemia = 1), 0 = no TBM

The study explored associations between the outcomes and independent variables at the individual, household, and community levels. An initial selection of variables was based on the WHO conceptual framework for the double burden of malnutrition (see Supplementary information B) and previous research looking at the prevalence and burden of DBM and TBM [4, 15, 21, 22]. These variables are listed in Table B.1 in Supplementary information B, along with the variable label. The community-level variables were aggregated from individual/household level to EA (i.e. cluster) level. For instance, we generated a new variable and assigned a 1 to the households in the middle, rich or richest wealth quintiles and a zero to the households in the poor/poorest quintiles. The variable capturing the percentage of households in at least the middle wealth quintile in a cluster was, therefore, computed by taking the sum of the new variable divided by the total number of households sampled in that cluster.

The following gridded raster covariates were used in the mapping analysis: precipitation, nightlights, elevation, temperature, aridity, antenatal visits during pregnancy, female literacy and percentage of children who had received all basic vaccinations. These variables have been shown to be associated with nutrition-related indicators and have been included in previous modelling and mapping studies [17, 23–28]. The variables were also selected based on the WHO conceptual framework (see Supplementary information B) of DBM and conceptual frameworks for mother-child pair DBMs from previous research [29, 30]. Details on the sources of the variables and their spatial and temporal resolution are given in Supplementary information B.

4.2.4 Statistical analyses

Outcome and independent variables were explored by tabulating frequencies for the categorical variables and computing the mean and standard deviations for the continuous variables. Bivariate analyses were carried out using a Chi-square test for categorical variables and t-test for continuous variables. To avoid including multiple independent variables that were highly correlated, we fitted a logistic regression model to the outcome variables and computed the variance inflation factors (VIF) of the independent variables. All the variables had a VIF of less than

3 and were included in the multilevel models [31, 32].

The multivariable analysis first considered a binomial mixed model (three-level multilevel model where mother-child pairs reside within a household that is in a community/EA) to investigate multilevel risk factors of TBM and DBM. Mixed models are valuable for analysing data with hierarchical or nested structures, allowing for the incorporation of both fixed and random effects [33]. The clustering variables that were used in the study were the household number and the EA which is coded as a cluster number in the DHS data. A three-level multivariable binomial mixed model was fitted to investigate the determinants of DBM and TBM at the individual, household and cluster levels where mother-child pairs (individual, Level 1) are in households (Level 2) which are clustered within an EA (community level, Level 3).

Spatial modelling

To map DBM and TBM, we adapted a methodology utilized in estimating and mapping of other health-related outcomes that are derived from multiple indicators [34, 35]. Each of the indicators contributing to DBM and TBM (child stunting, child wasting, child underweight, child overnutrition, child anaemia, maternal undernutrition, maternal short height, maternal anaemia, and maternal overnutrition) was individually considered in our approach [34]. We examined multicollinearity and selected between closely associated spatial covariates as described above [32, 36]. We then fitted binomial generalized mixed models to each of the nine indicators. Random effects extracted from the binomial mixed models were used to fit variograms, which confirmed the absence of spatial correlation in all nine indicators, as depicted in Figure B.5 and Figure B.6 in the Supplementary information. We, therefore, mapped DBM and TBM using the non-spatial binomial mixed models as done in other non-spatial mapping studies since spatial dependence is required when using geospatial methods [36–38]. We computed the predicted prevalence of each of the nine indicators and combined them to produce the predicted prevalence of DBM and TBM at pixel level (3km x 3km grid) and district level [34]. We utilized the multivariate Gaussian approximation of the maximum likelihood estimator to compute confidence intervals for the estimates as employed in previous non-spatial mapping studies [38].

We generated new level-weights according to the recent guidance by the DHS Survey methodology team [39]. All analyses adjusted for the weights and were carried out at a 5% significance level using Stata version 15 and R. We further adjusted for the stratification. Statistical details for the multilevel model and mapping analyses are available in the supplementary information.

4.3 Results

We analysed data for a total of 4,618 for DBM and 4,209 for TBM (Level 1 data) mother-child pairs. The DBM mother-child sample were from 3,661 households (Level 2) within 848 communities (Level 3). The TBM mother-child sample was from 3,442 unique households (Level 2) within 844 communities (Level 3). Among the 4,618 mother-child pairs in the DBM sample, there were 4,473 (97%) unique women, and in the TBM sample, 4,089 (97%) were unique women, meaning than less than 3% of children in the analysis pairs shared the same mother. The distribution of the DBM and TBM samples is displayed in Table 4.2. Briefly, about a fifth of the children in both samples (DBM 21%, TBM 23%) were aged between 12 months and 23 months and a slight majority were female (DBM 52%, TBM 52%). Furthermore, three-quarters of the mothers were aged between 20 and 34 years (DBM 74%, TBM 74%) and two thirds had attended up to primary school education at the time of the survey (DBM 66%, TBM 66%).

Table 4.2: Socio-demographic characteristics among mother-child pair analysis samples for DBM (n=4,618) and TBM (n=4,209) from the Malawi 2015-2016 DHS.

Variable	**Sample for DBM estimation**			**Sample for TBM estimation**		
	Total (N)	**%**	**95% CI**	**Total (N)**	**%**	**95% CI**
Age of child						
<12 months	901	19.51	(18.16,20.93)	532	12.65	(11.46,13.93)
12-23 months	989	21.41	(19.99,22.90)	979	23.27	(21.75.24.87)
24 - 35 months	907	19.65	(18.29,21.08)	900	21.38	(19.90,22.95)
36 to 47 months	945	20.47	(19.08,21.93)	936	22.24	(20.72,23.83)
48 to 59 months	876	18.97	(17.68,20.33)	861	20.46	(19.09,21.90)
Sex of child						
Male	2,234	48.38	(46.66,50.10)	2,031	48.25	(46.52,49.98)
Female	2,384	51.62	(49.90,53.34)	2,178	51.75	(50.02,53.48)
Parity						
1 or 2	1,902	41.19	(39.05,43.36)	1,695	40.28	(38.05,42.55)
3+	2,716	58.81	(56.64,60.95)	2,514	59.72	(57.45,61.95)
Age of mother in years						
15-19	318	6.88	(5.98,7.00)	247	5.88	(5.06,6.82)
20-34	3,427	74.20	(72.42,75.92)	3,134	74.45	(72.60,76.23)

Continued on next page

Table 4.2 – continued from previous page

Variable	Sample for DBM estimation			Sample for TBM estimation		
	Total (N)	**%**	**95% CI**	**Total (N)**	**%**	**95% CI**
35+	874	18.92	(17.39,20.55)	828	19.67	(18.04,21.41)
Mother's marital status						
Not in a union	722	15.63	(14.12, 17.27)	667	15.85	(14.29, 17.54)
In a union	3,896	84.37	(82.73, 85.88)	3,542	84.15	(82.46, 85.71)
Highest level of completed education of mother						
None	596	12.91	(11.53,14.43)	545	12.94	(11.53,14.50)
Primary	3,039	65.80	(63.77,67.77)	2,770	65.80	(63.75,67.78)
Secondary and higher	983	21.29	(19.56,23.12)	895	21.26	(19.50,23.14)
Mother's employment status						
Not working	1,508	32.66	(30.46, 34.93)	1,311	31.14	(29.87, 34.50)
Working	3,110	67.34	(65.07, 69.54)	2,856	67.86	(65.50, 70.13)
Mother attended < 4 ANC visits						
No	2,739	59.31	(57.48,61.11)	2,526	60.01	(58.06,61.93)
Yes	1,879	40.69	(38.89,42.52)	1,683	39.99	(38.07,41.94)
Household wealth index						
Low	1,573	34.06	(32.17,36.01)	1,422	33.79	(31.83,35.80)
Middle	1,527	33.07	(31.23,34.97)	1,388	32.98	(31.06,34.97)

Continued on next page

Table 4.2 – continued from previous page

Variable	Sample for DBM estimation			Sample for TBM estimation		
	Total (N)	**%**	**95% CI**	**Total (N)**	**%**	**95% CI**
High	1,518	32.87	(30.83,34.97)	1,399	33.23	(31.11,35.42)
Area of residence						
Urban	626	13.55	(12.36,14.84)	563	13.38	(12.09,14.79)
Rural	3,992	86.45	(85.16,87.64)	3,646	86.62	(85.21,87.91)
Community-level variables (mean & std dev)*						
Proportion of households belonging to the middle, rich or richest wealth quintiles	0.577 (0.31)	(0.557,0.575)		0.565 (0.31)	(0.556,0.574)	
Proportion of women with fewer than 4 ANC visits	0.592 (0.23)	(0.586,0.600)		0.591 (0.23)	(0.584,0.598)	

Note: DBM = Double burden of malnutrition; TBM =Triple burden of malnutrition;
N = Total sample size across outcome variable; n: Column sample size across independent variable;
Std. dev = Standard deviation; ANC = Antenatal care; % = Percentage; CI = Confidence interval.

The prevalence of DBM among mother-child pairs in Malawi was 5.5% (95% Confidence interval, CI: 4.7, 6.4) and the prevalence of TBM was 3.1% (95% CI: 2.5%, 4.0%). Figure 4.1 shows the prevalence of the components of DBM and TBM. Anaemia prevalence was 63.4% among children 6-59 months, and among child undernutrition indicators (child stunted, child underweight, and child wasted), stunting was highest at 36.8%.

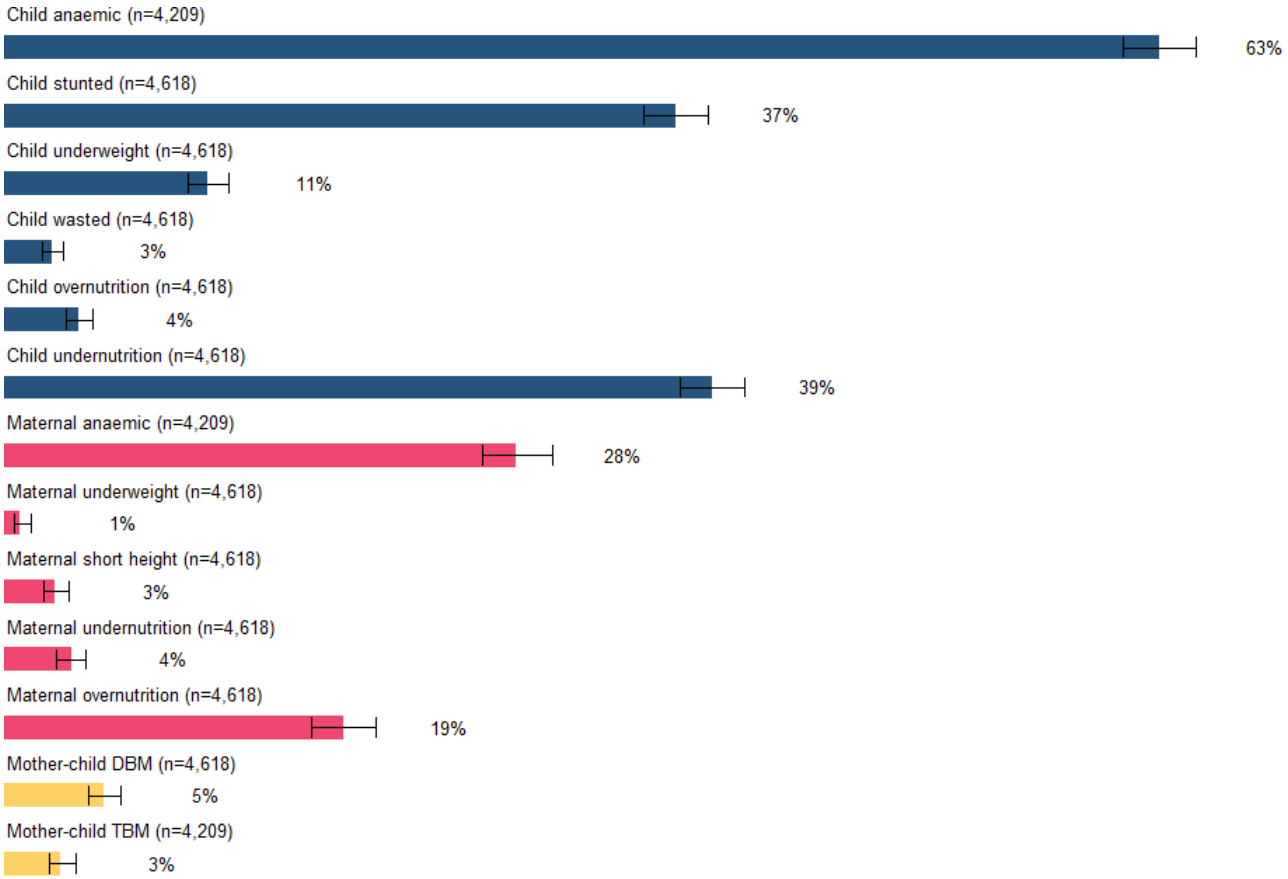

Figure 4.1: Prevalence of measures of malnutrition among mother-child pairs included in the analysis in Malawi (DBM: n=4,618, TBM: n=4,209).

4.3.1 Factors associated with DBM and TBM

The distribution of characteristics of mother-child pairs according to whether they had DBM or TBM is displayed in Table 4.3. In the bi-variate analysis, DBM was associated with older ages of the child and mother, parity, and highest level of education of the mother. On the community level, the percentage of households in at least middle wealth quintile were associated with higher DBM, whilst the proportion of women who attended fewer than 4 antenatal care (ANC) visits during

pregnancy were also associated with lower DBM. In the bi-variate analysis of TBM, only parity and maternal education were significantly associated with TBM. Mothers with some education had lower odds of mother-child TBM.

Table 4.3: Bivariate and multivariable analyses (from the multilevel logistic regression model) of the individual, household and community-level variables associated with mother-child pair DBM (n=4,618) and TBM (n=4,209), 2015-16 MDHS.

Variable	DBM Sample		DBM Model	TBM Sample		TBM Model
	n (%)	p-val	AOR (95% CI)	n (%)	p-val	AOR (95% CI)
Individual-level variables						
Age of child		**<0.001**			0.354	
<12 months	30 (11.2)		Ref	15 (10.5)		Ref
12-23 months	32 (11.7)		1.1 (0.4, 2.9)	24 (17.0)		0.8 (0.3, 2.2)
24 - 35 months	69 (25.9)		**4.3 (1.6, 11.8)**	40 (28.4)		1.7 (0.6, 4.9)
36 to 47 months	60 (22.4)		2.1 (0.8, 5.2)	28 (19.6)		0.8 (0.3, 2.4)
48 to 59 months	77 (28.8)		**4.1 (1.4, 11.7)**	35 (24.5)		1.1 (0.3, 3.8)
Sex of child		0.691			0.670	
Male	125 (46.8)		Ref	64 (45.5)		Ref
Female	143 (53.2)		0.9 (0.5, 1.8)	77 (54.5)		1.2 (0.7, 2.1)
Parity		**<0.001**			**0.010**	
1 or 2	72 (26.8)		Ref	36 (25.5)		Ref
3+	196 (73.2)		1.9 (0.9, 4.1)	105 (74.5)		2.0 (1.0, 4.2)
Age of mother in years		**<0.001**			0.101	
35+	8 (3.0)		Ref	6 (4.2)		Ref

Continued on next page

Table 4.3 – continued from previous page

Variable	DBM Sample		DBM Model	TBM Sample		TBM Model
	n (%)	**p-val**	**AOR (95% CI)**	**n (%)**	**p-val**	**AOR (95% CI)**
15-19	174 (65.1)		0.4 (0.1, 2.2)	94 (66.3)		1.1 (0.3, 5.1)
20-34	86 (31.9)		**0.4 (0.2, 0.8)**	42 (29.5)		0.8 (0.3, 2.0)
Mother's marital status		0.165			0.570	
Not in a union	32 (11.8)		Ref	19 (13.6)		Ref
In a union	236 (88.0)		0.5 (0.2, 1.2)	122 (86.4)		0.7 (0.3, 1.6)
Highest completed level of education of mother		**<0.001**			**<0.001**	
None	47 (17.4)		Ref	34 (24.4)		Ref
Primary	159 (59.4)		0.6 (0.2, 1.3)	91 (64.8)		**0.4 (0.2, 0.8)**
Secondary and higher	62 (23.2)		0.9 (0.3, 2.8)	15 (10.8)		**0.2 (0.1, 0.6)**
Mother's employment status		0.988			0.967	
Not working	87 (32.6)		Ref	46 (32.4)		Ref
Working	181 (67.4)		0.8 (0.4, 1.9)	95 (67.6)		0.8 (0.4, 1.6)
Mother attended at least 4 ANC visits		0.876			0.393	
Yes	160 (59.9)		Ref	79 (55.9)		Ref
No	108 (40.1)		1.1 (0.6, 2.0)	62 (44.1)		0.8 (0.4, 1.4)
Household-level variables						
Household wealth index		0.594			0.388	

Continued on next page

Table 4.3 – continued from previous page

Variable	DBM Sample		DBM Model	TBM Sample		TBM Model
	n (%)	**p-val**	**AOR (95% CI)**	**n (%)**	**p-val**	**AOR (95% CI)**
Low	81 (30.1)		Ref	56 (39.8)		Ref
Middle	91 (34.1)		0.9 (0.4, 2.0)	48 (34.0)		0.7 (0.3, 1.5)
High	96 (35.8)		0.7 (0.2, 1.9)	37 (26.2)		0.6 (0.2, 1.7)
Area of residence		0.414			0.147	
Urban	43 (15.9)		Ref	12 (8.3)		Ref
Rural	225 (84.1)		1.2 (0.6, 2.6)	129 (91.7)		1.5 (0.5, 4.4)
Community-level variables						
Proportion of households belonging to the middle, rich or richest wealth quintiles	0.63 (0.30)	**<0.001**	**2.8 (1.1, 7.3)**	0.58 (0.29)	0.499	1.5 (0.4, 4.8)
Proportion of women with fewer than 4 ANC visits	0.56 (0.24)	**0.032**	0.5 (0.2, 1.6)	0.58 (0.24)	0.570	0.4 (0.1, 2.1)
Random effects						
Cluster variance			33.9 (22.7, 50.5)			1.4 (0.5, 4.3)
Household variance			2.1 (0.8, 5.6)			4.1 (1.2, 13.5)

Note: DBM = Double burden of malnutrition; TBM =Triple burden of malnutrition; n: Column sample size across independent variable (DBM = yes, TBM = yes); p-val = Chi-square *P-value* for categorical variables and t-test *P-value* for continuous (cluster-level) variables; AOR = Adjusted Odds Ratio; CI = Confidence Interval; Ref = Reference category, AOR = 1; Std dev = Standard deviation; ANC = Antenatal care; % = Percentage; Bold = statistical significance in the multilevel modelling (*P-value* < 0.05).

In the adjusted multilevel logistic regression, an increase in the child's age was associated with an increase in adjusted odds of DBM. Being a child between age 24 and 35 months is associated with 4-times higher odds of DBM [adjusted odds ratio, AOR: 4.3; 95% CI: 1.6, 11.8] compared to children aged below 11 months. Similarly, being a child between age 48 and 59 months is associated with 4 times higher odds of DBM [AOR: 4.1, 95% CI: 1.4, 11.7] compared to children aged below 11 months. Conversely, the odds of having a mother-child pair with DBM was 60% lower [AOR: 0.4; 95% CI: 0.2, 0.8] among mother-child pairs with women aged between 20 and 34 years compared to women aged 35 years and above. In the multilevel logistic regression for TBM among mother-child pairs, mother's level of educational attainment was the only factor statistically significantly associated with the outcome (*p-value* <0.01). The odds of TBM were 60% lower [AOR: 0.4; 95% CI: 0.2, 0.8] among mother-child pairs where the women had primary education and 80% lower [AOR: 0.2; 95% CI: 0.1, 0.6] among mother-child pairs where the women had secondary or tertiary education, compared to mother-child pairs where the mother had no education.

Sex of the child, birth order, mother's marital status, mother's employment status, number of attended ANC visits, household wealth quintile and community level of ANC attendance were not associated with DBM or TBM (p>0.05). DBM among mother-child pairs was significantly associated with the household wealth in the community. A one-unit increase in the percentage of households in the middle wealth quintile or wealthier in the community was associated with a three-fold increase in the adjusted odds of DBM [AOR: 2.8 95% CI: 1.1, 7.3]. No statistically significant community-level effects were seen in the TBM model.

In the spatial analysis, an increase in the nightlights value in an area, which is a proxy for the wealth index, was associated with decreased odds of child-level outcomes such as child stunting, anaemia and underweight. On the other hand, an increase in nightlight value in an area was associated with increased odds of maternal overnutrition (overweight and obese). Likewise, an increase in the proportion of literate women was associated with reduced odds of childhood stunting, wasting, underweight and anaemia. With respect to maternal-level outcomes, an increase in the proportion of literate women was associated with increased odds of maternal

overnutrition. Climatic variables such as precipitation, temperature and aridity index were not significantly associated with any of the child-level and maternal-level outcomes.

4.3.2 Geographic distribution of DBM and TBM

The predicted prevalence of DBM was heterogeneous, ranging from 1.2% to 8.2% across the pixels in Malawi (Figure 4.2). The predicted prevalence of TBM ranged from 0.9% to 3.5% at the pixel level. The highest prevalence of mother-child pairs with DBM and TBM was estimated in cities (Figure 4.3). All the four cities in Malawi (Blantyre, Lilongwe, Mzuzu and Zomba) had a DBM prevalence of greater than 5.1% and a TBM predicted prevalence greater than 2.4%. Maps illustrating the uncertainties in the predicted prevalence of DBM and TBM are included in Supplementary information B.

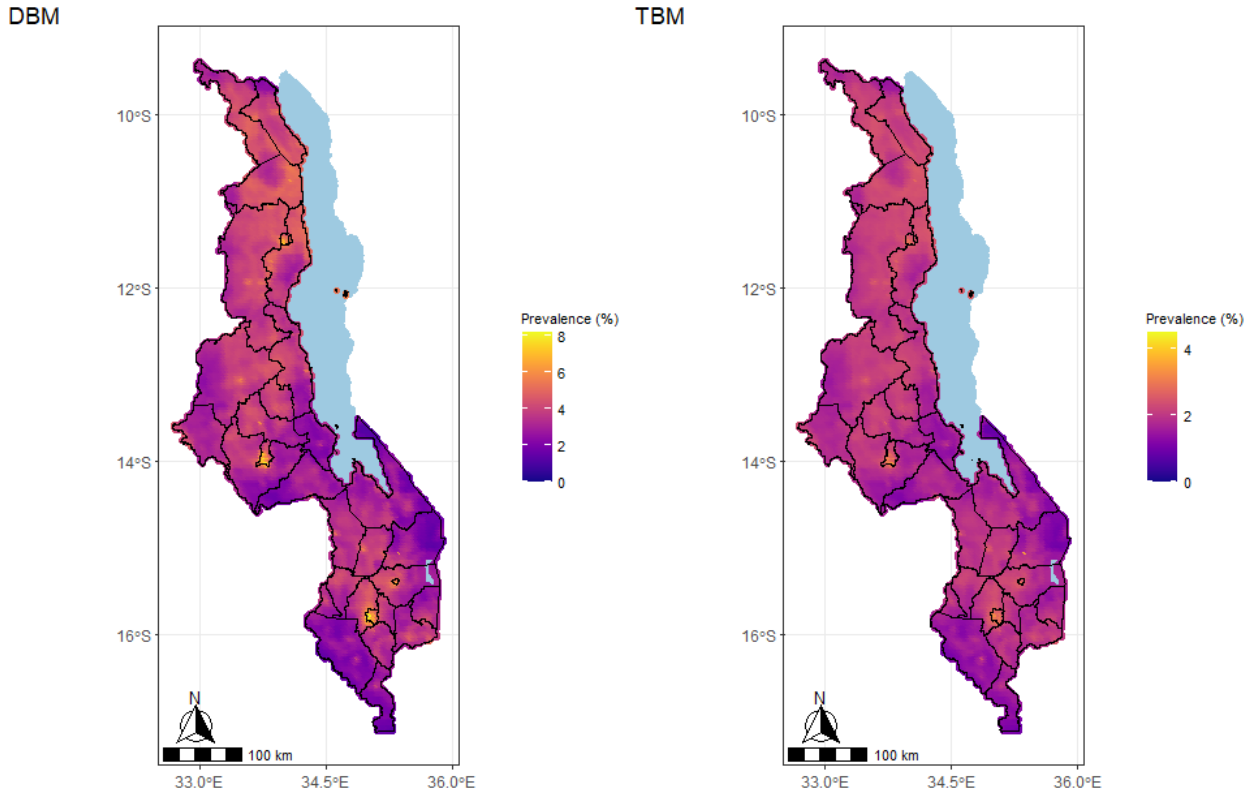

Figure 4.2: Predicted prevalence of the double burden of malnutrition (DBM) and triple burden of malnutrition (TBM) among mother-child pairs in Malawi.

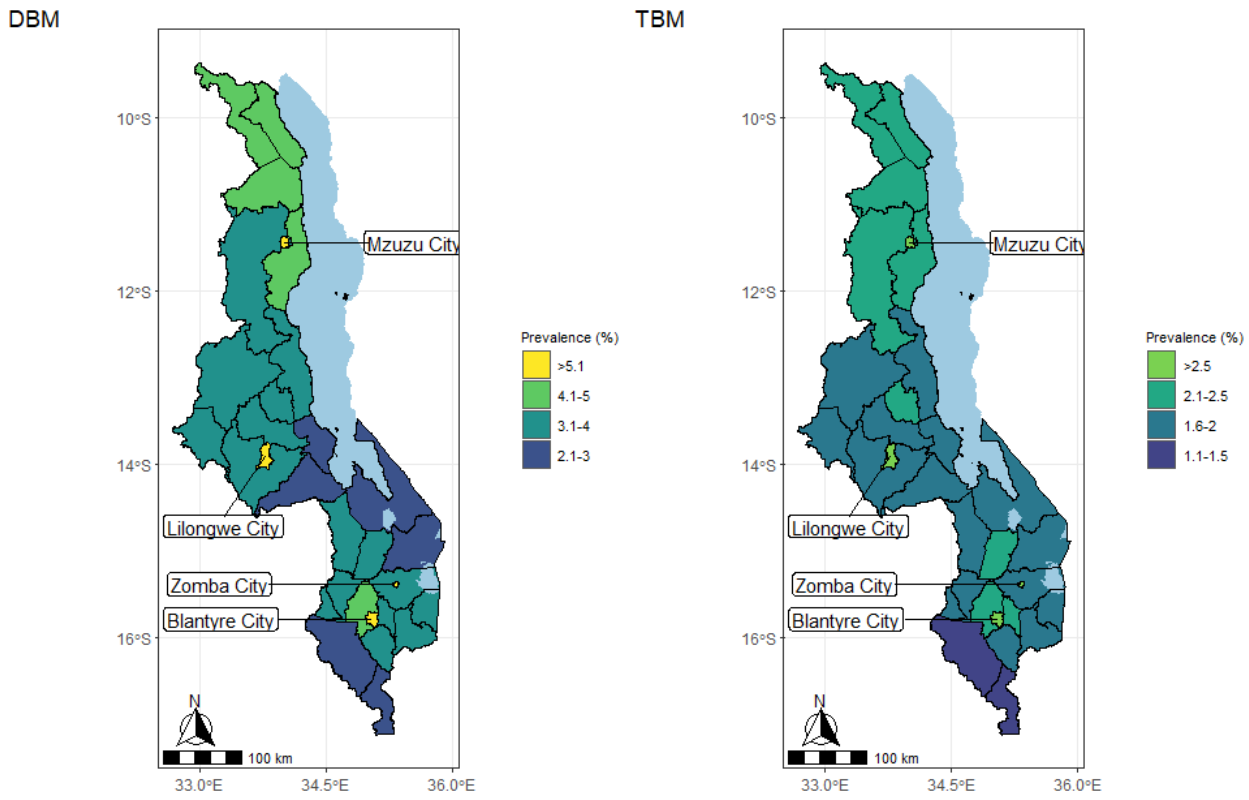

Figure 4.3: District-level predicted prevalence of the double burden of malnutrition (DBM) and triple burden of malnutrition (TBM) among mother-child pairs in Malawi.

4.4 Discussion and Conclusion

4.4.1 Discussion

This study examined the prevalence of and spatial variation in DBM and TBM, as well as individual, household and community-level correlates of DBM and TBM among mother-child pairs in Malawi. Our study contributes to the current literature by investigating the multilevel factors associated with DBM and TBM and generating high spatial resolution maps of DBM and TBM in Malawi. The prevalence of DBM and TBM among mother-child pairs were 5.5% and 3.1%, respectively. These prevalence values are higher than the reported DBM and TBM prevalence in Ethiopia (3.1% and 1.6%, respectively), and similar to the prevalence reported in Nepal (6.6% and 7.0%) and India (TBM 5%), respectively [4, 15, 21]. Although the overlap between child undernutrition and maternal overweight was small, two-fifths of children were malnourished (39%) and 19% of women were overweight or obese. While there was absence of spatial correlation in our data,

our findings suggest that despite the relatively low prevalence, the geographic distribution of mother-child DBM and TBM varied greatly within Malawi, thereby suggesting that environmental conditions conducive for DBM and TBM exist throughout Malawi, particularly in cities. We found a strong positive association between the increased age of a child and the odds of having mother-child pair DBM and TBM. This result is consistent with results from other studies where children older than 24 months had higher odds of being in a mother-child pair with DBM and TBM compared to younger ones [4, 9, 15, 21, 23, 40]. Regarding age, our results also revealed that mother-child pairs where the mother was aged between 20-34 had reduced odds of DBM compared to women aged at least 35 years. This finding is consistent with previous studies where the odds of household-level DBM and TBM was higher among older women than younger women due to the older women being overweight and obese [9, 41]. Some studies have postulated that sedentary lifestyles and the increased likelihood of overnutrition (overweight and obesity) among older women might be some of the leading drivers of DBM and TBM [4, 9, 15].

Educational attainment was another predictor of TBM in our study. Mothers with at least primary and secondary education had reduced odds of TBM. This result further corresponds with the spatial modelling where an increase in the proportion of literate women was associated with reduced burden of childhood malnutrition (underweight, stunting, and wasting). These results align with previous evidence that women with higher education have reduced odds of household-level DBM and TBM as higher education attainment is associated with reduced risk of childhood malnutrition [9, 23].

Our results show that an increase in the proportion of wealthy households at community level increased the odds of DBM, thereby suggesting a higher burden in relatively wealthier areas [8, 21]. The spatial analysis further decomposes this finding by showing that an increase in the nightlights value (a proxy for level of wealth) is associated with reduced odds of child-level outcomes and increased odds of maternal overnutrition. No significant community-level effects were seen in the TBM model. Previous work has postulated that higher income levels may enhance the ability to buy food, thereby shaping dietary habits and preferences [41].

Furthermore, previous studies have hypothesized that inadequate physical activity, sedentary lifestyles and westernized dietary behaviors adopted by wealthier households and communities drive adult overnutrition [8, 21, 41, 42]. Previous reports have shown that these behaviors tend to be more common in urban settings, which could be reflected in our results [43]. Further research is needed in Malawi to identify more community-level determinants of DBM and TBM.

The absence of spatial correlation in our data suggests that the spatial correlation in our DBM and TBM samples can be explained by the environmental covariates that were used in the models. This notwithstanding, the maps generated in this study suggest that DBM and TBM are not homogeneously distributed in Malawi. These results underscore the potential existence of spatial drivers influencing the observed patterns of DBM and TBM, which should be further investigated. In studies where the data exhibits spatial correlation, the mixed effects models used in this study could be further extended to explicitly incorporate the spatial correlation using geostatistical methods [36, 37].

4.4.2 Limitations and strengths of the study

The main limitation of this study is that the DHS was a cross-sectional study, therefore, no causal associations could be inferred from the study. Secondly, our analysis included categorizing continuous variables such as wasting and stunting which have been shown to lead to some loss of information on the variable [44]. An additional constraint pertains to the use of outdated DHS data. Nevertheless, this research remains valuable as it demonstrates the application of mixed effects models for estimating factors at multiple levels and mapping health outcomes like DBM and TBM. Another limitation with DHS data is the unavailability of most of the variables in the WHO DBM conceptual framework. Furthermore, the DHS data lack information on dietary behaviors which are immediately related to nutritional outcomes. This shortcoming creates the need to have more nutrition-related datasets in Malawi to better understand the determinants of nutrition-related issues such DBM and TBM in the country. Another important limitation is the lack of micronutrient data. Thus, we used anaemia status as a proxy for estimating TBM prevalence. We, therefore, recommend collecting nationally representative data on micronutrient deficiencies, and not necessarily

through DHS.

Given that common household-level risk factors for child undernutrition are often protective for maternal overweight and vice-versa, most of the risk factors identified for DBM and TBM in this study were driven either by strong positive associations with child undernutrition (and weaker or null negative associations with maternal overnutrition) or by strong positive associations with maternal overnutrition (and weaker or null negative associations with child undernutrition). These limitations might explain why the prevalence of DBM and TBM appears possibly lower than what would be observed by chance. These methodological challenges are, nevertheless, common when studying DBM and TBM.

Despite these limitations, to our knowledge, this study is the first to establish the prevalence of mother-child pair DBM and TBM in Malawi. Our study also presents an important finding that the co-existence of undernutrition in children and overnutrition in mothers is associated with both individual and community-level factors. Furthermore, our study is the first to illustrate the geographical disparities in the distribution of mother-child DBM and TBM in Malawi. It shows that the burden of DBM and TBM is higher in cities than other areas. Future studies could further extend this work by building a spatio-temporal model to assess whether the burden of DBM and TBM has also been higher in cities than in other areas over time. This is because a recent study has shown a shifting spatio-temporal trend in DBM from metropolitan areas to other regions in Guatemala [45]. The findings from the spatio-temporal modelling could help the government and implementing partners anticipate which districts/areas in Malawi might get an increasing burden of DBM and TBM in the future. The study results could also inform interventions about specific areas where attention is needed to target DBM and TBM and control the burden of these conditions. The findings from this study could also be used to inform a new hypothesis of shifting spatial DBM and TBM trends in other countries.

4.4.3 Conclusion

This study highlighted that the mother-child pair prevalence of DBM and TBM in Malawi was relatively low, with the highest burden in large cities. It further

showed that the increased age of the child and mother also increased the odds of a mother-child pair having DBM. Additionally, individual-level maternal educational attainment was shown to have a protective effect against TBM. Our study also highlighted that there are community-level determinants of DBM such as household wealth that are associated with increased odds of DBM. These results emphasize the need to not neglect wealthier communities in disseminating and implementing adult-related nutrition-related interventions in Malawi. These results could be explicitly adopted and translated in national documents such as the "Malawi National Multi-Sector Nutrition Policy 2018-2022" which acknowledges the existence of both undernutrition and overnutrition, but does not provide any recommendations on addressing their co-existence such as DBM and TBM [46].

Through the revised Malawi National Multi-Sector Nutrition Policy, the government of Malawi can tackle DBM and TBM among mother-child pairs in Malawi by implementing comprehensive, integrated nutrition programs that simultaneously address undernutrition in children and overnutrition in mothers. The government can also promote a multisectoral response to dealing with mother-child DBM and TBM in Malawi. A multisectoral approach can ensure that policies across multiple sectors, such as health and agriculture, work together to ensure access to supplements and nutritious foods, address food insecurity, and promote healthy habits among women. These interventions could, as a priority, begin to focus on the Malawian cities where double and triple burden of malnutrition among mother-child pairs is the highest.

References

[1] W. H. Organization et al. *Double-duty actions for nutrition: policy brief*. Tech. rep. World Health Organization, 2017.

[2] V. Mannar, R. Micha, L. Allemandi, A. Afshin, et al. *2020 global nutrition report: action on equity to end malnutrition*. Tech. rep. 89023, 2020.

[3] W. H. Organization et al. *The double burden of malnutrition: policy brief*. Tech. rep. World Health Organization, 2016.

[4] D. R. Sunuwar, D. R. Singh, and P. M. S. Pradhan. "Prevalence and factors associated with double and triple burden of malnutrition among mothers and children in Nepal: evidence from 2016 Nepal demographic and health survey". In: *BMC Public Health* 20 (2020), pp. 1–11.

[5] T. Vassilakou. *Childhood malnutrition: time for action*. 2021.

[6] C. E. Coimbra, F. G. Tavares, A. A. Ferreira, J. R. Welch, et al. "Socioeconomic determinants of excess weight and obesity among Indigenous women: findings from the First National Survey of Indigenous People's Health and Nutrition in Brazil". In: *Public Health Nutrition* 24.7 (2021), pp. 1941–1951.

[7] P. Seferidi, T. Hone, A. C. Duran, A. Bernabe-Ortiz, et al. "Global inequalities in the double burden of malnutrition and associations with globalisation: a multilevel analysis of Demographic and Health Surveys from 55 low-income and middle-income countries, 1992–2018". In: *The Lancet Global Health* 10.4 (2022), e482–e490.

[8] B. M. Popkin, C. Corvalan, and L. M. Grummer-Strawn. "Dynamics of the double burden of malnutrition and the changing nutrition reality". In: *The Lancet* 395.10217 (2020), pp. 65–74.

[9] A. K. Christian and F. A. Dake. "Profiling household double and triple burden of malnutrition in sub-Saharan Africa: prevalence and influencing household factors". In: *Public Health Nutrition* 25.6 (2022), pp. 1563–1576.

[10] Malawi National Statistical Office (NSO). *The 2015-16 Malawi Demographic and Health Survey (MDHS)*. https://dhsprogram.com/pubs/pdf/FR319/FR319.pdf. 2016.

[11] Malawi National Statistical Office (NSO). *The 2010 Malawi Demographic and Health Survey (MDHS)*. dhsprogram.com/pubs/pdf/fr247/fr247.pdf. 2011.

[12] P. A. M. Ntenda and J. F. Kazambwe. "A multilevel analysis of overweight and obesity among non-pregnant women of reproductive age in Malawi: evidence from the 2015–16 Malawi Demographic and Health Survey". In: *International health* 11.6 (2019), pp. 496–506.

[13] A. Ngwira. "Shared geographic spatial risk of childhood undernutrition in Malawi: An application of joint spatial component model". In: *Public Health in Practice* 3 (2022), p. 100224.

[14] K. Machira and T. Chirwa. "Dietary consumption and its effect on nutrition outcome among under-five children in rural Malawi". In: *PLoS One* 15.9 (2020), e0237139.

[15] B. T. Tarekegn, N. T. Assimamaw, K. A. Atalell, S. F. Kassa, et al. "Prevalence and associated factors of double and triple burden of malnutrition among child-mother pairs in Ethiopia: Spatial and survey regression analysis". In: *BMC nutrition* 8.1 (2022), p. 34.

[16] J. M. Kasomo and E. Gayawan. "Spatial location, temperature and rainfall diversity affect the double burden of malnutrition among women in Kenya". In: *SSM-Population Health* 16 (2021), p. 100939.

[17] H. Andriani, E. Friska, M. Arsyi, A. E. Sutrisno, et al. "A multilevel analysis of the triple burden of malnutrition in Indonesia: trends and determinants from repeated cross-sectional surveys". In: *BMC Public Health* 23.1 (2023), p. 1836.

[18] The DHS Program. *The DHS Program*. https://www.dhsprogram.com/. 2022.

[19] C. R. Burgert, J. Colston, T. Roy, and B. Zachary. *Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys*. Icf International, 2013.

[20] T. Croft, A. Marshall, C. K. Allen, F. Arnold, et al. "Guide to DHS statistics: DHS-7 (version 2)". In: *Rockville, MD: ICF* (2020).

[21] P. Kumar, S. Chauhan, R. Patel, S. Srivastava, et al. "Prevalence and factors associated with triple burden of malnutrition among mother-child pairs in India: a study based on National Family Health Survey 2015–16". In: *BMC Public Health* 21 (2021), pp. 1–12.

[22] O. C. Kurian and S. Suri. *Weighed down by the gains: India's twin double burdens of malnutrition and disease*. Vol. 193. Observer Research Foundation, 2019.

[23] B. O. Ahinkorah, I. Amadu, A.-A. Seidu, J. Okyere, et al. "Prevalence and factors associated with the triple burden of malnutrition among mother-child pairs in Sub-Saharan Africa". In: *Nutrients* 13.6 (2021), p. 2050.

[24] F. A. Kanu, M. E. D. Jefferds, A. M. Williams, O. Y. Addo, et al. "Association between hemoglobin and elevation among school-aged children: a verification of proposed adjustments". In: *The American Journal of Clinical Nutrition* 118.1 (2023), pp. 114–120.

[25] L. S. Tusting, J. Bradley, S. Bhatt, H. S. Gibson, et al. "Environmental temperature and growth faltering in African children: a cross-sectional study". In: *The Lancet Planetary Health* 4.3 (2020), e116–e123.

[26] M. W. Cooper, M. E. Brown, S. Hochrainer-Stigler, G. Pflug, et al. "Mapping the effects of drought on child stunting". In: *Proceedings of the National Academy of Sciences* 116.35 (2019), pp. 17219–17224.

[27] P. M. Amegbor, Z. Zhang, R. Dalgaard, and C. E. Sabel. "Multilevel and spatial analyses of childhood malnutrition in Uganda: examining individual and contextual factors". In: *Scientific reports* 10.1 (2020), p. 20019.

[28] J. R. Khan, M. M. Islam, A. S. M. Faisal, H. Islam, et al. "Quantification of urbanization using night-time light intensity in relation to women's overnutrition in Bangladesh". In: *Journal of Urban Health* 100.3 (2023), pp. 562–571.

[29] W. H. Organization et al. *Strategic action plan to reduce the double burden of malnutirtion in the south-east Asia region 2016–2025*. 2016.

[30] P. K. Masibo, F. Humwa, and T. N. Macharia. "The double burden of overnutrition and undernutrition in mother- child dyads in Kenya: demographic and health survey data, 2014". In: *Journal of nutritional science* 9 (2020), e5.

[31] Advanced Research Computing: Statistical Methods and Data Analytics. *Regression with STATA Chapter 2: Regression diagnostics. Checking for multicollinearity*.
https://stats.oarc.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/. 2022.

[32] O. Wariri, C. E. Utazi, U. Okomo, C. J. E. Metcalf, et al. "Mapping the timeliness of routine childhood vaccination in the Gambia: a spatial modelling study". In: *Vaccine* 41.39 (2023), pp. 5696–5705.

[33] D. Bates, M. Mächler, B. Bolker, and S. Walker. "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823* (2014).

[34] N. de Silva and A. Hall. "Using the prevalence of individual species of intestinal nematode worms to estimate the combined prevalence of any species". In: *PLoS Negl Trop Dis* 4.4 (2010), e655.

[35] H. O. Mogaji, O. O. Johnson, A. B. Adigun, O. N. Adekunle, et al. "Estimating the population at risk with soil transmitted helminthiasis and annual drug requirements for preventive chemotherapy in Ogun State, Nigeria". In: *Scientific Reports* 12.1 (2022), p. 2027.

[36] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

[37] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

[38] K. Deribe, A. Mbituyumuremyi, J. Cano, M. J. Bosco, et al. "Geographical distribution and prevalence of podoconiosis in Rwanda: a cross-sectional country-wide survey". In: *The Lancet Global Health* 7.5 (2019), e671–e680.

[39] M. Elkasabi, R. Ren, and T. W. Pullum. *Multilevel modeling using DHS surveys: a framework to approximate level-weights*. Tech. rep. ICF, 20200.

[40] S. Das, S. M. Fahim, M. S. Islam, T. Biswas, et al. "Prevalence and sociodemographic determinants of household-level double burden of malnutrition in Bangladesh". In: *Public health nutrition* 22.8 (2019), pp. 1425–1432.

[41] D. Chilot, D. G. Belay, M. W. Merid, A. A. Kibret, et al. "Triple burden of malnutrition among mother–child pairs in low-income and middle-income countries: a cross-sectional study". In: *BMJ open* 13.5 (2023), e070978.

[42] R. A. Mahumud, B. W. Sahle, E. Owusu-Addo, W. Chen, et al. "Association of dietary intake, physical activity, and sedentary behaviours with overweight and obesity among 282,213 adolescents in 89 low and middle income to high-income countries". In: *International journal of obesity* 45.11 (2021), pp. 2404–2418.

[43] J. Crush, B. Frayne, and M. McLachlan. "Rapid urbanization and the nutrition transition in Southern Africa". In: (2011).

[44] I. Kyomuhangi and E. Giorgi. "A threshold-free approach with age-dependency for estimating malaria seroprevalence". In: *Malaria Journal* 21 (2022), pp. 1–12.

[45] D. Sagastume, J. L. Peñalvo, M. Ramírez-Zea, K. Polman, et al. "Dynamics of the double burden of malnutrition in Guatemala: a secondary data analysis of the demographic and health surveys from 1998–2015". In: *Public Health* 229 (2024), pp. 135–143.

[46] Ministry of Health, Malawi. *The Malawi National Multi-Sector Nutrition Policy 2018–2022*. Accessed July 30, 2023. 2018. URL: https://leap.unep.org/countries/mw/national-legislation/national-multi-sector-nutrition-policy-2018-2022.

Chapter 5

Paper 3: Using ESPEN Data for Evidence-Based Control of Neglected Tropical Diseases in sub-Saharan Africa: a Comprehensive Model-based Geostatistical Analysis of Soil-Transmitted Helminths

Jessie J. Khaki [1,2,3], Mark Minnery[4], Emanuele Giorgi[1].

[1] Lancaster Medical School, Lancaster University, Lancaster, United Kingdom.

[2] Malawi-Liverpool-Wellcome Trust Programme, Blantyre, Malawi.

[3] School of Global and Public Health, Kamuzu University of Health Sciences, Blantyre, Malawi.

[4] Evidence Action, Washington DC, United States.

Summary

The Expanded Special Project for the Elimination of Neglected Tropical Diseases (ESPEN) was launched in 2019 by the World Health Organization and African nations to combat Neglected Tropical Diseases (NTDs), including Soil-transmitted helminths (STH), which affect over 1.5 billion people globally. In this study, we present a comprehensive geostatistical analysis of publicly available STH survey data from ESPEN to delineate inter-country disparities in STH prevalence and its environmental drivers while highlighting the strengths and limitations that arise from the use of the ESPEN data. To achieve this, we also propose the use of calibration validation methods to assess the suitability of geostatistical models for disease mapping at the national scale.

We analysed the most recent survey data with at least 50 geo-referenced observations, and modelled each STH species data (hookworm, roundworm, whipworm) separately. Binomial geostatistical models were developed for each country, exploring associations between STH and environmental covariates, and were validated using the non-randomized probability integral transform. We produced pixel-, subnational-, and country-level prevalence maps for successfully calibrated countries. All the results were made publicly available through an R Shiny application.

Among 35 countries with STH data that met our inclusion criteria, the reported data years ranged from 2004 to 2018. Models from 25 countries were found to be well-calibrated. Spatial patterns exhibited significant variation in STH species distribution and heterogeneity in spatial correlation scale (1.14 km to 3,027.44 km) and residual spatial variation variance across countries. This study highlights the utility of ESPEN data in assessing spatial variations in STH prevalence across countries using model-based geostatistics. Despite the challenges posed by data sparsity which limit the application of geostatistical models, the insights gained remain crucial for directing focused interventions and shaping future STH assessment strategies within national control programs.

Keywords: ESPEN; model-based geostatostics, neglected tropical diseases; STH.

5.1 Introduction

Soil-transmitted Helminthiases (STH) are the most common type of Neglected Tropical Diseases (NTDs) and are caused by parasitic worms, including whipworms (*Trichuris trichiura*), hookworms (*Necator americanus* and *Ancylostoma duodenale*), and roundworms (*Ascaris lumbricoides*) [1, 2]. Approximately 24% (1.5 billion) of the global population experiences annual infections of STH, with high prevalences among children and women of reproductive age, who are at the highest risk for morbidity associated with STH [1–3]. Populations that mostly suffer from STH infections are found in China, sub-Saharan Africa, East Asia, and the Americas [2, 4]. In sub-Saharan Africa, STH affect more than 11% of the population [3]. However, the STH burden greatly varies both between and within each country of the African continent [3, 4]. Although the STH mortality rate is low, STH are associated with both lower health outcomes (such as anemia and malnutrition) and poor cognitive performance [5–7].

One of the interventions for controlling the transmission of STH is mass drug administration (MDA), otherwise known as preventive chemotherapy (PC). The PC drugs are primarily given to preschool and school-age children and pregnant women to contribute to reducing STH-related morbidities. The frequency of the MDA programs is usually determined according to prevalence classes defined by the WHO, namely <2%, 2%-10%, 10%-20%, 20%-50% and >50% [1, 8, 9]. Understanding the level of burden of STH is thus crucial to assist the efficient allocation of drugs.

The Expanded Special Project for the Elimination of Neglected Tropical Diseases (ESPEN) was established in 2016 as a collaborative effort between the World Health Organization (WHO) African region office, African NTD endemic countries and other NTDs partners [10]. The ESPEN was instituted to help mobilize financial, political, and technical resources. ESPEN aims to contribute to mitigating the effects of the 5 most prevalent NTDs in Africa which, in addition to STH, are trachoma, lymphatic filariasis, schistosomiasis, and onchocerciasis. The ESPEN electronic data portal contains publicly available geo-located sub-national prevalence data on the aforementioned high-burden NTDs, as well as Loiasis. The

ESPEN portal also provides both spatial and time-referenced information for some countries. Historical applications of ESPEN data have involved the application of geostatistical mapping of diseases such as schistosomiasis, onchocerciasis, and STH at both country and continent (Africa) levels to inform survey designs and strategies for preventive therapy [3, 11–14].

Model-based geostatistics (MBG) has become an established methodology for prevalence mapping and for better understanding the spatial distribution of disease risk [15–17], thus providing valuable insights for guiding interventions, survey designs, and resource allocations [18–21]. MBG methods for global disease mapping has been instrumental in studying disease distribution across Africa; see, for example, the extensive application of MBG from the Institute of Health Metrics (IHME) in the mapping of HIV/AIDS, onchocerciasis, lymphatic filariasis, maternal and child health, and other health-related indicators [12, 22–27]. Several studies have utilized geostatistical methods to map STH and inform interventions by fitting either a single continent-wide model or have limited their analysis to a single country model [3, 28–30].

The view adopted in this study is that developing a single model for the entire African continent might prove unsuitable, given the diverse climatic and geopolitical landscapes across countries which could be excessively complex to fully capture in a single model using spatially sparse survey data. To address the disparities across countries in relation to STH risk, the adoption of a single Africa-wide model needs to carefully consider two fundamental aspects: the extensive use of spatial risk factors that can best capture the environmental and socioeconomic variation across the continent; and the use of complex covariance structure that accounts for non-stationary residual effects. Prior analyses of STH data incorporated a diverse set of covariates, including socio-economic indicators, (e.g. nightlights and gross domestic product), climatic variables (e.g. precipitation and temperature), and environmental variables (e.g. soil components and elevation) [3, 14, 28, 30]. Of the studies that provided details on the type of covariance function used, most have adopted stationary Matérn and exponential correlation functions [14, 28, 29, 31, 32]. Similarly, in the studies carried out by IHME on mapping other health outcomes at the continent level, a stationary

Matérn function was adopted and approximated using stochastic partial differential equations [12, 23, 26, 27]. The adoption of a stationary Matérn becomes more justifiable if the study area is relatively small and/or the covariates have allowed us to account for most of the non-stationary effects from the variation of the outcome. In this study, we pursue a simpler modelling approach to global mapping that aims at formulating context-specific geostatistical models tailored to individual countries, thereby enhancing our understanding of soil-transmitted helminths (STH) dynamics and their differences across countries. In contrast to the use of a single African-wide model, we show that this approach allows us to account for the spatially heterogeneous effects of spatial covariates as well as to better understand the differences in the predictive performance of MBG methods across the continent.

Most of the MBG mapping for STH have adopted cross-validation methods to assess the performance of the fitted geostatistical models [3, 14, 29–33]. In these studies, the focus was primarily on quantifying the accuracy and precision of point predictions through receiver operating curves, root mean square error summaries and mean absolute error [3, 14, 29–34]. One of the issues inherent to these cross-validation approaches is that they treat the observed fraction of positive cases as the true disease prevalence against which the model predictions are assessed [35, 36]. This assumption is especially problematic in low-prevalence settings, where the observed fraction is often zero, making it a poor proxy for the true prevalence [35, 37]. Furthermore, commonly used metrics such as mean square error (MSE) focus solely on the accuracy of point estimates, failing to account for the uncertainty in predictions. In geostatistical modeling, uncertainty quantification is crucial, as it reflects the variability and reliability of predictions across the study area, which point-based metrics like the MSE cannot capture. In this study, we use an alternative approach that uses the non-random probability integral transform (nrPIT) method originally proposed to calibrate count data models [36]. We show that one of the main advantages of the nrPIT is that it enables us to evaluate the overall consistency between the data and the predictive distribution of prevalence which is essential to establish the reliability of the predictive inferences derived from geostatistical models [35].

The majority of prior studies on the mapping of STH prevalence did not attempt to classify sub-national units according to the WHO STH prevalence classes [28, 31–34], except for Sartorius et al. [3] where a single threshold of 20% prevalence was used for the classification. In this study, we show how geostatistical models can be used to classify sub-national units based on the WHO STH prevalence classes (<2%, 2%-10%, 10%-20%, 20%-50%, and >50%) that are used to inform the frequency of MDA and other interventions.

In summary, the specific objectives of this paper are as follows:

- to demonstrate how to make the best use of publicly available STH survey data from the ESPEN portal;
- to highlight between countries differences in terms of the importance of environmental risk factors and spatial correlation structure in STH prevalence;
- to highlight the limitations of global mapping when using spatially sparse data, through the non-randomized integral probability transform (nrPIT).

5.2 Materials and methods

5.2.1 Analysis Outline

The workflow of the geostatistical analysis is summarised in Figure 5.1 and consists of the following steps:

1. We extracted the latest STH prevalence data for each country and only considered data-sets that provided information on the year of data collection and geo-referenced coordinates for the sample locations.
2. We extracted climatic and environmental covariates and merged these with the STH prevalence data.
3. We assessed the relationships between covariates and STH prevalence, separately for each species. For countries where prevalence data were not available for each species, we instead used the prevalence of infection with any STH.

4. We tested for residual spatial correlation using the variogram computed on the random effects of a non-spatial Binomial mixed model.

5. The prevalence data were fitted to a Binomial geostatistical model via the Monte Carlo maximum likelihood method.

6. The calibration of the models was validated using the non-randomized probability integral transform.

7. If the model successfully passed the previous validation step, we then used this to generate predictive inferences at country-, sub-national- and pixel-level.

In the following paragraphs, we provide more details for each of the steps outlined above.

5.2.2 The ESPEN data on STH prevalence

The geographical area of interest in this study is the sub-Saharan region. Publicly available geo-referenced prevalence soil-transmitted Helminthiases (STH) survey data were extracted from the Expanded Special Project for Elimination of Neglected (ESPEN) tropical diseases database (https://espen.afro.who.int/). The ESPEN database is a publicly available database that stores data for several neglected tropical diseases. The most recent survey data were retrieved from the website for each country. Full details of data reporting to ESPEN can be found at https://espen.afro.who.int/. Our requirement for inclusion of a country was a sample size of at least 50 observations with complete information on the STH species (hookworm, roundworm, whipworm) or overall STH (any STH), the year of data collection, and geo-coordinates (longitude and latitude). The 50-sample size criterion was based on previous studies showing that small sample sizes of fewer than 50 data points in geostatistical data lead to issues such as overly noisy variograms. Furthermore, in geostatistical studies, small sample sizes (fewer than 50) result in variograms displaying little or no spatial correlation [38–41]. In total, 35 countries complied with this requirement. Figure 5.2 is a point map illustrating the locations of the observations that were used in the study. For the countries in grey, either the STH data were unavailable, or the sample size was less than 50.

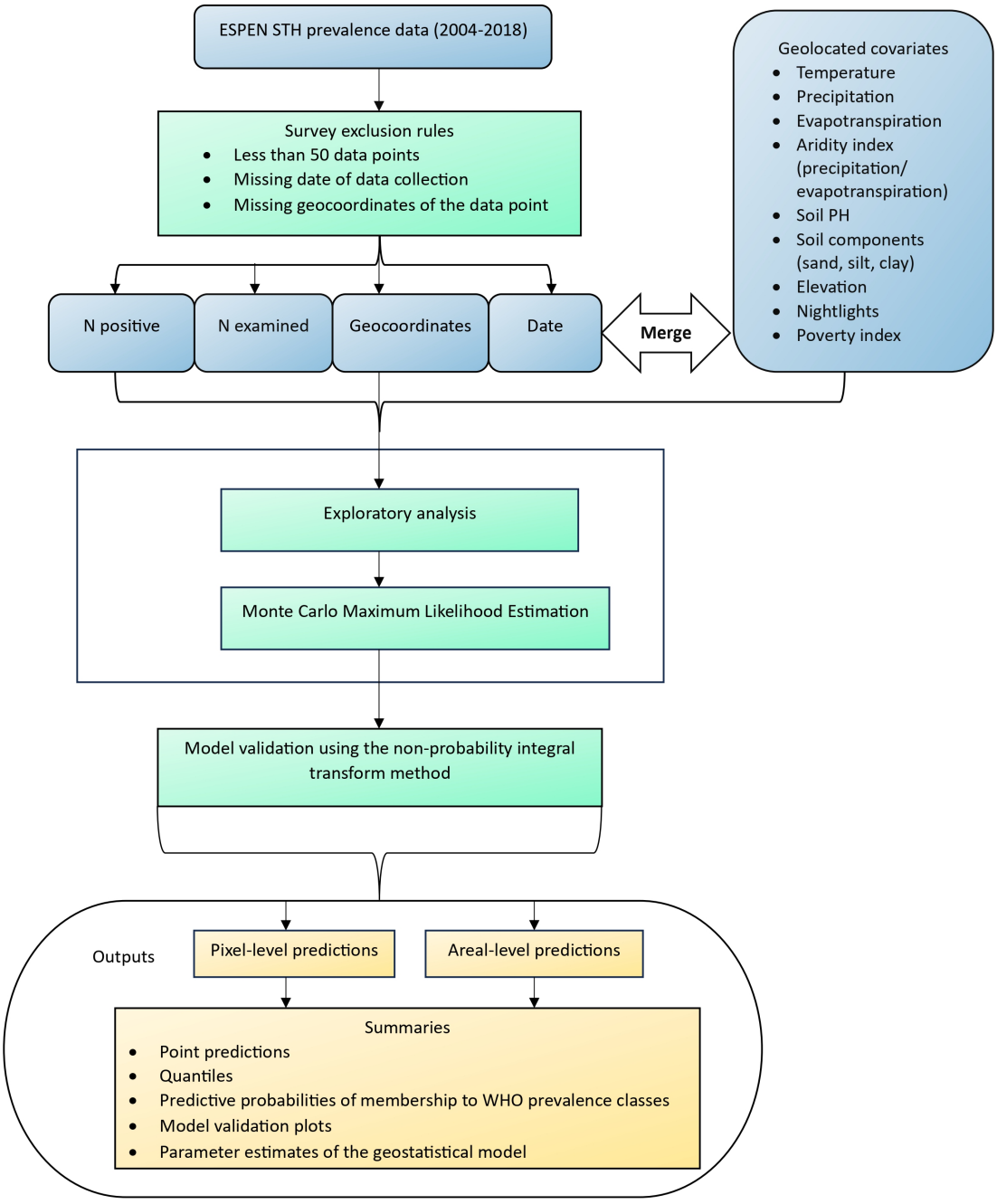

Figure 5.1: Schematic overview of the modelling and mapping procedures and techniques.
The blue boxes denote the input data or materials. The green boxes indicate processes, procedures, and models. The orange boxes describe the output data.

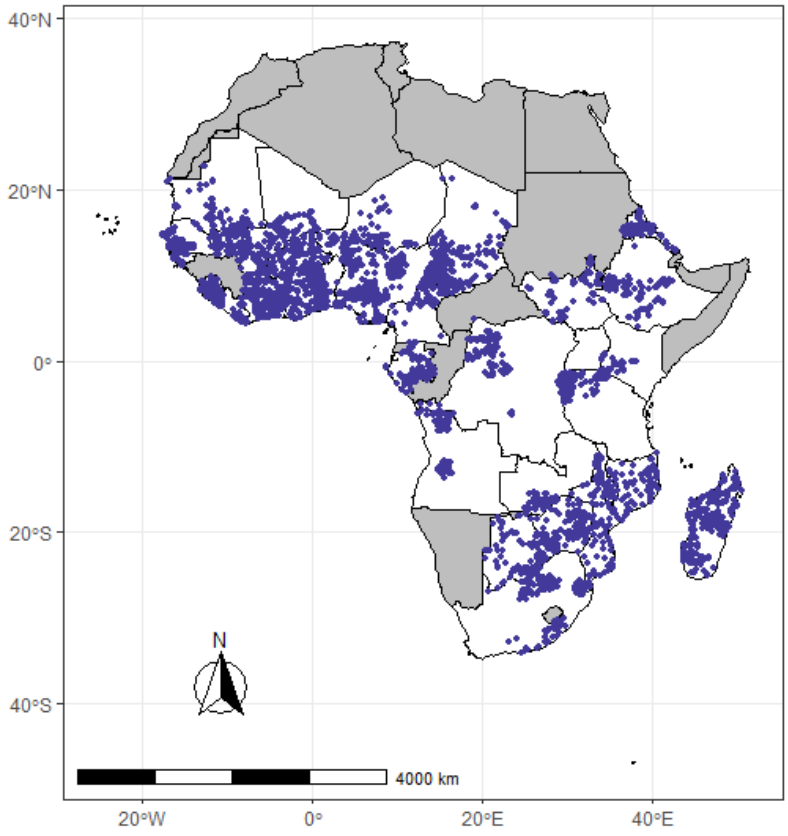

Figure 5.2: Map illustrating the locations of STH cases. The shaded areas represent countries with no data.

5.2.3 Climatic and environmental data

Our analysis uses spatially referenced climatic and environmental covariates that have been previously used to map STH prevalence [3]. More precisely, we considered maximum temperature, mean precipitation, and evapotranspiration, which were obtained from TerraClimate database [42]. An aridity index variable was derived by computing the proportion of the precipitation to the evapotranspiration of a country. An increase in the levels of climatic variables such as precipitation and aridity index have been shown in other studies to also increase the prevalence of STH [3]. Previous studies have also shown that the prevalence of STH decreases with an increase in the amount of soil PH and soil texture (clay, sand, silt). We, therefore, extracted covariates on soil acidity and soil texture (clay, sand, and silt) from the International and Soil Reference and Information Centre (ISRIC) [43]. Lastly, we downloaded elevation, nightlight and poverty index data from the Worldpop website [44]. Empirically, it has been found that higher altitudes are generally associated with lower STH risk, especially for *Trichiura* [28].

As expected, it has also been reported that an increase in wealth-related indicators is associated with a decrease in the prevalence of STH [3]. In this study, we used nightlights and poverty indices as proxies for estimating the level of wealth.

The spatial resolution and data sources for the covariates considered in this study are given in Table 5.1. The geographical locations (longitude and latitude) and year of data collection of the implementation units in the survey data were used to link the survey data to the spatial covariates.

Table 5.1: List of explanatory covariates used in the study and their spatial resolutions.

Name	Spatial resolution	Source
Soil type and content		
Soil PH in water	250 m	World Soil Information [43]
Soil type/texture fraction (sand, silt, clay)	250 m	World Soil Information [43]
Climatic variables		
Mean precipitation	4 km	TerraClimate [42]
Maximum temperature	4 km	TerraClimate [42]
Potential Evapotranspiration (PET)	4 km	TerraClimate [42]
Aridity index	4 km	Ratio of mean precipitation to PET
Other variables		
Elevation	100 m	Worldpop [44]
Nightlights	100 m	Worldpop [44]
Poverty index	1 km	Worldpop [44]

5.2.4 Data analysis

We first carry out an exploratory analysis to assess the relationship between STH prevalence (species-specific or overall STH) and the spatial covariates. We investigated multicollinearity and chose among highly correlated covariates (those with a correlation surpassing 0.6, following the recommendations and methodologies observed in prior research [45]. To select covariates, we fitted a Binomial generalized linear mixed model where, conditional on mutually-independent distributed Gaussian variables, Z_i, the logit linear predictor

for prevalence, for a given STH species, is defined as:

$$\log\left\{\frac{p_j\left(\mathbf{x}_i\right)}{1-p_j\left(\mathbf{x}_i\right)}\right\} = d\left(\mathbf{x}_i\right)\beta + Z_i \tag{5.1}$$

where $d\left(\mathbf{x}_i\right)$ is the vector of explanatory variables to be selected and β is a vector of regression coefficients.

The selection of covariates was carried out using a backward stepwise approach, in which the models were compared using the likelihood ratio test. After carrying out the selection of covariates, we tested for residual spatial correlation using the empirical variogram based on the random effects Z_i using a permutation test [35, 46]. If the residual spatial correlation was detected, we then fitted a geostatistical model, which is obtained by introducing a spatial Gaussian process, $S\left(\mathbf{x}_i\right)$ and, hence, we modify equation 5.1 as:

$$\log\left\{\frac{p_j\left(\mathbf{x}_i\right)}{1-p_j\left(\mathbf{x}_i\right)}\right\} = d\left(\mathbf{x}_i\right)\beta + S\left(\mathbf{x}_i\right) + Z_i \tag{5.2}$$

In the above equation, $S\left(\mathbf{x}_i\right)$ is a zero-mean stationary and isotropic Gaussian process with an exponential function with variance σ^2, hence

$$\mathrm{Cov}\left\{S\left(\mathbf{x}_i\right), S\left(\mathbf{x}_j\right)\right\} = \sigma^2 \exp\left\{-u_{ij}/\phi\right\}$$

where u_{ij} denotes any distance between any two areas $\mathbf{x}_i$ and $\mathbf{x}_j$ and ϕ is a scale parameter that determines the rate at which the spatial correlation decays to 0 as the distance u_{ij} increases. The exponential covariance function used in this study is a specific case of the Matérn covariance function, where the parameter *kappa* (κ) is set to 0.5 [46].

In countries where species-specific data were available, we fitted model 5.2 to each of the three species. For Mozambique, Togo, and Zimbabwe only the overall STH prevalence was available, hence, we fitted a single geostatistical model to this outcome. When fitting the model to species separately, we obtained the prevalence of infection with any STH species as:

$$1 - \{(1 - p_{\text{HK}}(x)) \times (1 - p_{\text{ASC}}(x)) \times (1 - p_{\text{TT}}(x))\}$$

where $p_{\text{HK}}(x)$, $p_{\text{ASC}}(x)$, and $p_{\text{TT}}(x)$ are the prevalence for hookworm, *Ascaris* and *Trichiura*, respectively. In the above equation, the expression for the prevalence of any STH species is obtained by assuming that the underlying spatial processes that modulate the three prevalences in the equation are independent conditionally on the spatial covariates used in the models. We point out that this assumption is less strong than the assumption of mutual independence between the three STH species that has been previously made in other studies [28, 47, 48].

The model parameters for equation 5.2 were estimated using a Monte Carlo maximum-likelihood (MCML) approach in the PrevMap package in R [49].

5.2.5 Model validation

To assess the model fit, we used the non-randomized probability integral transform (nrPIT) method that was first proposed for count data models and later adapted to validate binomial geostatistical models [35, 36]. If we let $Y = \{Y_i; 1 =, \ldots, n\}$ denote the vector of random variables of the number of STH (any STH or species-specific) positive cases; Y_i^* denote the random variable of the positive tested STH (any STH or species-specific) cases at a set of hold-out locations say x_j^* for $j =, \ldots, q$; and $Q(Z)$ denote the cumulative density function of a random variable Z; the nrPIT is defined as:

$$\text{nrPIT}(u \mid y_j^*, y) = \begin{cases} 0 & \text{if } u \leq Q(y_j^* - 1 \mid y) \\ \frac{[u - Q(y_j^* - 1 \mid y)]}{[Q(y_j^* \mid y) - Q(y_j^* - 1 \mid y)]} & \text{if } Q(y_j^* - 1 \mid y) \leq u \leq Q(y_j^* \mid y) \\ 1 & \text{if } u \geq Q(y_j^* \mid y) \end{cases}. \quad (5.3)$$

A detailed explanation of the nrPIT can be found in Appendix C for this paper and other work [35, 36]. Briefly, the nrPIT method uses the following steps:

1. Divide the dataset into a training set and a test set using a random approach.
2. Use the binomial geostatistical models that have been fitted to generate the predictive distribution of prevalence for the locations within the test set.
3. Employ the nrPIT to the positive cases observed in the test set.
4. Evaluate whether the transformed data from the nrPIT method conform to a uniform distribution by analyzing the cumulative density function.

The steps above were implemented for 30%, 40%, and 50% hold-out samples for each model.

For countries and species that validation indicated that the geostatistical models were well calibrated, we then proceeded to carry out predictions as explained in the next section.

5.2.6 Spatial prediction and policy-relevant criteria for STH interventions

For country and species data-sets analysed, we use the fitted geostatistical models to carry out inferences on the following predictive targets.

1. The **spatially continuous surface** of prevalence defined as:

$$p(A) = \{p(\mathbf{x}) : \mathbf{x} \in A\} \tag{5.4}$$

   where A denotes the area encompassed by the boundaries of a given country.

2. The **district-level prevalence**, which we define as follows. Let D_k be the set of spatial regions that partition the study country A into $k = 1, 2, .., K$ subunits. Then the predictive target for subunits was defined as:

$$p(D_k) = \frac{1}{|D_k|} \int_{D_k} p(\mathbf{x}) d\mathbf{x} \tag{5.5}$$

   where $|D_k|$ is the area for subunit k. The above integral is approximated using a regular grid covering $|D_k|$ with a spatial resolution of 95%. In this study, we used second-level administrative units from the Global Administrative Areas (GADM) website for each country as sub-national boundaries [50].

3. The **country-level prevalence**, which we define as:

$$p(A) = \frac{1}{|A|} \int_A p(\mathbf{x}) d\mathbf{x} \tag{5.6}$$

where A represents the area encompassed by the boundaries of a given country, as defined above.

We sample from the joint distribution of prevalence at all pixels and then aggregate according to equation 5.5 and equation 5.6 for the administrative-level and country-level predictions.

We obtained 10,000 predictive samples using the Laplace sampling approach implemented in the PrevMap package [49]. For the spatial continuous surface of prevalence, we use a regular grid covering a given country, whose spatial correlation (ϕ) is chosen so that the correlation between adjacent pixels is 95% [35, 51].

To classify the districts of a country into predefined classes of prevalence, we compute the predictive probability of falling in any given class based on the fitted models. For this, we use the WHO classification for STH prevalence, namely less than 2%, 2% to 10%, 10% to 20%, 20% to 50%, and greater than 50%. Hence, we allocate a district to one of those classes' prevalence based on the highest predictive probability.

5.3 Results

A total of 35 countries had STH data with at least 50 observations on the ESPEN database. The year of the last reported data-set on ESPEN varied from 2004 to 2018. About 67% of the data-sets are from 2014 onwards. The number of data points per country ranged from 50 to 1,054, with a median of 129 and an interquartile range of 86 to 265. The list of countries with their sample size and year of data collection can be found in the Shiny applications associated with this paper (Pixel-level results application and Subnational-level and other results application).

In the remainder of this section, we provide a summary of the results at the national level and provide a comprehensive summary of model validation for each country.

Taking Rwanda as a representative case, we further explain how to interpret the findings for each of the 35 countries, which can be accessed using the Shiny application at the links Pixel-level results application and Subnational-level and other results application.

5.3.1 Country-level results

5.3.1.1 Country-level predictions

Figure 5.3 shows the spatial distribution of the species-specific observed prevalence and overall STH prevalence at the country level in the countries where the models were calibrated. The binomial regression models indicate 11 of the 26 countries with a high prevalence (>20%) of any STH species and overall STH in countries such as Sierra Leone, Mozambique, Rwanda, and Zambia. The figure shows that the highest Hookworm prevalence was observed in the eastern and western parts of Africa. Conversely, the highest *Ascaris* prevalence was observed in southern and eastern Africa. The central and eastern parts of Africa had the highest predicted *Trichiura* prevalence. Overall, the highest prevalence of any STH was in western and eastern Africa, and it was predicted in Sierra Leone, Mozambique, Rwanda, and Zambia. The level of uncertainty, however, varied widely per species and within each country, as seen in the 95% confidence intervals of the estimates (Table 5.2) and associated uncertainty maps (Figure 5.4).

The uncertainty maps also illustrate the countries where predictions were produced with low confidence (indicated by high standard deviation, s.d.) and high confidence (indicated by low standard deviation). The levels of uncertainty were generally low (s.d. < 33) for all the species and countries.

Table 5.2 shows the overall prevalence and confidence intervals for the well-calibrated country models.

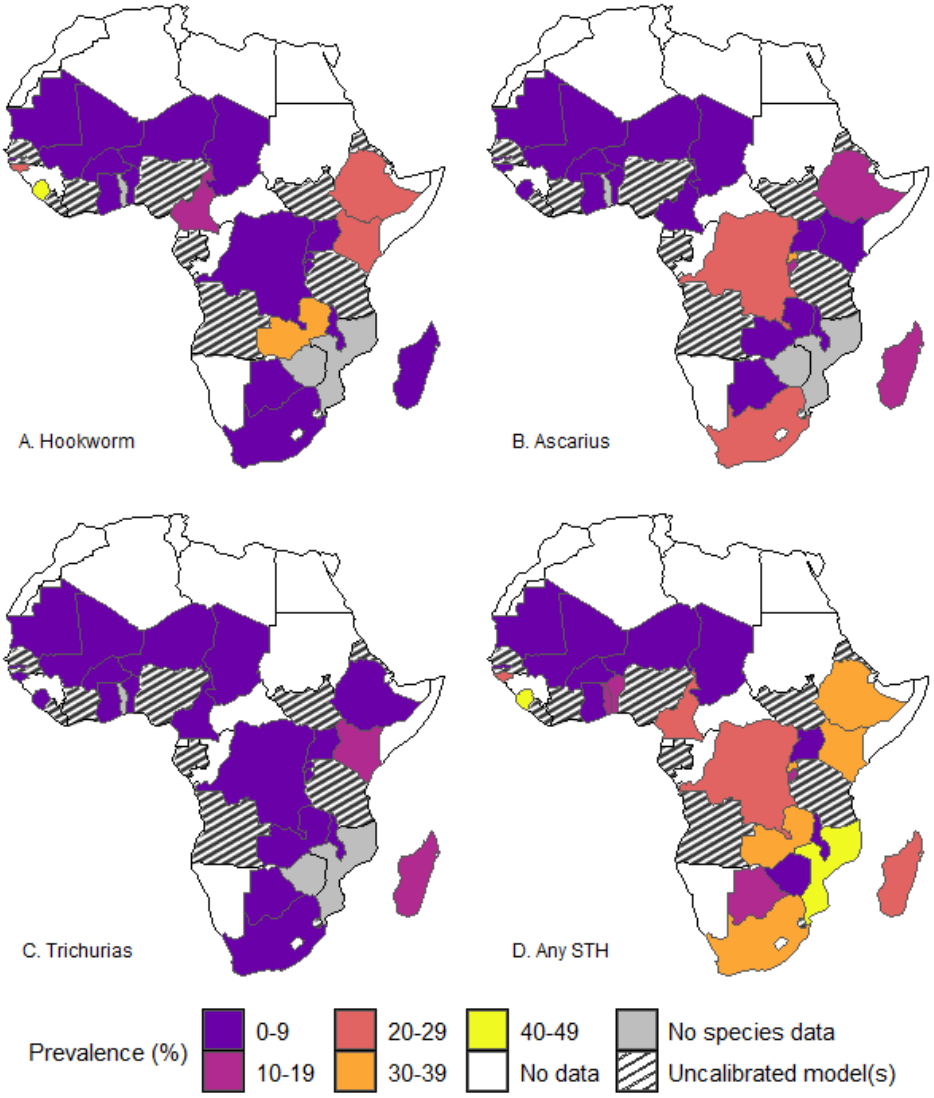

Figure 5.3: Map showing the country-level predicted geographic distribution of Hookworm (A), *Ascaris* (B), *Trichiura* (C), and overall STH (D).

5.3.1.2 Geostatistical model parameter estimates at country-level

Variable selection was performed for each country and species. The final selected covariates were utilized to construct predictive geostatistical models specific to each of the three STH species or any STH. In general, there was a negative association between nightlights and all of the species (Table C.3 in Appendix C). Similarly, the amount of soil PH and soil content (silt, sand, or clay) had a negative association with all three species. On the other hand, an increase in the aridity index and precipitation was associated with an increased risk of STH. Furthermore, an increase in the poverty index was associated with an increase in the odds of Hookworm (Table C.3 in Appendix C). The variance and scale of spatial correlation varied extensively

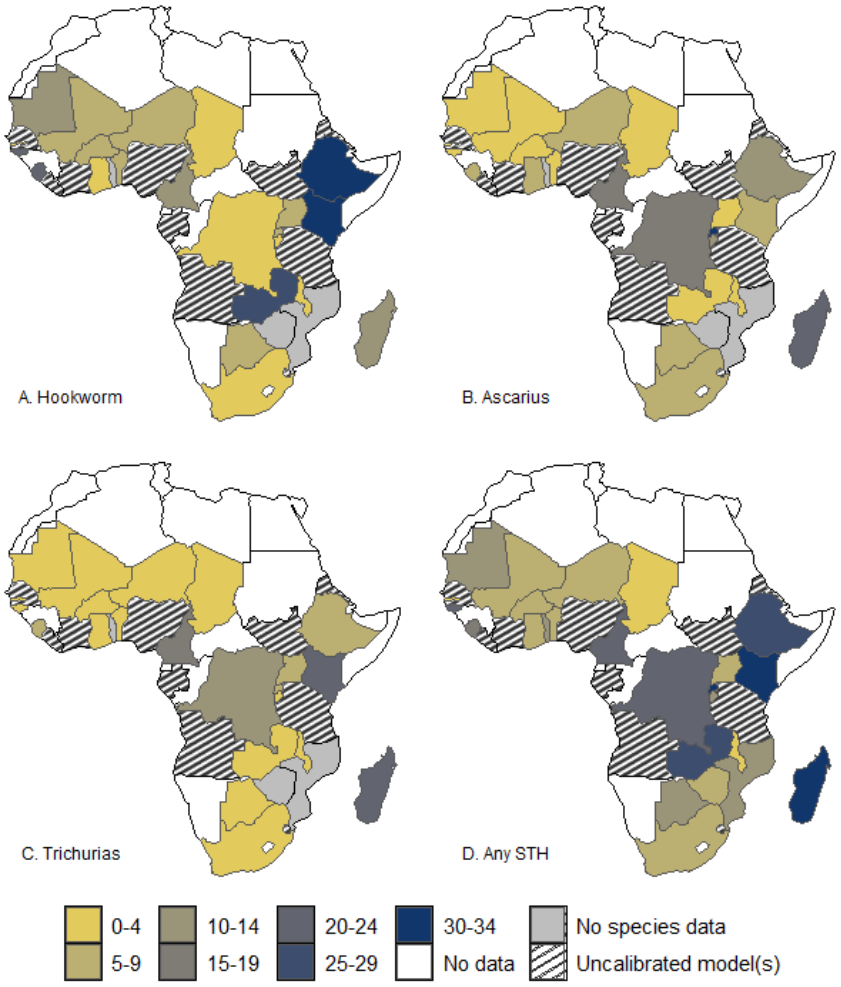

Figure 5.4: Maps showing the uncertainty (standard deviations) of the country-level predicted prevalence for Hookworm (A), *Ascaris* (B), *Trichiura* (C), and overall STH (D).

by country and species (exponents of coefficients in Figure C.1, Figure C.2, Figure C.3 and Figure C.4 in Appendix C).

5.3.1.3 Summaries of model validation at country-level

Table 5.3 shows the summary information on model validation for each country. A country was classified as having an uncalibrated model(s) if the validation for at least one of the hold-out samples in each model did not meet the criteria for being well-calibrated. Overall, 29% (10) of the 35 fitted country-models were uncalibrated in at least one of the holdout samples.

Table 5.2: Country-level predicted prevalence estimates and associated 95% confidence intervals.

Country	**Year**	**Hookworm** Estimate (95% CI)	***Ascaris*** Estimate (95% CI)	***Trichiura*** Estimate (95% CI)
Southern Africa				
Botswana	2015	5.2% (0.1%,35.4%)	5.6% (0.1%, 34.8%)	1.1% (0.0%,9.9%)
South Africa	2017	3.2% (0.7%, 8.4%)	27.8% (14.2%,45.2%)	0.1% (0.1%,0.2%)
Central Africa				
Cameroon	2012	13.0% (0.6%,44.3%)	5.6% (0.0%,66.9%)	8.3% (0.0%,69.9%)
DRC	2015	1.1% (0.4%, 2.5%)	20.9% (1.3%, 68.6%)	6.3% (0.0%, 58.8%)
Chad	2015	0.2% (0.0%,1.5%)	0.8% (0.0%,5.2%	0.1% (0.0%,0.4%)
Eastern Africa				
Burundi	2014	4.6% (0.8%,14.2%)	12.5% (1.2%,38.2%)	2.9% (0.2%,12.1%)
Ethiopia	2009	23.2% (0.0%,97.0%)	10.3% (0.2%,54.1%)	3.0% (0.0%,22.8%)
Kenya	2015	24.6% (0.0%, 95.4%)	2.6% (0.0%,18.0%)	10.3% (0.0%,87.5%)
Madagascar	2015	5.3% (0.0%,43.6%)	15.4% (0.1%, 80.6%)	14.2% (0.0%,90.6%)
Malawi	2018	0.9% (0.1%,3.6%)	1.7% (0.1%,7.3%)	0.1% (0.0%, 0.2%)
Mozambique	2007	NA	NA	NA
Rwanda	2014	4.7% (0.3%, 20.3%)	33.7% (1.1%, 98.8%)	1.8% (0.0%, 18.7%)
Uganda	2016	3.5% (0.1%,18.1%)	0.9% (0.0%,6.3%)	2.3% (0.0%,18.3%)
Zambia	2005	36.0% (1.2%,93.6%)	0.9% (0.0%,4.9%)	0.2% (0.0%,1.0%)
Zimbabwe	2010	NA	NA	NA
Western Africa				
Benin	2017	9.4% (1.3%,27.5%)	0.9% (0.0%,7.8%)	0.4% (0.3%,0.4%)
Burkina Faso	2004	3.0% (0.0%,20.3%)	0.0% (0.0%,0.0%)	0.4% (0.2%,0.6%)
Ghana	2008	2.9% (0.2%,11.8%)	3.3% (0.0%,25.1%)	0.3% (0.0%,1.4%)
Guinea-Bissau	2018	26.6% (1.2%,81.5%)	0.1% (0.0%,0.2%)	0.2% (0.1%, 0.4%)
Mali	2004	2.2% (0.0%,25.9%)	0.0% (0.0%,0.1%)	0.2% (0.0%,0.6%)
Mauritania	2015	7.2% (0.1%,49.2%)	1.8% (0.8%,3.4%)	0.7% (0.0%,4.7%)
Niger	2006	3.3% (0.0%,25.7%)	1.0% (0.0%,8.2%)	0.1% (0.0%,0.4%)
Sierra Leone	2008	40.9% (9.1%,82.5%)	7.7% (0.8%, 26.0%)	3.3% (0.1%, 18.2%)
The Gambia	2015	0.3% (0.1%,1.2%)	0.4% (0.0%,2.3%)	0.1% (0.0%, 0.1%)
Togo	2015	NA	NA	NA

CI = Confidence interval.
DRC = Democratic Republic of the Congo (Congo Kinshasa).
NA = Not available.

Table 5.3: Summary of model validation analyses per country.

Country	Year	Prevalence (%)				ϕ (km)				Calibrated
		HK	ASC	TT	Any STH	HK	ASC	TT	Any STH	Model (s)
Southern Africa										
Botswana	2015	55.0	41.0	54.0		391.9	34.2	445.7		Yes
South Africa	2017	24.0	70.0	3.0		384.9	67.7	12.3		Yes
Swaziland	2015	20.0	90.0	70.0		154.4	56.7	263.3		No
Central Africa										
Angola	2014	80.0	100.0	43.0		173.1	181.4	69.0		No
Cameroon	2012	60.0	73.0	72.0		94.1	642.2	213.9		Yes
Chad	2015	20.0	36.0	26.0		58.4	134.4	28.7		Yes
DRC	2015	60.0	88.0	94.0		3,027.4	100.4	1,102.2		Yes
Gabon	2015	91.0	100.0	100.0		65.2	8.3	28.4		No
Eastern Africa										
Burundi	2014	38.0	70.0	36.0		43.1	49.2	21.3		Yes
Eritrea	2015	4.0	2.0	8.0		61.5	26.9	475.3		No
Ethiopia	2009	75.0	57.0	54.0		32.3	114.5	128.9		Yes

Continued on next page

Table 5.3 – continued from previous page

Country	Year	Prevalence (%)				ϕ (km)				Calibrated
		HK	ASC	TT	Any STH	HK	ASC	TT	Any STH	Model (s)
Kenya	2015	50.0	32.0	50.0		52.5	241.7	174.5		Yes
Madagascar	2015	52.0	96.0	98.0		25.9	144.0	169.1		Yes
Malawi	2018	20.0	37.0	7.0		12.3	13.1	11.8		Yes
Mozambique	2007				82				158.2	Yes
Rwanda	2014	44.0	100.0	100.0		21.7	19.8	72.9		Yes
South Sudan	2018	67.0	36.0	29.0		279.9	81.9	268.1		No
Tanzania (Mainland)	2018	50.0	27.0	43.0		78.5	366.2	153.9		No
Uganda	2016	23.0	9.0	12.0		91.8	643.1	202.8		Yes
Zambia	2005	87.0	33.0	12.0		50.7	97.5	157.3		Yes
Zimbabwe	2010				78				50.8	Yes
Western Africa										
Benin	2017	45.0	34.0	4.0		90.8	123.0	29.6		Yes
Burkina Faso	2004	75.0	2.0	5.0		107.1	19.6	716.2		Yes
Cote d'Ivoire	2014	78.0	56.0	74.0		98.0	626.2	88.0		No
The Gambia	2015	12.0	60.0	8.0		44.8	11.0	4.9		Yes
Ghana	2008	27.0	20.0	5.0		141.8	32.9	120.5		Yes

Continued on next page

Table 5.3 – continued from previous page

Country	Year	Prevalence (%)				ϕ (km)				Calibrated
		HK	ASC	TT	Any STH	HK	ASC	TT	Any STH	Model (s)
Guinea-Bissau	2018	100.0	4.0	12.0		21.6	26.6	1.1		Yes
Liberia	2015	42.0	100.0	16.0		94.6	19.8	91.6		No
Mali	2004	100.0	6.0	7.0		104.1	370.9	229.1		Yes
Mauritania	2015	100.0	8.0	10.0		68.9	92.1	915.6		Yes
Niger	2006	5.0	27.0	8.0		193.5	67.5	1,316.5		Yes
Nigeria	2014	86.0	94.0	77.0		112.8	58.5	228.3		No
Senegal	2013	61.0	64.0	78.0		43.3	34.9	43.1		No
Sierra Leone	2008	95.0	25.0	30.0		43.9	22.3	24.0		Yes
Togo	2015				100				25.8	Yes

HK = Hookworm, ASC = *Ascaris*, TT = *Trichiura*.

ϕ = Estimated scale of spatial correlation.

DRC = Democratic Republic of the Congo (Congo Kinshasa).

5.3.2 Country example: Rwanda

5.3.2.1 Predicted prevalence of STH in Rwanda

The predicted point prevalence of both STH species and overall STH in Rwanda are presented in Figure 5.5. Overall, the predicted prevalence of any STH species and overall STH is heterogeneously distributed across Rwanda. A notably heightened burden of STH infections was documented in the western regions of Rwanda, with *Ascaris* demonstrating the highest prevalence, closely followed by *Trichiura* (Figure 5.5). These findings are also evident in the sub-national predicted prevalence maps (Figure 5.6). The confidence intervals for both the point and sub-national prevalence maps are given in the Shiny application.

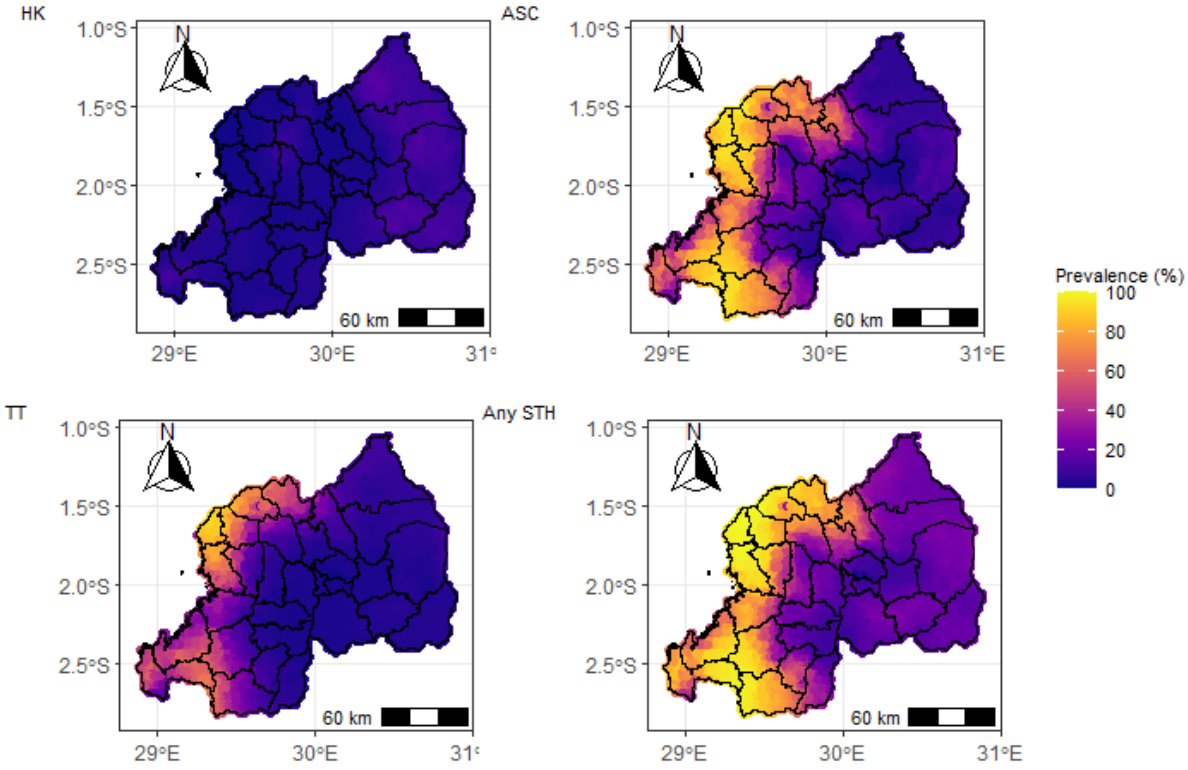

Figure 5.5: Map showing the pixel-level predicted geographic distribution of the prevalence of STH in Rwanda (HK = Hookworm, ASC = *Ascaris*, TT = *Trichiura* and Any STH = Overall STH)

5.3.2.2 Point and exceedance probability maps of soil-transmitted helminths in Rwanda

The binomial regression models indicate a lot of areas with a high prevalence (> 20%) of any STH species and overall STH in Rwanda. Figures 5.7 and 5.8 show the

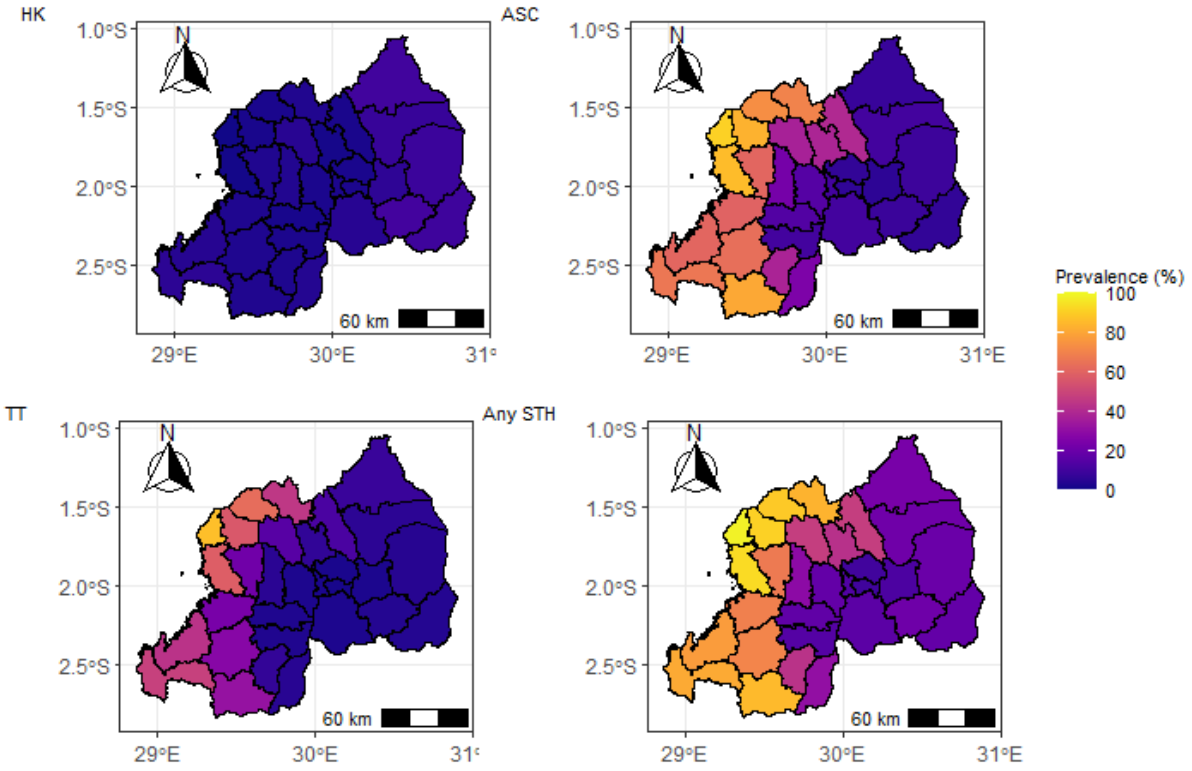

Figure 5.6: Map showing the subnational-level predicted geographic distribution of the prevalence of STH in Rwanda (HK = Hookworm, ASC = *Ascaris*, TT = *Trichiura* and Any STH = Overall STH)

WHO predicted endemicity class STH treatment at pixel and sub-national levels. The maps depict high exceedance probabilities in the central and the western sides of Rwanda. These are, therefore, the treatment priority areas for STH.

5.3.2.3 Geostatistical model parameter estimates for Rwanda

The modeling suggests a strong relationship between rainfall and *Ascaris* and *Trichiura* in Rwanda (Table 5.4). An increase in the amount of rainfall was seen to increase the prevalence of the two species. Conversely, soil content (sand, silt, clay) was associated with a reduction in the prevalence of all STH species. Likewise, an increase in nightlights was associated with a reduction in the prevalence of *Ascaris* and *Trichiura*.

Table 5.4 also shows the differences in the covariance parameters for the three species. The point estimates for the scale parameter were 21.73 km (Hookworm), 19.77 km (*Ascaris*), and 72.91 km (*Trichiura*).

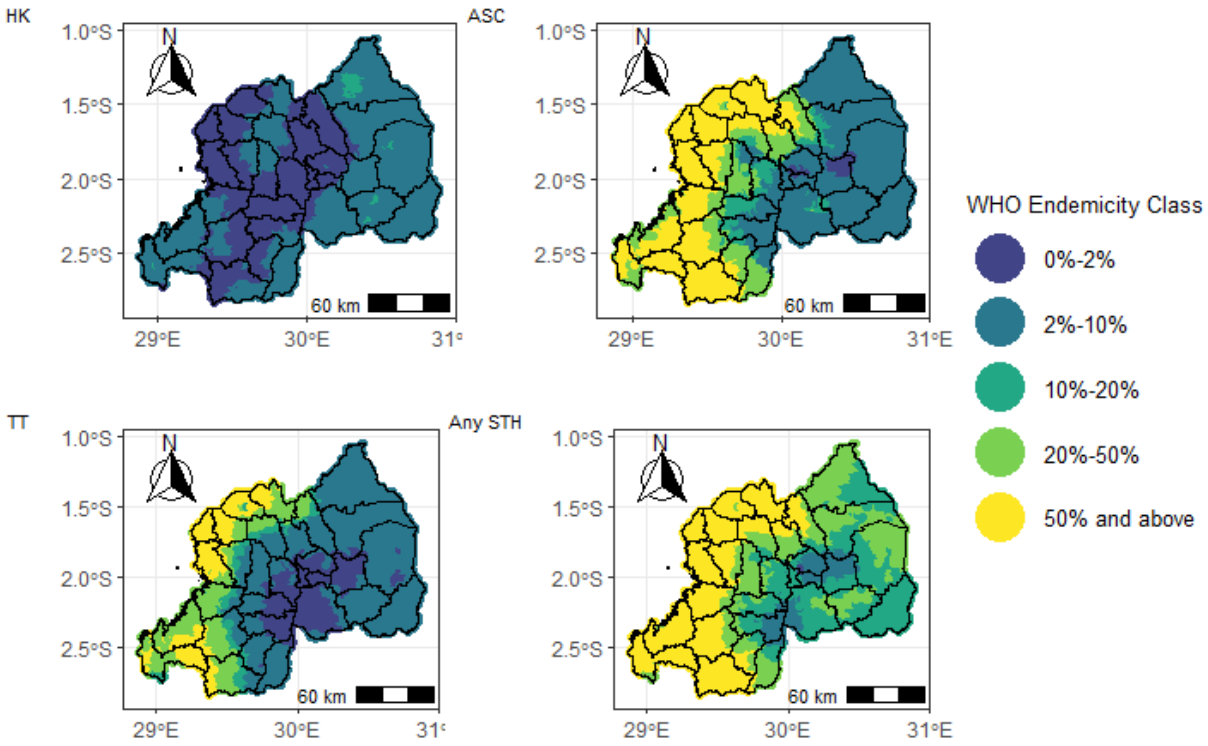

Figure 5.7: Map showing the predicted STH (HK = Hookworm, ASC = *Ascaris*, TT = *Trichiura*, STH = any STH) endemicity class in Rwanda at the pixel level from the Binomial regression model in 5.2.

5.3.2.4 Model validation for Rwanda

Figure 5.9 illustrates the model validation plots for the Rwanda models. The figure shows that the observed nrPIT curves (represented by the solid black line) from the three hold-out samples for all three species fall within the 95% envelope (denoted by the dashed lines). We, therefore, conclude that we do not have enough evidence to reject the null hypothesis of well-calibrated models.

5.4 Discussion

In this study, we have carried out a comprehensive geostatistical analysis of soil-transmitted infections data from the ESPEN database. We developed geostatistical models separately for each country, so as to tailor the selection of spatial covariates and estimation of covariance parameters to the heterogeneous spatial patterns across countries. In countries where the geostatistical models were validated successfully, we proceeded to generate predictions of STH prevalence at

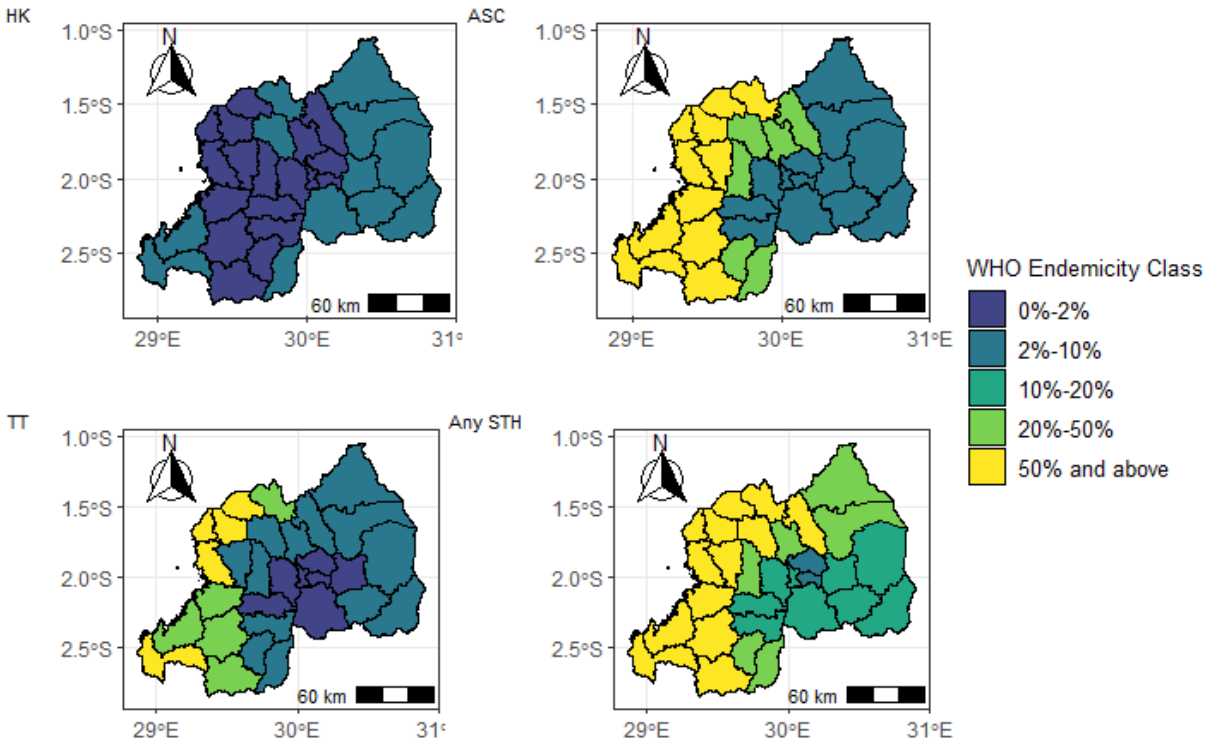

Figure 5.8: Map showing the predicted STH (HK = Hookworm, ASC = *Ascaris*, TT = *Trichiura*, STH = any STH) endemicity class in Rwanda at the subnational level from the Binomial regression model in 5.2.

both national and sub-national levels.

The selection of covariates used to assist in the geostatistical prediction of prevalence showed different results across countries. However, notably, due to the weak empirical strength of association with disease prevalence, only a few covariates were selected for most countries. The low predictive power of the spatial covariates may be attributed to the relatively low prevalence levels that are observed in most countries, which make the estimation of regression relationships more cumbersome. Despite these challenges, where predictors were included, they provided some comparable estimates with findings from previous studies. For instance, areas with increased precipitation were associated with a higher likelihood of all STH species, consistent with existing research indicating higher prevalence in wetter regions [3, 30, 31, 52]. Similarly, the observation that an increased amount of nightlights, serving as a proxy for wealth status, decreased the likelihood of all STH species aligns with the established notion of higher

Table 5.4: Monte Carlo maximum likelihood estimates and associated 95% confidence intervals for the model in Equation 5.2 for Rwanda.

Parameter	**Hookworm** Estimate (95% CI)	***Ascaris*** Estimate (95% CI)	***Trichiura*** Estimate (95% CI)
Intercept	-1.555 (-2.739, -0.372)	-6.153 (-6.883, -5.422)	-5.307 (-7.481, -3.133)
Soil	-0.007 (-0.012, -0.001)	-0.008 (-0.010, -0.006)	-0.005 (-0.007, -0.002)
Precipitation	NA	0.096 (0.092, 0.101)	0.064 (0.039, 0.090)
Nightlights	NA	-0.168 (-0.227, -0.108)	-0.259 (-0.420, -0.100)
σ^2	1.131 (0.529,2.420)	1.441 (0.858,2.418)	1.750 (0.459, 6.670)
ϕ	21.727 (8.313, 56.785)	19.774 (8.972,43.584)	72.933 (14.913,356.672)

CI = Confidence interval.
Soil = sand, clay, or silt content.
NA corresponds to the situation where the covariate was not included in the model.
σ^2 = Estimated variance; ϕ = Estimated scale of spatial correlation.

prevalence in economically disadvantaged areas [3, 29]. Additionally, the finding that soil pH and content (sand, silt, clay) reduced the likelihood of STH is also consistent with previous research findings [3, 28, 52, 53].

The analysis reveals significant heterogeneity in the estimates of the scale of spatial correlation and the variance of residual spatial variation across countries. The scale of spatial correlation ranged from 1.14 km to 3,027.44 km, while the variance ranged from 0.02 to 95.01 across the countries. Therefore, the wide variations in the estimates of spatial correlation across countries, coupled with observed non-stationarity, further justify the use of species-specific, single-country models for this STH data. The non-stationarity is likely driven by differing control intervention histories across countries, which are challenging to capture adequately using the available spatial covariates. These intervention histories can significantly influence the spatial distribution and prevalence of STH, leading to localized variations that a global model might fail to account for.

Moreover, it was observed that the geostatistical models exhibited inadequate

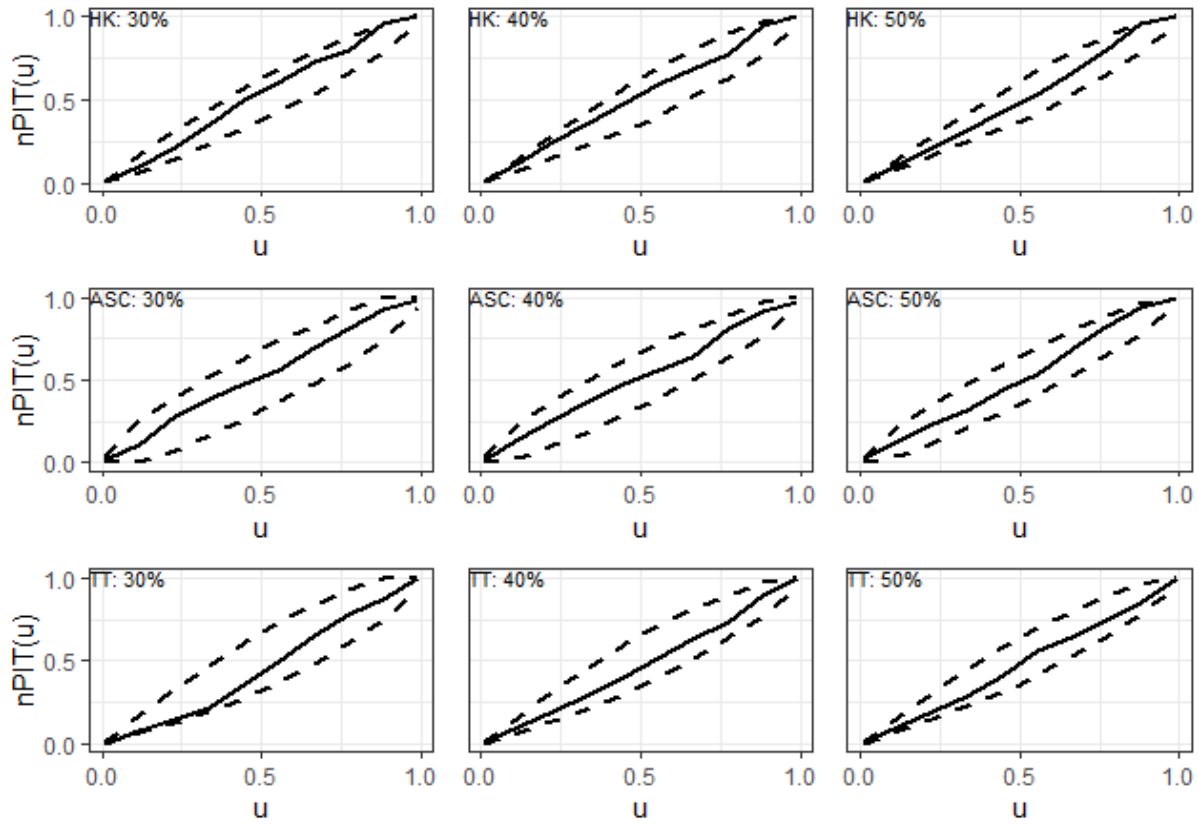

Figure 5.9: Plots of the non-randomized probability integral transform (nrPIT) calculated for three (30%, 40%, 50%) hold-out samples for Hookworm (HK), *Ascaris* (ASC), and *Trichiura* (TT).

calibration in certain countries, prohibiting spatial predictions at unsampled locations. This issue may be attributed to a combined effect of very sparse data and small estimated spatial correlations relative to the study area. For some countries where the estimated variance of the residual spatial process is relatively small, an additional explanation for the poor calibration of the geostatistical models might be the presence of strong noise components that diminish the spatial signal within the data. These findings consequently urge caution in developing an Africa-wide model based solely on ESPEN data, given the observed heterogeneity in the model parameter estimates and the challenges encountered in model calibration across different regions and species.

In our study, we used data for a single time point for all the countries, namely the most recent survey. Hence, one of the main limitations is the absence of a spatio-temporal geostatistical model that could make full use of all the historical

data. However, the availability of data over time varies from country to country, with some countries providing only a single survey. The average number of surveys per country was 6, with the minimum being 1 survey and the maximum being 15 surveys per country. An additional challenge in building credible spatio-temporal models for STH is the effect on prevalence trends due to mass drug administration (MDA). Information on the frequency and coverage of MDA is an essential element that should be incorporated in such models; however, not all countries provide this information at suitable spatial and temporal resolutions for geostatistical models. Future research should aim to bridge geostatistical models with mathematical models capable of integrating MDA data, offering a valuable approach for combining information from baseline to impact surveys.

5.5 Conclusion

This study demonstrates the use of model-based geostatistics to harness ESPEN data, offering valuable insights into the spatial distribution of STH prevalence across countries. While ESPEN data serve as a crucial resource for understanding spatial patterns in STH prevalence through geostatistical models, inherent limitations arise from the sparsity of data, both temporally and spatially in certain countries, constraining the applicability of such models. Nevertheless, the predictive inferences derived from these models, where possible, provide useful information for national control programs, facilitating targeted interventions and informing survey designs for future STH assessments.

References

[1] A. Montresor, D. Mupfasoni, A. Mikhailov, P. Mwinzi, et al. "The global progress of soil-transmitted helminthiases control in 2020 and World Health Organization targets for 2030". In: *PLOS Neglected Tropical Diseases* 14.8 (2020), e0008505.

[2] World Health Organization (WHO). *Soil-transmitted helminth infections*. Accessed May 2022. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections.

[3] B. Sartorius, J. Cano, H. Simpson, L. S. Tusting, et al. "Prevalence and intensity of soil-transmitted helminth infections of children in sub-Saharan Africa, 2000–18: a geospatial analysis". In: *The Lancet Global Health* 9.1 (2021), e52–e60.

[4] R. L. Pullan, J. L. Smith, R. Jasrasaria, and S. J. Brooker. "Global numbers of infection and disease burden of soil transmitted helminth infections in 2010". In: *Parasites & Vectors* 7 (2014), pp. 1–19.

[5] N. Pabalan, E. Singian, L. Tabangay, H. Jarjanazi, et al. "Soil-transmitted helminth infection, loss of education and cognitive impairment in school-aged children: A systematic review and meta-analysis". In: *PLoS Neglected Tropical Diseases* 12.1 (2018), e0005523.

[6] S. Novianty, Y. Dimyati, S. Pasaribu, and A. P. Pasaribu. "Risk factors for soil-transmitted helminthiasis in preschool children living in farmland, North Sumatera, Indonesia". In: *Journal of Tropical Medicine* (2018).

[7] G. Raso, P. Vounatsou, L. Gosoniu, M. Tanner, et al. "Risk factors and spatial patterns of hookworm infection among schoolchildren in a rural area of western Côte d'Ivoire". In: *International Journal for Parasitology* 36.2 (2006), pp. 201–210.

[8] B. Levecke, L. E. Coffeng, C. Hanna, R. L. Pullan, et al. "Assessment of the required performance and the development of corresponding program decision rules for neglected tropical diseases diagnostic tests: Monitoring and evaluation of soil-transmitted helminthiasis control programs as a case study". In: *PLoS Neglected Tropical Diseases* 15.9 (2021), e0009740.

[9] World Health Organization. *Preventive chemotherapy in human helminthiasis. Coordinated use of anthelminthic drugs in control interventions: a manual for*

health professionals and programme managers. World Health Organization, 2006.

[10] A. D. Hopkins. "Neglected tropical diseases in Africa: a new paradigm". In: *International Health* 8.1 (2016), pp. i28–i33.

[11] K. M. Fornace, C. Fronterre, F. M. Fleming, H. Simpson, et al. "Evaluating survey designs for targeting preventive chemotherapy against Schistosoma haematobium and Schistosoma mansoni across sub-Saharan Africa: a geostatistical analysis and modelling study". In: *Parasites & Vectors* 13 (2020), pp. 1–13.

[12] C. A. Schmidt, E. A. Cromwell, E. Hill, K. M. Donkers, et al. "The prevalence of onchocerciasis in Africa and Yemen, 2000–2018: A geospatial analysis". In: *BMC Medicine* 20.1 (2022), p. 293.

[13] O. A. Eneanya, C. Fronterre, I. Anagbogu, C. Okoronkwo, et al. "Mapping the baseline prevalence of lymphatic filariasis across Nigeria". In: *Parasites & Vectors* 12.1 (2019), pp. 1–13.

[14] M. O. Afolabi, A. Adebiyi, J. Cano, B. Sartorius, et al. "Prevalence and distribution pattern of malaria and soil-transmitted helminth co-endemicity in sub-Saharan Africa, 2000–2018: A geospatial analysis". In: *PLoS Neglected Tropical Diseases* 16.9 (2022), e0010321.

[15] P. J. Diggle, J. A. Tawn, and R. A. Moyeed. "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3 (1998), pp. 299–350.

[16] P. J. Diggle and E. Giorgi. "Model-based geostatistics for prevalence mapping in low-resource settings". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1096–1120.

[17] R. J. S. Magalhães, A. C. Clements, A. P. Patil, P. W. Gething, et al. "The applications of model-based geostatistics in helminth epidemiology and control". In: *Advances in parasitology* 74 (2011), pp. 267–296.

[18] C. Fronterre, B. Amoah, E. Giorgi, M. C. Stanton, et al. "Design and analysis of elimination surveys for neglected tropical diseases". In: *The Journal of Infectious Diseases* 221.5 (2020), S554–S560.

[19] O. Johnson, C. Fronterre, B. Amoah, A. Montresor, et al. "Model-based geostatistical methods enable efficient design and analysis of prevalence surveys for soil-transmitted helminth infection and other neglected tropical diseases". In: *Clinical Infectious Diseases* 72.3 (2021), S172–S179.

[20] P. J. Diggle, B. Amoah, C. Fronterre, E. Giorgi, et al. "Rethinking neglected tropical disease prevalence survey design and analysis: a geospatial paradigm". In: *Transactions of The Royal Society of Tropical Medicine and Hygiene* 115.3 (2021), pp. 208–210.

[21] B. Amoah, C. Fronterre, O. Johnson, M. Dejene, et al. "Model-based geostatistics enables more precise estimates of neglected tropical-disease prevalence in elimination settings: mapping trachoma prevalence in Ethiopia". In: *International Journal of Epidemiology* 51.2 (2022), pp. 468–478.

[22] B. Sartorius, J. D. VanderHeide, M. Yang, E. A. Goosmann, et al. "Subnational mapping of HIV incidence and mortality among individuals aged 15–49 years in sub-Saharan Africa, 2000–18: a modelling study". In: *The Lancet HIV* 8.6 (2021), e363–e375.

[23] E. A. Cromwell, C. A. Schmidt, K. T. Kwong, D. M. Pigott, et al. "The global distribution of lymphatic filariasis, 2000–18: a geospatial analysis". In: *The Lancet Global Health* 8.9 (2020), e1186–e1194.

[24] N. V. Bhattacharjee, L. E. Schaeffer, and S. I. Hay. "Mapping inequalities in exclusive breastfeeding in low-and middle-income countries, 2000–2018". In: *Nature Human Behaviour* 5.8 (2021), pp. 1027–1045.

[25] N. Golding, R. Burstein, J. Longbottom, A. J. Browne, et al. "Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals". In: *The Lancet* 390.10108 (2017), pp. 2171–2182.

[26] N. Graetz, J. Friedman, A. Osgood-Zimmerman, R. Burstein, et al. "Mapping local variation in educational attainment across Africa". In: *Nature* 555.7694 (2018), pp. 48–53.

[27] A. Osgood-Zimmerman, A. I. Millear, R. W. Stubbs, C. Shields, et al. "Mapping child growth failure in Africa between 2000 and 2015". In: *Nature* 555.7694 (2018), pp. 41–47.

[28] H. O. Mogaji, O. O. Johnson, A. B. Adigun, O. N. Adekunle, et al. "Estimating the population at risk with soil transmitted helminthiasis and annual drug requirements for preventive chemotherapy in Ogun State, Nigeria". In: *Scientific Reports* 12.1 (2022), p. 2027.

[29] R. B. Yapi, F. Chammartin, E. Hürlimann, C. A. Houngbedji, et al. "Bayesian risk profiling of soil-transmitted helminth infections and estimates

of preventive chemotherapy for school-aged children in Cote d'Ivoire". In: *Parasites & Vectors* 9 (2016), pp. 1–9.

[30] R. L. Pullan, P. W. Gething, J. L. Smith, C. S. Mwandawiro, et al. "Spatial modelling of soil-transmitted helminth infections in Kenya: a disease control planning tool". In: *PLoS Neglected Tropical Diseases* 5.2 (2011), e958.

[31] S.-Y. Huang, Y.-S. Lai, and Y.-Y. Fang. "The spatial-temporal distribution of soil-transmitted helminth infections in Guangdong Province, China: A geostatistical analysis of data derived from the three national parasitic surveys". In: *PLoS Neglected Tropical Diseases* 16.7 (2022), e0010622.

[32] M. Assoum, G. Ortu, M.-G. Basáñez, C. Lau, et al. "Spatiotemporal distribution and population at risk of soil-transmitted helminth infections following an eight-year school-based deworming programme in Burundi, 2007–2014". In: *Parasites & Vectors* 10 (2017), pp. 1–12.

[33] T. Tsheten, K. A. Alene, A. C. Restrepo, M. Kelly, et al. "Risk mapping and socio-ecological drivers of soil-transmitted helminth infections in the Philippines: a spatial modelling study". In: *The Lancet Regional Health–Western Pacific* 43 (2024).

[34] D. J. Gerber, S. Dhakal, M. N. Islam, A. Al Kawsar, et al. "Distribution and treatment needs of soil-transmitted helminthiasis in Bangladesh: A Bayesian geostatistical analysis of 2017-2020 national survey data". In: *PLoS Neglected Tropical Diseases* 17.11 (2023), e0011656.

[35] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

[36] C. Czado, T. Gneiting, and L. Held. "Predictive model assessment for count data". In: *Biometrics* 65.4 (2009), pp. 1254–1261.

[37] G. Varoquaux. "Cross-validation failure: Small sample sizes lead to large error bars". In: *Neuroimage* 180 (2018), pp. 68–77.

[38] R. Kerry and M. Oliver. "Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood". In: *Geoderma* 140.4 (2007), pp. 383–396.

[39] M. M. Rufino, V. Stelzenmüller, F. Maynou, and G.-P. Zauke. "Assessing the performance of linear geostatistical tools applied to artificial fisheries data". In: *Fisheries Research* 82.1-3 (2006), pp. 263–279.

[40] R. Webster and M. A. Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, 2007.

[41] R. Webster and M. A. Oliver. "Sample adequately to estimate variograms of soil properties". In: *Journal of soil science* 43.1 (1992), pp. 177–192.

[42] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". In: *Nature Scientific Data* 5.1 (2018), pp. 1–12.

[43] *ISRIC World Soil Information*. Accessed June 2022. 2022. URL: https://www.isric.org/.

[44] A. J. Tatem. "WorldPop, open data for spatial demography". In: *Scientific data* 4.1 (2017), pp. 1–4.

[45] O. Wariri, C. E. Utazi, U. Okomo, C. J. E. Metcalf, et al. "Mapping the timeliness of routine childhood vaccination in the Gambia: a spatial modelling study". In: *Vaccine* 41.39 (2023), pp. 5696–5705.

[46] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

[47] N. de Silva and A. Hall. "Using the prevalence of individual species of intestinal nematode worms to estimate the combined prevalence of any species". In: *PLoS Negl Trop Dis* 4.4 (2010), e655.

[48] ESPEN. *Soil Transmitted Helminths site-level data codebook*. https://espen.afro.who.int/diseases/soil-transmitted-helminthiasis. Accessed June 2022. 2022.

[49] E. Giorgi and P. J. Diggle. "PrevMap: an R package for prevalence mapping". In: *Journal of Statistical Software* 78 (2017), pp. 1–29.

[50] GADM. *Global Administrative Areas (GADM) maps and data*. https://gadm.org. Accessed August 2024. 2022.

[51] M. Sasanami, B. Amoah, A. N. Diori, A. Amza, et al. "Using model-based geostatistics for assessing the elimination of trachoma". In: *PLoS Negl Trop Dis* 17.7 (2023), e0011476.

[52] D.-A. Karagiannis-Voules, P. Biedermann, U. F. Ekpo, A. Garba, et al. "Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis". In: *The Lancet Infectious Diseases* 15.1 (2015), pp. 74–84.

[53] R. Wardell, A. C. Clements, A. Lal, D. Summers, et al. "An environmental assessment and risk map of Ascaris lumbricoides and Necator americanus distributions in Manufahi District, Timor-Leste". In: *PLoS Neglected Tropical Diseases* 11.5 (2017), e0005565.

Chapter 6

Paper 4: Disentangling Outbreak Patterns of Dengue Fever in Nepal: A District-Level Analysis from 2006 to 2022

Jessie J. Khaki [1,2,3], Bipin K. Acharya[4], Basu D. Pandey[5,6], Kouichi Morita[6], Emanuele Giorgi[1].

[1] Centre for Health Informatics, Computing and Statistics, Lancaster University, Lancaster, United Kingdom.

[2] Malawi Liverpool Wellcome Programme, Blantyre, Malawi.

[3] School of Global and Public Health, Kamuzu University of Health Sciences, Blantyre, Malawi.

[4] School of Public Health, Sun Yat Sen University, Guangzhou, China.

[5] Everest International Clinic and Research Center, Kathmandu, Nepal.

[6] DEJIMA Infectious Disease Research Alliances, Nagasaki University, Japan.

Summary

Dengue is the fastest-growing mosquito-borne disease globally and one of the top ten threats to global health, as declared by the World Health Organization. In Nepal, dengue cases have surged since the first recorded case in 2004, with seasonal outbreaks occurring every two to three years. This study aimed to identify the three major outbreaks within each Nepalese district and analyze their duration.

We applied modified Negative Binomial models to district-level data collected from Nepalese health facilities between 2006 and 2022, validating the models using the Chi-square goodness of fit test. Our results showed significant variation in outbreak occurrence across districts.

Notably, 65% of districts experienced their second outbreak in 2019, and 95% faced their third outbreak in 2022. In 79% of districts, the third outbreak contributed the most to the overall dengue burden. Outbreak durations also varied across districts.

Tailored strategies are essential for preventing and controlling dengue transmission at the district level. The proposed modelling framework is flexible and can be applied to other diseases, and can also be extended to include more than three outbreaks.

Keywords: dengue, multiple Outbreaks, Nepal, Negative Binomial model, peak, duration

6.1 Introduction

Dengue is the most rapidly growing mosquito-borne disease in the world and can be caused by four viruses (DENV): DENV-1, DENV-2, DENV-3, and DENV-4 [1, 2]. The number of dengue infections has been steadily increasing in the past 30 years [3]. It is estimated that there are more than 100 million dengue cases yearly, about 20,000 of which result in death [1, 4]. Recent reports show that the death rate due to dengue has increased by almost 70%, from 0.31 per 100,000 population in 1990 to 0.53 per 100,000 population in 2017 [3]. The disability-adjusted life years (DALYs) due to dengue have also increased more than a hundred-fold [3, 5]. Furthermore, recent studies estimate that the potential number of people at risk of dengue could rise by an additional 4.7 billion individuals by 2070 due to climate change [2, 6]. Due to its high burden level, the World Health Organization (WHO) has declared dengue one of the top ten threats to global health and upgraded to highest threat level 3 [7]. Dengue is endemic in more than 100 countries, with the highest cases occurring in tropical and subtropical regions such as Caribbeans, South Asia, Southeast Asia, and Latin America [1, 5]. Approximately 70% of the burden of dengue is estimated to be in Asia [8].

Urbanization, climate change, and co-circulation of multiple DENV serotypes are hypothesized to be part of the variables driving the dengue epidemic in Nepal and elsewhere [5, 9]. Previous studies have shown that the density of *Aedes aegypti* mosquitoes, which transmit dengue viruses (DENV 1-4), is highly influenced by environmental and climatic variables such as rainfall, temperature, and precipitation [9–11]. An increase in climatic variables such as temperature and precipitation is associated with an increase in the incidence of dengue [11–14]. However, other studies in Asian countries such as Thailand found that the relationship between dengue and climatic variables such as precipitation and rainfall is complex [15]. For instance, although precipitation can create a breeding ground for mosquitoes, excessive rainfall can wash away their breeding sites and thus reduce dengue incidence [11, 16, 17]. In addition to varying over time, the incidence of dengue has been shown to vary geographically [10, 12, 18]. Furthermore, studies in other countries such as China have shown notable variations in the peaks and intensity of dengue outbreaks [19].

In Nepal, the first case of dengue was recorded in 2004 [20, 21]. Since then, the number of dengue cases in Nepal has increased rapidly from around 32 cases in 9 districts in 2006 to over 50,000 cases in 2022, with all four DENV serotypes co-circulating in the country [9, 22–26]. Although dengue has become endemic in Nepal, cyclic outbreaks of dengue cases are observed every two to three years [22, 26, 27]. Nepal's first national dengue prevention, control, and management guidelines were developed in 2008 and revised in 2011 and 2019 [28]. Despite the successful implementation of dengue control measures in Nepal, dengue remains a public health concern.

Describing the dynamics of disease outbreaks is essential for investigating the underlying factors that drive them, enhancing the precision of outbreak predictions, and devising effective strategies to manage and mitigate the outbreaks [29]. Two key characteristics of disease outbreaks are the peak in the number of infected cases and the duration of the outbreak. In this paper, we define the "duration of an outbreak" as the length of time from the onset of the outbreak to the point when the number of new cases returns to baseline or significantly decreases. In other words, the duration of the outbreak is the period that includes the rise, peak, and decline of the outbreak, encompassing all phases of the epidemic curve. Understanding these characteristics is especially important for formulating timely response measures, allocating resources efficiently, and implementing appropriate control strategies.

Existing modelling methods used to describe these aspects of disease outbreaks have focused mainly on the use of compartmental models such as the Susceptible-Exposed- Infectious- Removed (SEIR) and Susceptible- Infectious- Quarantined-Recovered (SIQR) [19, 30, 31]. However, these models have inherent limitations, as they do not enable the inclusion of spatial risk factors unless the fitting is done by stratifying for the covariates of interest [32], and cannot easily accommodate the occurrence of multiple outbreaks. Other studies have employed epidemic-endemic models such as the modified Poisson and Negative Binomial to describe outbreaks of diseases such as COVID-19 and measles [33, 34]. Prior research on the

characterization dengue outbreaks has focused on the rate at which individuals susceptible to dengue become infected, a parameter also referred to as the force of infection [35, 36]. Whilst this parameter affects the peak and duration of an outbreak, these two characteristics are not explicitly modelled and explained in epidemic-endemic models. Moreover, existing epidemic models do not allow for multiple outbreaks to be estimated in a single model.

In this study, we use a data-driven approach to extend the standard class of Negative Binomial regression models to estimate the peaks and duration of dengue fever outbreaks at the district level in Nepal using yearly reported cases. Our proposed modelling framework allows for the estimation of multiple outbreaks through the specification of an outbreak function, for which we consider three different specifications. In our modelling framework, we interpret the outbreak function as the unexpected rise in the reported cases that cannot be explained by the covariates used in the regression. The Negative Binomial models are validated using the Chi-square goodness of fit test [37].

The structure of the paper is as follows. In Section 6.2, we describe the data and outline the modelling framework for characterising the dengue outbreak in Nepal. In Section 6.3, we illustrate the results from the descriptive analysis. We also illustrate the timing and duration of 3 outbreaks within each district. The discussion and conclusions of the study are presented in Section 6.4.

6.2 Methods and materials

6.2.1 Study site and Data collection

Nepal, a country of approximately 147,200 km^2 [38], borders India and China and is located at a longitude of 84.12° and a latitude of 28.39°. The elevation in Nepal ranges from 60 meters (m) to 8,848 m above sea level [38].

Topographically, Nepal is divided into 11 main ecological zones, with the most prominent being the Terai region (67 m to 300 m above sea level), the Siwalik Hills (700 m to 1,500 m), the Mahabharat region (1,500 m to 2,700 m), and the

Himalayan zone (above 4,000 m) [39]. The Nepal climate, therefore, varies significantly due to the wide variations in altitude [38–40]. Approximately 80% of Nepal's precipitation occurs from June to September in the summer monsoon season. The average annual rainfall in Nepal also varies widely and ranges from 295 millimetres (mm) to 3,345 mm, depending on the ecological zone [39]. Furthermore, the average temperature decreases by 6 ^{o}C for every 1,000 meters increase in the altitude [39].

Nepal is administratively divided into 7 provinces and 77 districts. We retrieved district-level annual dengue fever case data reported between 2006 and 2022 from the Epidemiology and Diseases Control Division (EDCD), Department of Health services, Ministry of Health and Population, Nepal responsible for outbreak preparedness and response in Nepal (https://www.edcd.gov.np/section/dengue-control-program). The EDCD is also responsible for the prevention, surveillance, and control of communicable and non-communicable diseases diseases in Nepal. Details on how the data is collated from the facility level (such as district hospitals and health centers) to a central location at the EDCD have been published elsewhere [22, 41]. Briefly, the official EDCD reports publish the district and year of every dengue case confirmed through laboratory tests. The annual number of dengue cases per district was defined as the sum of confirmed cases within a particular district. Figure 6.1 shows the study location.

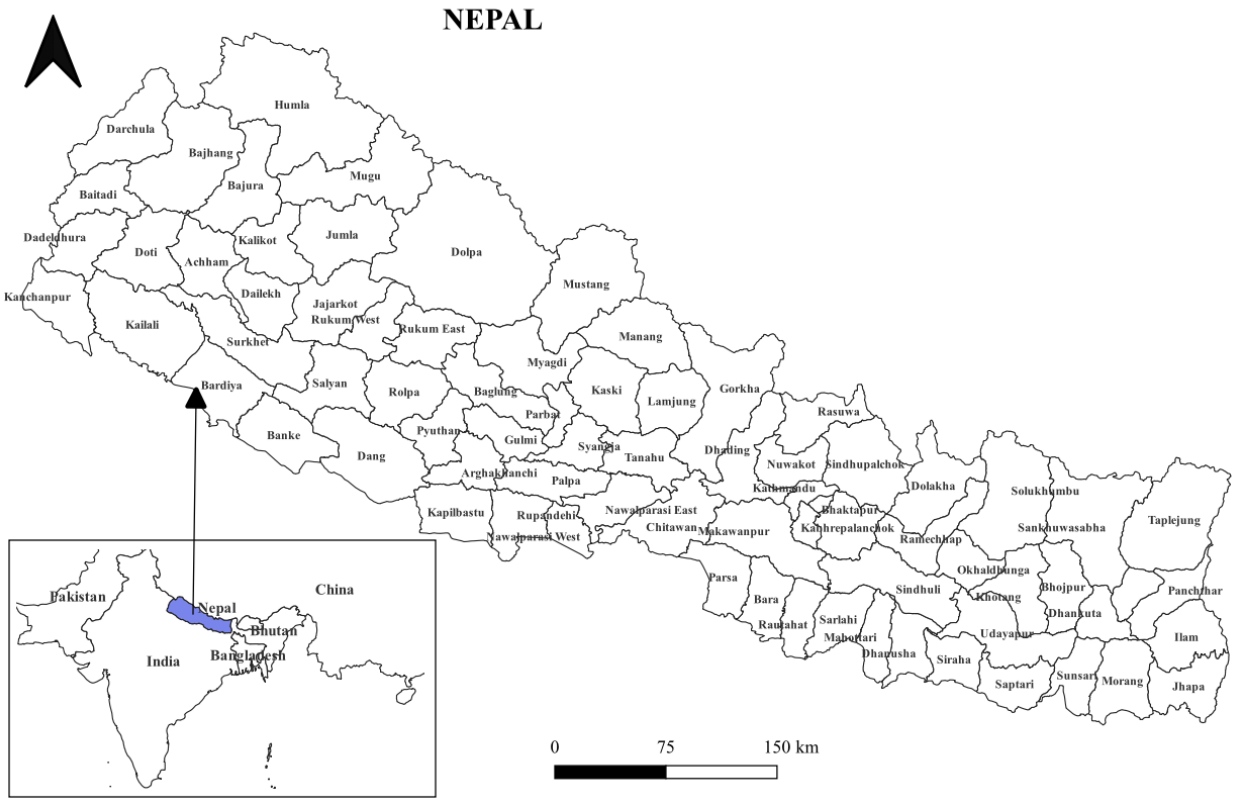

Figure 6.1: Map of Nepal showing the location of Nepal and the boundaries (black lines) of the 77 districts.

6.2.2 Spatio-temporal covariates, covariate processing and population data

We extracted a list of spatio-temporal environmental variables for which a significant association with dengue fever risk has been previously reported [9–12]. Table 6.1 outlines the list of environmental variables, their sources, and temporal resolutions.

Table 6.1: List of environmental covariates and their sources

Covariate	**Temporal Resolution**	**Source of Data**
Minimum, mean & maximum Precipitation	2006-2022	TerraClimate [42]
Minimum & maximum temperature	2006-2022	TerraClimate [42]
Potential evapotranspiration (PET)	2006-2022	TerraClimate [42]
Aridity index	2006-2022	Mean precipitation ÷ PET

Following previous modelling studies, we explored the relationships between the covariates and dengue fever incidence using scatter plots to assess the strength of association [43–45]. We combined the covariates using principal components analysis (PCA) to develop an index of environmental exposure. The PCA was carried out to reduce the number of covariates in our final model to avoid overfitting due to the small sample size per district. Appendix D presents the exploratory analysis and the PCA results. The variables were standardized prior to carrying out the PCA. Standardizing variables before PCA ensures that all variables are on the same scale, preventing those with larger variances from disproportionately influencing the principal components [46, 47]. The first component, accounting for 48% of the variation in the PCA, was selected as our environmental exposure index (Figure 6.4).

We also obtained the yearly population counts per district from 2006 to 2020 from the Worldpop website [48]. The 2021 and 2022 population counts were derived by scaling the 2020 population using the Nepal population growth rate (0.92 per cent) reported in the 2021 Nepal National Population and Housing Census [49].

6.2.3 Statistical modelling

Let Y_t denote the yearly dengue case count at the district level in year t (t = 1 (2006), 2,3, . . . ,17 (2022)). To account for the overdispersion, we then assume that Y_t follows a Negative Binomial (NB) distribution, i.e. $Y_t \sim \text{NB}(\lambda_t, \alpha)$, where λ_t denotes the annual dengue incidence, and α is the dispersion parameter. Let θ denote the vector of unknown model parameters; the likelihood function is then defined as [50]:

$$L(\theta) = \prod_{t=1}^{17} p\left(y_t\right) = \prod_{t=1}^{17} \frac{\Gamma\left(y_t + 1/\alpha\right)}{\Gamma\left(y_t + 1\right)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\lambda_t}\right)^{1/\alpha} \left(\frac{\alpha\lambda_t}{1+\alpha\lambda_t}\right)^{y_t} \tag{6.1}$$

where λ_t is the mean number of dengue cases at time t and was defined as follows in Table 6.2. To characterize the outbreaks in Nepal and by assuming that there were 3 outbreaks within each district, we extend the Negative Binomial regression model (equation 6.1) as follows. The assumption of 3 main outbreaks within each district was based on expert recommendations and preliminary evaluation of annual incidence trends per district.

Table 6.2: Specification of parameters used in characterizing the dengue outbreak in Nepal

Model	**Outbreak parameter specification**	
1	$\lambda_t = m_t . \exp\left(d'\beta\right) + \sum_{i=1}^{3} \gamma_i f_{(i,t)}$	(6.2)
2	$\lambda_t = m_t . \exp\left(d'\beta + \sum_{i=1}^{3} \gamma_i f_{(i,t)}\right)$	(6.3)
3	$\lambda_t = m_t . \exp\left(d'\beta\right) \times \left(1 + \sum_{i=1}^{3} \gamma_i f_{(i,t)}\right)$	(6.4)

In equations 6.2, 6.3, and 6.4, the population count (m_t) is used as an offset in the model, $\boldsymbol{d}_t = (1, d_{t_1})$ is the vector of the intercept and the environmental exposure index, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the coefficient vector associated with the intercept and environmental exposure index covariate. Additionally, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)'$ is a coefficient vector for the three outbreaks, and f_t is a function used to quantify the *intensity* of the outbreak in the district being considered for the analysis. Henceforth, we refer to f_t as the outbreak intensity function (OIF). In this setting, we used the squared exponential function to define each OIF for each district at time t as follows.

$$f_t = \exp\left(-\frac{(t-\mu)^2}{2\omega^2}\right), \tag{6.5}$$

where μ and ω denote the peak and the scale parameter of the OIF.

In the three models considered in Table 6.2, the γ parameters reflect the contribution of the corresponding OIF to the mean dengue incidence, λ_t. In Model 1 (equation 6.2), the effect of γ_i $f_{(i,t)}$ is additive, meaning the OIF quantifies the excess number of cases that are not due to covariates effects.

In Models 2 (equation 6.3) and 3 (equation 6.4), the effect is multiplicative. In Model 2, $\gamma_i f_{(i,t)}$ is incorporated into the linear predictor and its effect can be interpreted in the same way as the other covariates, i.e. by using $\exp\{\gamma_i\}$ to

quantify the multiplicative increase in the number of cases due to the OIF. In Model 3 (equation 6.4), the term $1 + \sum_{i=1}^{3} \gamma_i f_{(i,t)}$ scales the overall mean dengue incidence. Here, $\gamma_i f_{(i,t)}$ acts as a proportion of the baseline incidence $m_t . \exp(d'\beta)$, adjusting the number of cases relative to the baseline level, where larger values of $\gamma_i f_{(i,t)}$ represent proportionally larger outbreaks.

We fitted the Negative Binomial outbreak model to each of the districts separately due to varying dengue fever incidence within each district.

Confidence intervals for parameters $\{\theta = (\beta_1, \beta_2, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\omega})\}$ estimated from the Negative Binomial models were constructed using a bootstrap procedure by following the steps below [51]:

- Construct the first bootstrap sample by selecting a random sample of 17 (N) observations with replacement from our original sample.
- Fit the model to the first bootstrap sample and store the model coefficients.
- Repeat steps 1 and 2 10,000 (K) times, drawing new random samples of size N with replacement each time to create subsequent bootstrap samples.
- Compute the confidence intervals using the Bias-Corrected and Accelerated (BCa) method, which utilizes the distribution of estimates obtained from the K bootstrap samples as the sampling distribution.

We assessed our models' goodness of fit (gof) using the Chi-square gof test [37]. The test, which has the null hypothesis that there is no significant difference between the observed and the expected dengue counts, is given as:

$$\sum_{t=1}^{T} \frac{(O_t - E_t)^2}{E_t} \tag{6.6}$$

where O_t are the observed dengue fever cases at time t and E_t are the predicted counts of dengue from the model.

6.3 Results

6.3.1 Descriptive analysis

A total of 78,819 dengue cases were reported over the 17 years in all 77 districts. Figure 6.2 shows the total number of dengue cases in each district over 17 years (2006 to 2022). As can be seen, the highest total number of cases was observed in the central areas of Nepal. In particular, the districts with the observed dengue cases of at least 2,000 over the study period were Kathmandu (15,940), Lalitpur (10,220), Chitawan (8,845), Makawanpur (8,237), Bhaktapur (6,515), Rupandehi (3,695), Kaski (3,654), Dang (2,585), and Jhapa (2,008).

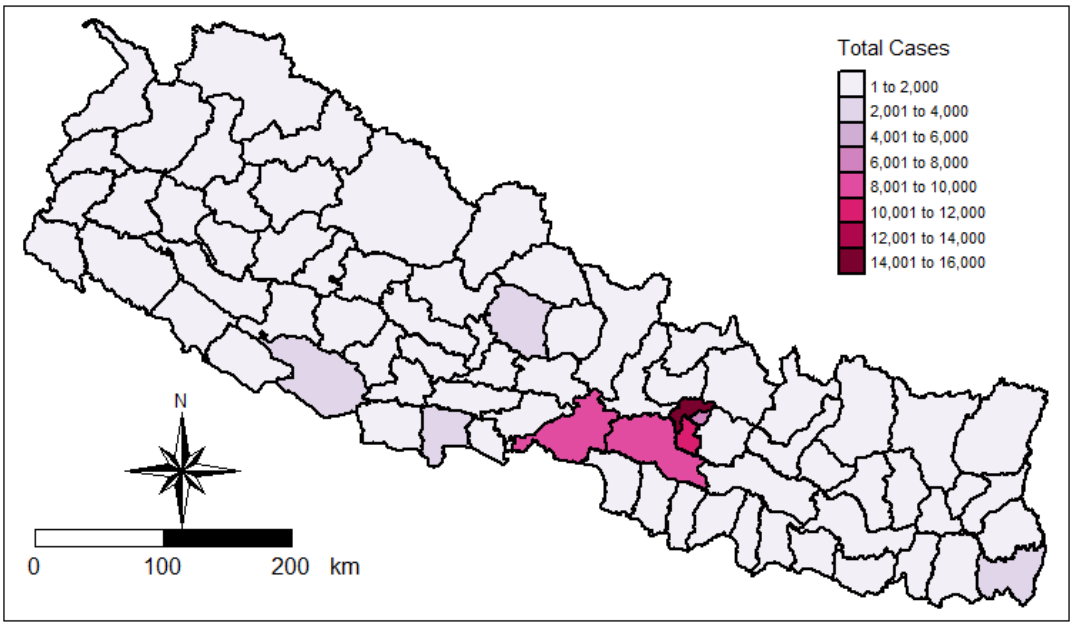

Figure 6.2: A map illustrating the total number of cases observed in each district over the study period.

In general, there was a rise in dengue fever incidence per 100,000 population in each district over time, as illustrated in Figure 6.3. A discernible pattern emerged, with notable peaks in 2010, 2013, 2016, 2019, and 2022. Notably, the year 2022 exhibited the highest incidence across most districts. Nevertheless, Figure 6.3 also highlights that the districts did not experience peaks in the same years. Our modelling framework accounts for the variations in peak occurrences by explicitly incorporating outbreak parameters such as peak and duration, in each district.

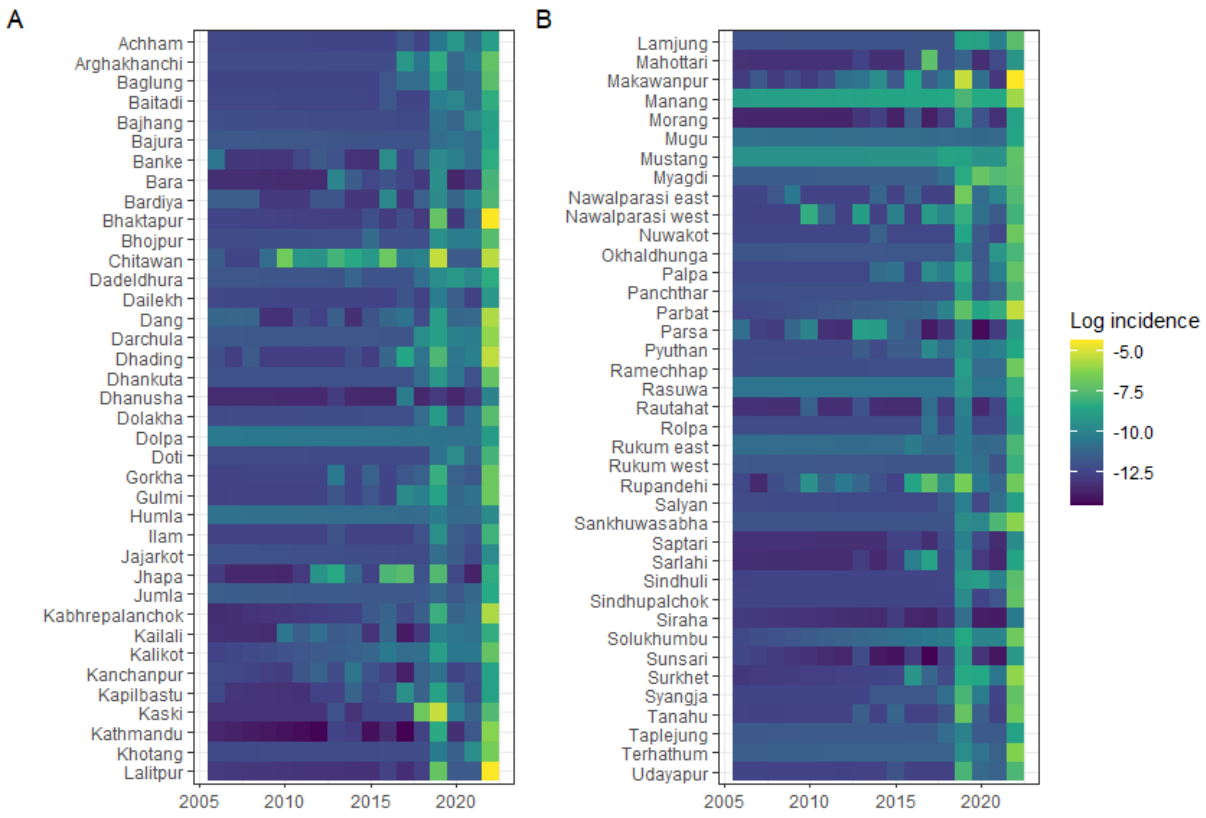

Figure 6.3: Heat maps illustrating the observed log-transformed cases of dengue fever per 100,000 population (log((count of dengue cases + 1) / population)) for each district from 2006 to 2022. The transformation adds 1 to the case counts before taking the logarithm to handle zero values.

6.3.2 Principal Components Analysis (PCA) results

An index of environmental exposure was derived from the above covariates using Principal Components Analysis (PCA). Figure 6.4 shows the variation percentage explained by the PCA components. Our environmental exposure score was based on the first principal component, which explained 48% of the variation in our data, due to previous epidemiological studies using the first component to derive an index or score [52, 53].

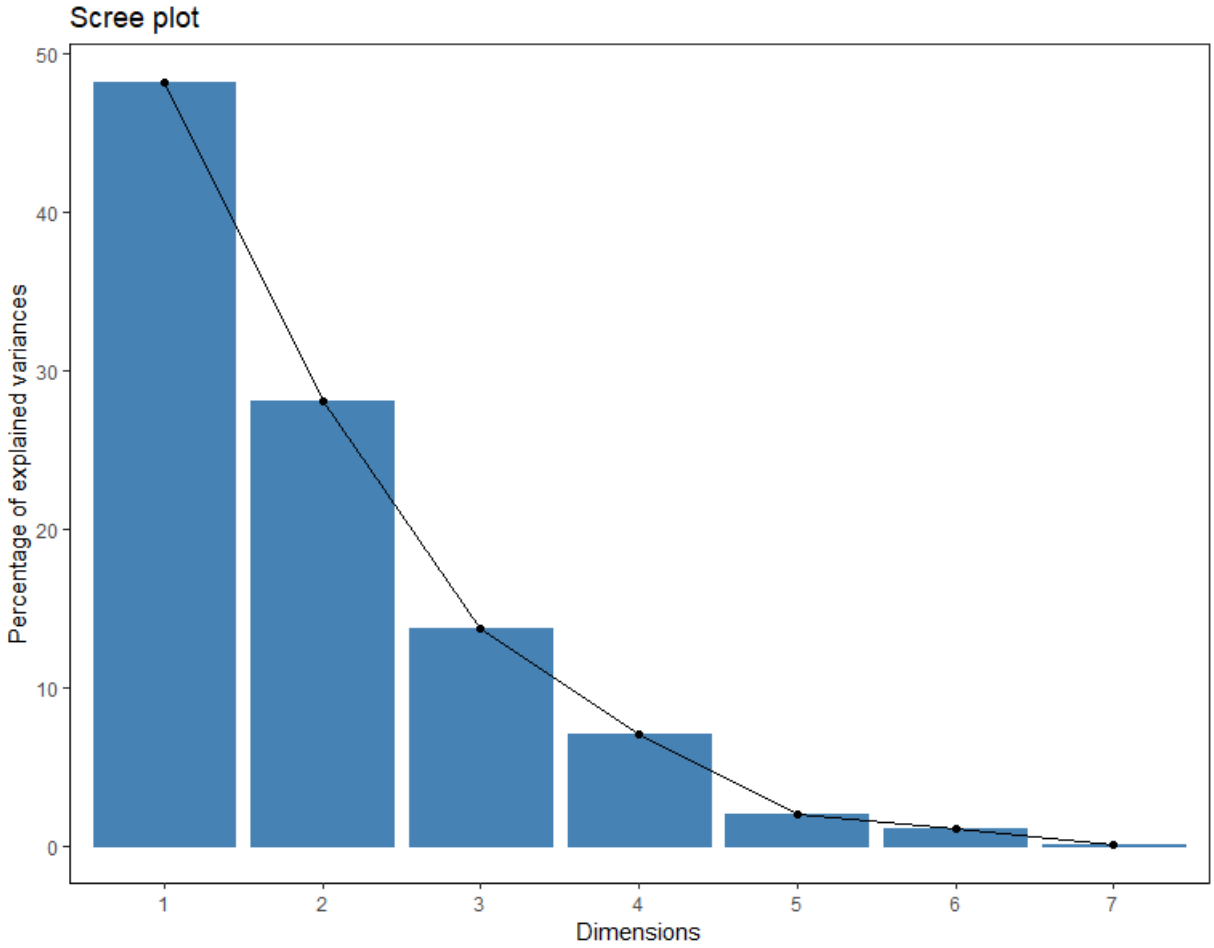

Figure 6.4: Eigen values illustrating the variance percentage explained by each component

Figure D.2 in Appendix D illustrates the loadings for Principal Component 1 (PC1). The most significant contributions come from mean precipitation (-0.48), the aridity index (-0.43), and maximum precipitation (-0.40). The weights in Figure D.2, therefore, indicate that the first principal component primarily captures variations related to aridity index and precipitation, suggesting that these factors are dominant in the environmental exposure index used in our modelling. Consequently, due to the negative loadings, an increase in the environmental exposure index (PC1) is associated with a reduction in covariates, such as precipitation, indicating that areas with high precipitation correspond to those with a lower exposure index (see Figure D.3 in Appendix D).

6.3.3 Model selection

We fitted the models in equations 6.2, 6.3 and 6.4 with and without an environmental exposure index covariate, and compared the models using the Akaike Information

Criteria (AIC). Among all the three models, those with the covariate provided the lowest AIC. Furthermore, equation 6.3 with the environmental exposure index as a covariate had the lowest AIC across all the 6 models. Therefore, the results in this paper are based on model 6.3.

Table D.2 in the supplementary information provides the estimates and confidence intervals of the environmental exposure index. Overall, an increase in the environmental exposure index was associated with an increase in the incidence of dengue in 36 districts and a decrease in 41 districts. However, the results for 82% (n=63) districts were not statistically significant as the confidence intervals for the coefficients span the null hypothesis value of zero.

6.3.4 Characterizing the dengue outbreaks in Nepal

Figure 6.5 shows the outbreak years in each district after adjusting for the effect of the environmental exposure index. The majority of the districts experienced their first outbreak (μ_1) in 2017 (16%), the second outbreak (μ_2) in 2019 (65%) and the third outbreak (μ_3) in 2022 (95%).

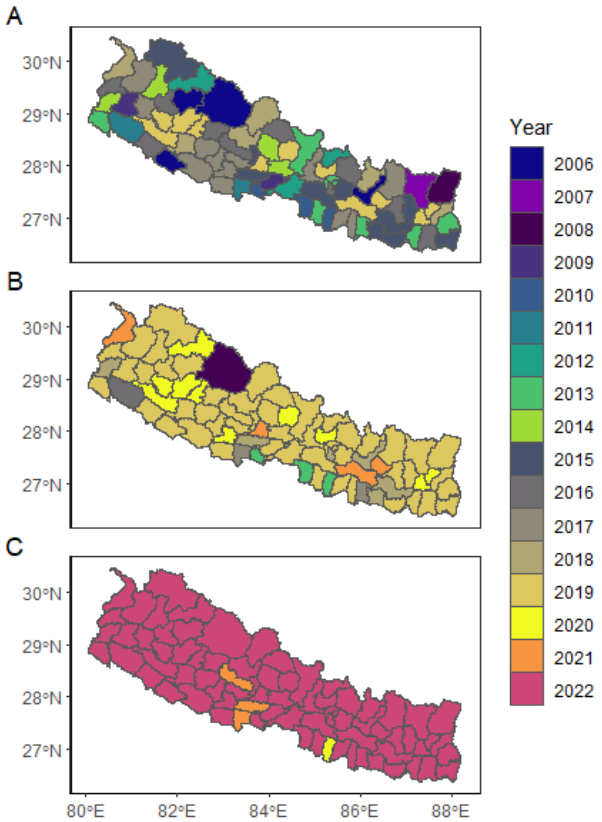

Figure 6.5: Maps showing the years that the districts had outbreak 1 (A = μ_1), outbreak 2 (B = μ_2) and outbreak 3 (C = μ_3)

The estimate of the scale parameter (the duration) of the 3 outbreak intensity functions also varied within each district. Figure 6.6 illustrates the duration of three distinct outbreaks across the 77 Nepalese districts. For the first outbreak, 82% of the districts experienced it for less than a year, while 18% had it for a year or more. The second outbreak affected 71% of the districts for less than a year and 29% for more than a year. For the third outbreak, 73% of the districts experienced it for less than a year, with 8% affected for more than 2 years.

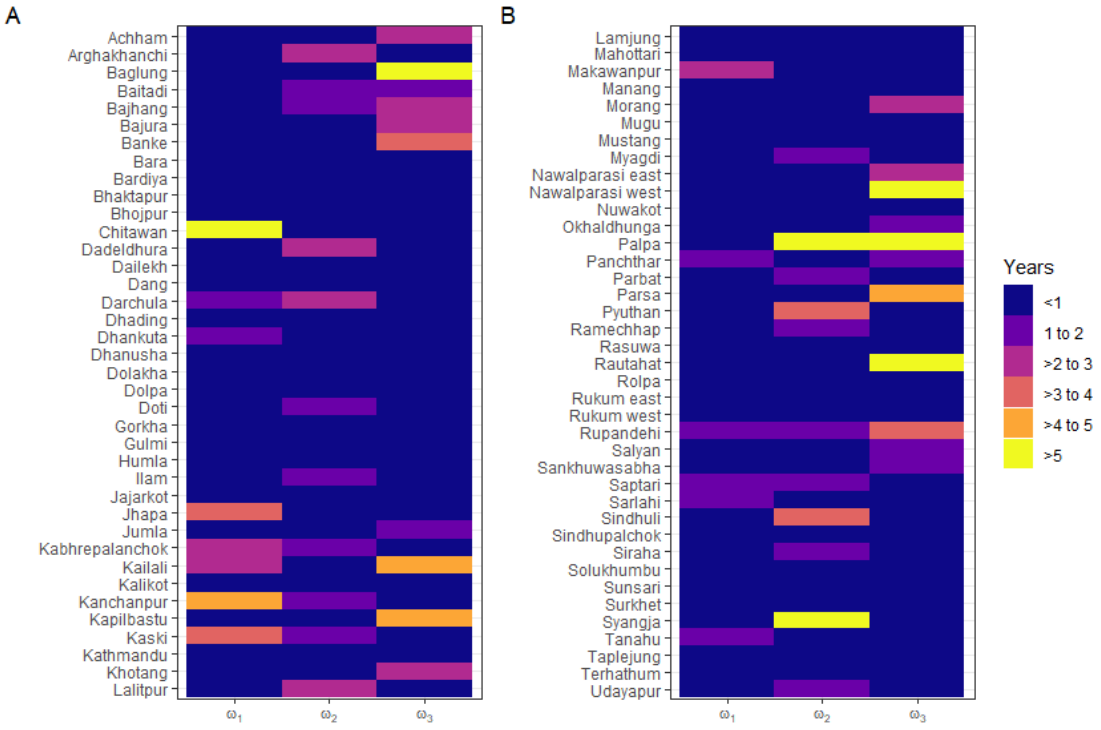

Figure 6.6: Heat maps showing the duration (ω's) of each outbreak in each district

The largest outbreak coefficient for 93% (n=72) of the districts was associated with the third outbreak (γ_3), followed by the second (γ_2), and then the first (γ_1). Figure 6.7 illustrates the approximated contribution of each outbreak to the overall dengue epidemic in Nepal. Overall, the two highest contributions to the dengue epidemic in the 72 districts were from the second (14%, n = 11) and third outbreaks (79%, n = 61). This result should, however, be cautiously interpreted as we did not constrain the outbreak coefficients (γ's in equation 6.3) in our study.

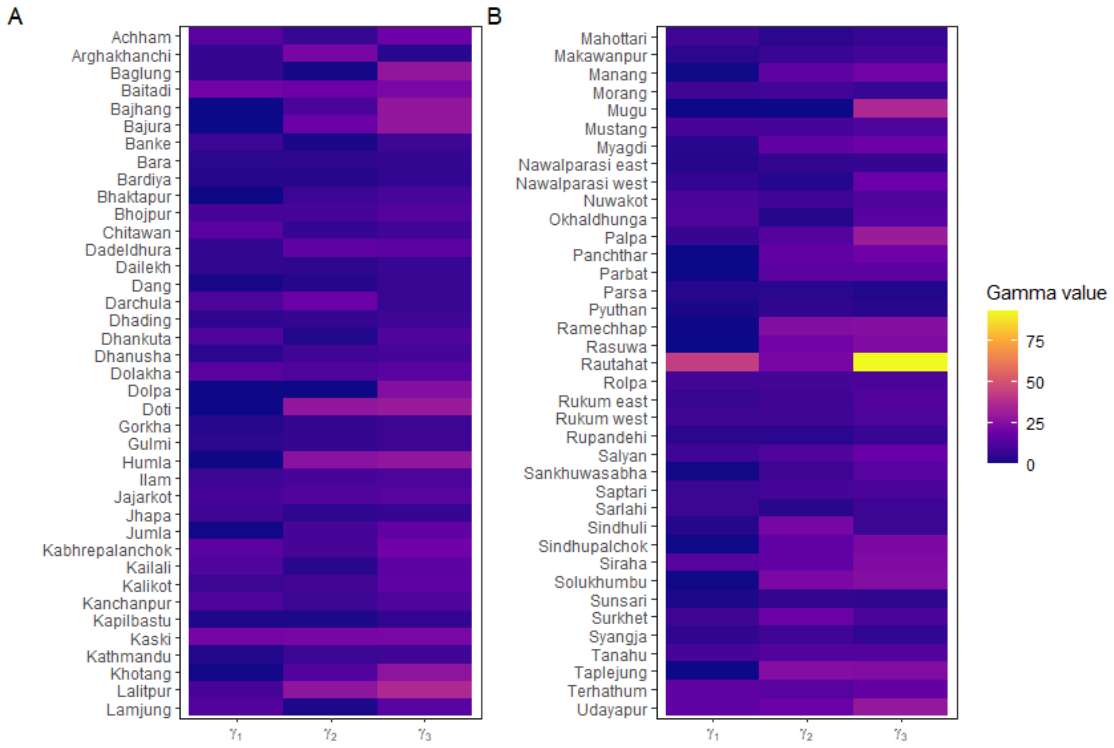

Figure 6.7: Heat maps showing the estimated coefficients (γ's) for each outbreak

The outbreak parameters (μ, ω and γ) in our models reveal distinct patterns of dengue incidence across Nepal's districts. Districts such as Okhaldunga, Sindhuli and Syangja experienced more recent outbreaks as characterized by higher values of μ_1 and μ_2 associated with later time functions $f_{(i,t)}$. Furthermore, higher values of the outbreak coefficients (γ) in these districts indicate that the dengue epidemic has intensified in recent years. Conversely, districts such as Chitawan and Kailali experienced more longstanding outbreaks, as evidenced by the high values of the scale parameter of the outbreak intensity function (ω) across multiple outbreaks, suggesting a prolonged epidemic with sustained cases over time. These results imply a heterogeneity of dengue dynamics across Nepal, with some districts facing newer outbreaks while others have endured more extensive and persistent epidemics over the study period.

6.3.5 Model validation

The model validation results are presented in Table D.3 in Appendix D. In this analysis, 87% (n=67) of the district models showed a satisfactory fit to the observed data. However, the model did not perform well for ten districts, all of

which had experienced more than three outbreaks. To address this, we extended model 6.3 to incorporate four outbreaks, leading to satisfactory fits for Banke, Kailali, and Rautahat. For Jhapa, Kapilbastu, Makawanpur, Nawalparasi West, and Parsa, incorporating five outbreaks produced good fits. The remaining two districts, Chitawan and Rupandehi, showed an improved fit based on the Akaike Information Criterion (AIC) as more outbreaks were included. Despite these AIC improvements, we chose not to extend the models further to avoid over-parameterization.

6.4 Discussion and Conclusion

In this study, we analyzed dengue outbreaks in Nepal, and extended the Negative Binomial regression model to allow for the inclusion of an outbreak intensity function (OIF) with parameters that directly express the timing and scale parameter of the OIF. To our knowledge, this is the first study to characterise dengue in Nepal using a model-based approach. Additionally, the proposed modelling framework effectively accommodates the occurrence of multiple outbreaks within a district. While models such as HHH4 and HHH4ZI are valuable for decomposing outbreak data into endemic and epidemic components [54, 55], they do not focus on estimating the scale of multiple outbreaks. Similarly, the approach by Anderson et al. identifies clusters in space but does not estimate their timing [56]. Additionally, previous approaches, such as the one used by Guzman et al., require data to be aggregated into spatio-temporal blocks (e.g., at 1 week, 3 weeks, and 5 weeks) to investigate outbreaks within those blocks [57]. In contrast, our model sought to contribute to the disease outbreak modelling body of knowledge by providing insights into the duration and size of each outbreak. Importantly, our model does not require aggregation of data into blocks, and it also offers a further understanding of dengue outbreak dynamics by estimating both the scale parameter of each OIF and approximating the contribution of each outbreak to the overall dengue epidemic.

Our results reveal that the association between dengue incidence and the environmental exposure index varies across districts. This is consistent with previous studies that have shown both positive and negative associations between environmental factors like precipitation and dengue incidence across different

regions [11, 13, 15–17]. This variability highlights the importance of district-specific models.

Another key finding is the occurrence of multiple outbreaks over time within a single district. Similar to the environmental exposure index, the peak and scale parameters of each outbreak intensity function varied across districts. An important extension of this work would be to introduce covariates that could account for the heterogeneity in the peak and scale parameters of each outbreak across districts. Incorporating district-level variables, such as nighttime light data as a proxy for urbanization, could help explain the observed differences in outbreak dynamics. These covariates would enable us to model how factors such as urbanization, human mobility, and climate conditions influence the timing and size of outbreaks. However, we did not pursue this approach here due to the limited amount of data available, which constrains our ability to estimate the effects of these variables. With information at higher temporal resolution, future research could expand on the work carried out in this study by incorporating additional data to better understand the drivers of outbreak variation across regions.

A further extension of the model proposed in this work would be to consider curves for the intensity of the outbreak that are not symmetric, such as the skew-Normal distribution. However, fitting such a model to the current data would be difficult because we only had yearly data and there were too few cases. This study did not find strong evidence of residual spatial correlation. However, in scenarios where spatial correlation is present, an extension of our proposed model could incorporate a Besag-York-Mollié (BYM) component.

The fact that some districts experienced outbreaks for which we estimated a scale parameter of the outbreak intensity function (OIF) of about five years suggests the presence of persistent dengue transmission in those areas. This persistence highlights the need for enhanced surveillance and targeted intervention strategies to reduce ongoing transmission and prevent future outbreaks.

This study has several limitations. First, the reliance on passive surveillance data from the Ministry of Health means that underreporting and misreporting of symptomatic cases is likely [9, 58, 59]. Another key limitation was the absence of certain risk variables, particularly urbanization levels, which are strongly associated with the density of *Aedes aegypti* mosquitoes and dengue transmission

[3, 60]. Additionally, we could not investigate the effects of lagged climate variables, such as delayed impacts of precipitation, due to the lack of monthly data. Future studies should explore the relationship between lagged environmental factors and dengue incidence, as suggested by other research [61, 62]. Moreover, future models could be extended to include outbreak-weighting parameters, allowing for the classification of outbreaks based on intensity relative to others.

We developed a modelling framework that identifies three main outbreaks in each district in Nepal. Applied to dengue data, this model improves upon previous outbreak modelling techniques by accounting for multiple outbreaks within the study period and describing the duration of each of the outbreaks. The results of this study could assist the Nepalese government in identifying districts with synchronized outbreaks and inform targeted intervention strategies. Additionally, our model may be applicable to other diseases, offering insight into their historical outbreak patterns. The proposed framework could thus be used to disentangle outbreaks of various infectious diseases and contribute to better public health responses.

Data availability

All the sources for the data used in this study have been cited in the main manuscript. The R code used to run the models in this study can be accessed on *Github*.

References

[1] X. Yang, M. B. Quam, T. Zhang, and S. Sang. "Global burden for dengue and the evolving pattern in the past 30 years". In: *Journal of travel medicine* 28.8 (2021), taab146.

[2] F. J. Colón-González, M. O. Sewe, A. M. Tompkins, H. Sjödin, et al. "Projecting the risk of mosquito-borne diseases in a warmer and more populated world: a multi-model, multi-scenario intercomparison modelling study". In: *The Lancet Planetary Health* 5.7 (2021), e404–e414.

[3] Z. Zeng, J. Zhan, L. Chen, H. Chen, et al. "Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017". In: *EClinicalMedicine* 32 (2021).

[4] Z. Zeng, J. Zhan, L. Chen, H. Chen, et al. "Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017". In: *EClinicalMedicine* 32 (2021), p. 100712.

[5] Gathsaurie, P. Neelika Malavige, K. Sjö, J.-M. Singh, et al. "Facing the escalating burden of dengue: Challenges and perspectives". In: *PLos Global Health* 3.12 (2023), e0002598.

[6] J. P. Messina, O. J. Brady, N. Golding, M. U. Kraemer, et al. "The current and future global distribution and population at risk of dengue". In: *Nature microbiology* 4.9 (2019), pp. 1508–1515.

[7] W. H. O. (WHO). *Ten threats to global health in 2019*. Accessed June 2023. 2019. URL: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

[8] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, et al. "The global distribution and burden of dengue". In: *Nature* 496.7446 (2013), pp. 504–507.

[9] B. D. Pandey and A. Costello. "The dengue epidemic and climate change in Nepal". In: *The Lancet* 394.10215 (2019), pp. 2150–2151.

[10] L. Kong, C. Xu, P. Mu, J. Li, et al. "Risk factors spatial-temporal detection for dengue fever in Guangzhou". In: *Epidemiology & Infection* 147 (2019).

[11] T. Phanitchat, B. Zhao, U. Haque, C. Pientong, et al. "Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016". In: *BMC infectious diseases* 19 (2019), pp. 1–12.

[12] W. Sun, L. Xue, and X. Xie. "Spatial-temporal distribution of dengue and climate characteristics for two clusters in Sri Lanka from 2012 to 2016". In: *Scientific reports* 7.1 (2017), p. 12884.

[13] F. I. Abdulsalam, P. Antunez, S. Yimthiang, and W. Jawjit. "Influence of climate variables on dengue fever occurrence in the southern region of Thailand". In: *PLOS Global Public Health* 2.4 (2022), e0000188.

[14] M. Dhimal, I. Gautam, A. Kreß, R. Müller, et al. "Spatio-temporal distribution of dengue and lymphatic filariasis vectors along an altitudinal transect in Central Nepal". In: *PLoS Neglected Tropical Diseases* 8.7 (2014), e3035.

[15] K. M. Campbell, C. Lin, S. Iamsirithaworn, and T. W. Scott. "The complex relationship between weather and dengue virus transmission in Thailand". In: *The American journal of tropical medicine and hygiene* 89.6 (2013), p. 1066.

[16] C. Li, T. Lim, L. Han, and R. Fang. "Rainfall, abundance of Aedes aegypti and dengue infection in Selangor, Malaysia." In: *The Southeast Asian journal of tropical medicine and public health* 16.4 (1985), pp. 560–568.

[17] H. Rozilawati, J. Zairi, C. Adanan, et al. "Seasonal abundance of Aedes albopictus in selected urban and suburban areas in Penang, Malaysia". In: *Trop Biomed* 24.1 (2007), pp. 83–94.

[18] R. F. do Carmo, J. V. J. Silva Júnior, A. F. Pastor, and C. D. F. de Souza. "Spatiotemporal dynamics, risk areas and social determinants of dengue in Northeastern Brazil, 2014–2017: an ecological study". In: *Infectious diseases of poverty* 9 (2020), pp. 1–16.

[19] Y. Chen, T. Liu, X. Yu, Q. Zeng, et al. "An ensemble forecast system for tracking dynamics of dengue outbreaks and its validation in China". In: *PLoS computational biology* 18.6 (2022), e1010218.

[20] B. K. Acharya, C. Cao, T. Lakes, W. Chen, et al. "Spatiotemporal analysis of dengue fever in Nepal from 2010 to 2014". In: *BMC Public Health* 16.1 (2016), pp. 1–10.

[21] B. D. Pandey, S. K. Rai, K. Morita, and I. Kurane. "First case of Dengue virus infection in Nepal." In: *Nepal Medical College Journal: NMCJ* 6.2 (2004), pp. 157–159.

[22] K. R. Rijal, B. Adhikari, B. Ghimire, B. Dhungel, et al. "Epidemiology of dengue virus infections in Nepal, 2006–2019". In: *Infectious diseases of poverty* 10.1 (2021), pp. 1–10.

[23] B. D. Pandey, T. Nabeshima, K. Pandey, S. P. Rajendra, et al. "First isolation of dengue virus from the 2010 epidemic in Nepal". In: *Tropical Medicine and Health* 41.3 (2013), pp. 103–111.

[24] S. P. Dumre, R. Bhandari, G. Shakya, S. K. Shrestha, et al. "Dengue virus serotypes 1 and 2 responsible for major dengue outbreaks in Nepal: clinical, laboratory, and epidemiological features". In: *The American journal of tropical medicine and hygiene* 97.4 (2017), p. 1062.

[25] S. Prajapati, R. Napit, A. Bastola, R. Rauniyar, et al. "Molecular phylogeny and distribution of dengue virus serotypes circulating in Nepal in 2017". In: *PloS One* 15.7 (2020), e0234929.

[26] M. M. Ngwe Tun, K. Pandey, T. Nabeshima, A. K. Kyaw, et al. "An outbreak of dengue virus serotype 2 cosmopolitan genotype in Nepal, 2017". In: *Viruses* 13.8 (2021), p. 1444.

[27] K. P. Acharya, B. Chaulagain, N. Acharya, K. Shrestha, et al. "Establishment and recent surge in spatio-temporal spread of dengue in Nepal". In: *Emerging Microbes & Infections* 9.1 (2020), pp. 676–679.

[28] W. H. O. (WHO). *National Guidelines On Prevention, Management And Control Of Dengue In Nepal*. Accessed August 2023. 2019. URL: https://www.edcd.gov.np/resource-detail/national-guidelines-of-prevention-control-and-management-of-dengue-in-nepal-2019-updated.

[29] C. Viboud, L. Simonsen, and G. Chowell. "A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks". In: *Epidemics* 15 (2016), pp. 27–37.

[30] C. F. Tovissodé, B. E. Lokonon, and R. Glèlè Kakaı=. "On the use of growth models to understand epidemic outbreaks with application to COVID-19 data". In: *Plos one* 15.10 (2020), e0240578.

[31] G. F. Cooper, R. Villamarin, F.-C. R. Tsui, N. Millett, et al. "A method for detecting and characterizing outbreaks of infectious disease from clinical reports". In: *Journal of biomedical informatics* 53 (2015), pp. 15–26.

[32] C. Franco, L. S. Ferreira, V. Sudbrack, M. E. Borges, et al. "Percolation across households in mechanistic models of non-pharmaceutical interventions in SARS-CoV-2 disease dynamics". In: *Epidemics* 39 (2022), p. 100551.

[33] M. Semakula, F. Niragire, S. Nsanzimana, E. Remera, et al. "Spatio-temporal dynamic of the COVID-19 epidemic and the impact of imported cases in Rwanda". In: *BMC Public Health* 23.1 (2023), pp. 1–13.

[34] A. S. Parpia, L. A. Skrip, E. O. Nsoesie, M. C. Ngwa, et al. "Spatio-temporal dynamics of measles outbreaks in Cameroon". In: *Annals of epidemiology* 42 (2020), pp. 64–72.

[35] O. Man, A. Kraay, R. Thomas, J. Trostle, et al. "Characterizing dengue transmission in rural areas: A systematic review". In: *PLOS Neglected Tropical Diseases* 17.6 (2023), e0011333.

[36] N. Imai, I. Dorigatti, S. Cauchemez, and N. M. Ferguson. "Estimating dengue transmission intensity from sero-prevalence surveys in multiple countries". In: *PLoS neglected tropical diseases* 9.4 (2015), e0003719.

[37] P. Roback and J. Legler. "Beyond multiple linear regression". In: *Applied Generalized Linear Models and Multilevel Models in R* (2021), p. 436.

[38] B. K. Acharya, C. Cao, M. Xu, L. Khanal, et al. "Present and future of dengue fever in Nepal: mapping climatic suitability by ecological niche model". In: *International journal of environmental research and public health* 15.2 (2018), p. 187.

[39] U. R. Bhuju, P. R. Shakya, T. B. Basnet, and S. Shrestha. *Nepal biodiversity resource book: protected areas, Ramsar sites, and World Heritage sites.* 2007.

[40] J. L. Nayava. "Rainfall in Nepal". In: *Himalayan Review* 12 (1980), pp. 1–18.

[41] W. H. Organization, S. P. for Research, T. in Tropical Diseases, W. H. O. D. of Control of Neglected Tropical Diseases, et al. *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization, 2019.

[42] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". In: *Nature Scientific Data* 5.1 (2018), pp. 1–12.

[43] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

[44] C. E. Utazi, J. Wagai, O. Pannell, F. T. Cutts, et al. "Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys". In: *Vaccine* 38.14 (2020), pp. 3062–3071.

[45] O. Wariri, C. E. Utazi, U. Okomo, C. J. E. Metcalf, et al. "Mapping the timeliness of routine childhood vaccination in the Gambia: a spatial modelling study". In: *Vaccine* 41.39 (2023), pp. 5696–5705.

[46] I. T. Jolliffe. "Principal component analysis: a beginner's guide—I. Introduction and application". In: *Weather* 45.10 (1990), pp. 375–382.

[47] S. Wold, K. Esbensen, and P. Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[48] A. J. Tatem. "WorldPop, open data for spatial demography". In: *Scientific data* 4.1 (2017), pp. 1–4.

[49] N. Nepal National Statistics Office. *2021 Nepal National Population and Housing Census*. Accessed April 2023. 2021. URL: https://nepal.unfpa.org/en/publications/12th-national-population-and-housing-census-2021.

[50] M. Zwilling. "Negative binomial regression". In: *The Mathematica Journal* 15 (2013).

[51] B. Efron. "The bootstrap and modern statistics". In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1293–1296.

[52] L. D. Howe, B. Galobardes, A. Matijasevich, D. Gordon, et al. "Measuring socio-economic position for epidemiological studies in low-and middle-income countries: a methods of measurement in epidemiology paper". In: *International journal of epidemiology* 41.3 (2012), pp. 871–886.

[53] L. Hjelm, A. Mathiassen, D. Miller, and A. Wadhwa. "Creation of a wealth index". In: *United Nations World Food Programme* (2017).

[54] S. Meyer, L. Held, and M. Höhle. "hhh4: Endemic-epidemic modeling of areal count time series". In: *J Stat Softw* 1 (2016), pp. 1–55.

[55] J. Lu and S. Meyer. "A zero-inflated endemic–epidemic model with an application to measles time series in Germany". In: *Biometrical Journal* 65.8 (2023), p. 2100408.

[56] C. Anderson, D. Lee, and N. Dean. "Identifying clusters in Bayesian disease mapping". In: *Biostatistics* 15.3 (2014), pp. 457–469.

[57] L. M. G. Rincón. "Statistical Methods for Campylobacter Outbreak Detection using Genomics and Epidemiological Data". PhD Thesis. University of Warwick, 2020.

[58] B. A. L. da Fonseca and S. N. S. Fonseca. "Dengue virus infections". In: *Current opinion in pediatrics* 14.1 (2002), pp. 67–71.

[59] B. P. Gupta, A. Haselbeck, J. H. Kim, F. Marks, et al. "The Dengue virus in Nepal: gaps in diagnosis and surveillance". In: *Annals of clinical microbiology and antimicrobials* 17.1 (2018), pp. 1–5.

[60] A. Kolimenakis, S. Heinz, M. L. Wilson, V. Winkler, et al. "The role of urbanisation in the spread of Aedes mosquitoes and the diseases they transmit—A systematic review". In: *PLoS neglected tropical diseases* 15.9 (2021), e0009631.

[61] M. Sugeno, E. C. Kawazu, H. Kim, V. Banouvong, et al. "Association between environmental factors and dengue incidence in Lao People's Democratic Republic: a nationwide time-series study". In: *BMC Public Health* 23.1 (2023), p. 2348.

[62] C. Li, Z. Liu, W. Li, Y. Lin, et al. "Projecting future risk of dengue related to hydrometeorological conditions in mainland China under climate change scenarios: a modelling study". In: *The Lancet Planetary Health* 7.5 (2023), e397–e406.

Chapter 7

Discussion, conclusions and future research

This thesis has developed novel and applied existing statistical methods for modelling and mapping health outcomes from publicly available data in resource-constrained settings. Although the methodologies discussed in this thesis were applied to specific diseases and health outcomes, the methods are broadly applicable to other health outcomes and thus have relevance beyond the scope of the specific health outcomes highlighted in this thesis. The methods developed and applied in this thesis can, therefore, be used to track other publicly available health-related SDGs data.

This chapter sums up the main contributions of each of the papers presented in the thesis and further discusses how each of the works presented in this thesis can be improved upon in future research. Further discussion and conclusions for each of the four papers can be found in the respective chapters of the thesis.

7.1 Extended discussion and future work on spatial and spatio-temporal modelling of multitype typhoid point pattern data

One of the main contributions of Paper 1 is developing a multiple marked inhomogeneous Poisson process model for typhoid that allows for the inclusion of both environmental variables, such as elevation, and individual-level variables, such as age. The modelling framework developed in this paper and applied to the data showed a lower typhoid incidence rate in areas with a high Water, Sanitation and Hygiene (WASH) score. The incidence rate of typhoid was also lower in adults compared to children under 18 years. Our study did not indicate any differences in typhoid incidence between males and females. The age-specific high-resolution maps generated in this study are useful to guide interventions aimed at reducing the incidence of typhoid in Blantyre, Malawi.

One of the main challenges with passive surveillance data, such as the data used in Paper 1 in this thesis, is the underreporting of individuals with typhoid who visit the health facility [1–3]. To account for under-reporting, the model proposed in Paper 1 can be extended using a so-called thinned inhomogeneous Poisson process model, whereby the intensity of the Poisson process in equation 3.3 would be modelled as follows for spatial and spatio-temporal data:

Spatial model:

$$\lambda_{ij}(x) = p_{ij}(x) \exp\left(\alpha_i + \gamma_j + d(x)^T\beta + \log m_{ij}(x)\right) \tag{7.1}$$

Spatio-temporal model:

$$\lambda_{ij}(x,t) = p_{ij}(x,t) \exp\left(\alpha_i + \gamma_j + d(x,t)^T\beta + \log m_{ij}(x,t)\right) \tag{7.2}$$

where $p_{ij}(x)$ and $p_{ij}(x,t)$ are the probabilities of attending the health facility in spatial and spatio-temporal processes, respectively, $m_{ij}(x)$ and $m_{ij}(x,t)$ are the

population of an individual with gender i and age j at location x and time t; and $d(x)^T$ and $d(x,t)^T\beta$ are covariates measured at location x and time t. This probability could then be modelled using a logit-linear regression where the distance from the hospital could also be included. However, one of the challenges of this approach is that some covariates can affect both typhoid fever risk and the probability of visiting a clinic, making the estimation of regression relationships more problematic. This issue has also been reported in ecology, where similar methods have been used in citizen science data [4]. Future research should, therefore, focus on better understanding the factors and mechanisms that drive the likelihood of attending health facilities to parameterize $p_{ij}(x)$ and $p_{ij}(x,t)$ better and overcome the aforementioned identifiability issues in the estimation.

7.2 Extended discussion and future work on modelling and mapping spatio-temporal malnutrition geostatistical data

In Paper 2, we utilized publicly available demographic and health survey data to investigate the multilevel factors of the double burden and triple burden of malnutrition among mother-child pairs in Malawi. The results from this work showed that the odds of DBM were three times higher with a higher proportion of wealthy households in a community. Furthermore, the odds of TBM were 60% lower among mother-child pairs where women had some education compared to pairs with women without education. Furthermore, it highlighted that DBM and TBM are higher in cities than in other areas. This finding is consistent with recent results on DBM in Guatemala [5]. The main contribution in this work was mapping DBM and TBM among mother-child pairs using model-based methods, as this had not been done before our work. Furthermore, this paper built on previous work that maps health outcome geospatial data in the absence of spatial correlation [6].

Future work could build a spatio-temporal model to assess whether the burden of DBM and TBM has also been higher in cities than in other areas over time. This is because some recent descriptive studies have shown a shifting trend in the

burden of DBM from metropolitan areas to other regions [5]. The findings from the spatio-temporal modelling could help the government and implementing partners anticipate which districts/areas in Malawi might get an increasing burden of DBM and TBM in the future.

Similar to the spatial setting, the first step in modelling spatio-temporal data is assessing the presence of residual variation that is not explained by covariates. This can be done by assessing spatio-temporal random effects using a spatio-temporal variogram at the set of points (x_i, t_i). Let $n(u, v)$ denote the pairs (i, j) such that $\|x_i - x_j\| = u$ and $|t_i - t_j| = v$; the empirical spatio-temporal variogram is mathematically defined as

$$\tilde{\gamma}(u, v) = \frac{1}{2|n(u, v)|} \sum_{(i,j) \in n(u,v)} \left\{ \tilde{Z}(x_i, t_i) - \tilde{Z}(x_j, t_j) \right\}^2, \tag{7.3}$$

where $|n(u, v)|$ is the number of pairs in the set.

A Monte Carlo procedure similar to the one presented in the introductory section of this thesis can be used to test for the presence of residual spatio-temporal correlation in the data for random effects arising from model 7.4. Briefly, the following steps are taken to assess the presence of spatio-temporal variation in the prevalence of the health outcome of interest.

(i) Fix the (x_i, t_i) and permute the data and $Z((x_i, t_i))$.

(ii) Calculate the variogram given in equation 7.3.

(iii) Repeat steps (i) and (ii) B times.

(iv) Generate 95% tolerance intervals using the variograms computed in steps (i) to (iii) above.

The data exhibit spatio-temporal correlation if some parts of the empirical variogram fall outside of the 95% tolerance intervals. In the absence of spatio-temporal correlation, the spatio-temporal mixed-effects model would be given as follows. Conditional on some unstructured random effects $Z(x_i, t_i)$, the observed data are mutually independent binomial distributions with a probability of being a case $(p(x; t))$ modelled as:

$$\log\left\{\frac{p\left(x_i,t_i\right)}{1-p\left(x_i,t_i\right)}\right\}=d\left(x_i,t_i\right)^{\top}\beta+Z\left(x_i,t_i\right) \tag{7.4}$$

where $p\left(x_i,t_i\right)$ is the probability of having a health outcome of interest for an individual at location x_i and at time t_i, and β is a vector of coefficients associated with the matrix of spatio-temporal covariates $d\left(x_i,t_i\right)^{\top}$.

7.3 Extended discussion and future work on mapping soil-transmitted helminths in developing countries

In Paper 3, we carried out geospatial modeling and mapping of soil-transmitted helminths (STH) in 35 sub-Saharan countries. Furthermore, the paper demonstrates how, beyond predicting normal World Health Organization (WHO) exceedance probabilities for neglected tropical diseases, one can also classify areas to the WHO endemic classes. The work also developed two R Shiny applications that showcase the results of our modeling. The third paper's main contribution is developing high-resolution maps for STH at the country, pixel, and subnational levels. The paper also proposes a new way of classifying a predictive target into the WHO STH endemicity classes.

One of the notable extensions proposed in the third paper is fitting spatio-temporal geostatistical models in countries with survey data at more than one point in time. Given that the spatio-temporal variogram given in equation 7.3 detects spatio-temporal correlation in the data after accounting for covariates, a spatio-temporal geostatistical model for mutually independent outcomes Y, conditional on unstructured random effects $(Z\left(x_i,t_i\right))$ and a spatio-temporal process $(S\left(x_i,t_i\right))$ can be defined as:

$$\log\left\{\frac{p\left(x_i,t_i\right)}{1-p\left(x_i,t_i\right)}\right\}=d\left(x_i,t_i\right)^{\top}\beta+S\left(x_i,t_i\right)+Z\left(x_i,t_i\right) \tag{7.5}$$

where $p\left(x_i,t_i\right)$ and $d\left(x_i,t_i\right)^{\top}$ are as defined in the non-spatial binomial mixed effects model (equation 7.4). In the above equation, the spatio-temporal process is assumed

to be a stationary and isotropic process with a variance of σ^2 and correlation of [7]:

$$\text{corr}\left\{S(x,t), S\left(x',t'\right)\right\} = \rho\left(x,x',t,t';\theta\right) \tag{7.6}$$

The third paper also briefly discussed the challenges of calibrating models when dealing with small sample sizes and low-prevalence geostatistical data. Future studies could further explore this hypothesis further to identify optimal geostatistical data conditions for predicting a health outcome in unsampled areas.

7.4 Extended discussion and future work on characterizing dengue outbreaks

One of the main contributions of Paper 4 is that it develops a modelling framework that allows for the characterisation of multiple disease outbreak patterns. In this paper, we proposed an extension of the Negative Binomial model to allow for estimating multiple outbreaks within a study region. Additionally, the model estimates the peak and duration of each outbreak intensity function (OIF). The main contribution of this work is the modelling framework that pinpoints multiple outbreaks for annual-level data because previous modelling approaches focussed on characterising more granular data, such as the one collected on a daily, weekly, or monthly basis.

Future studies on disentangling disease outbreaks using annual-level data can extend the modelling in the fourth paper by replacing the outbreak coefficients in equation 6.3 with parameters that determine the intensity of each OIF. Specifically, equation 6.3 can be rewritten as:

$$\lambda_t = \exp\left(\boldsymbol{d}_t\boldsymbol{\beta} + \sum_{i=1}^{3}\pi_i f_{(i,t)}\right) \tag{7.7}$$

where the π_i's ($\pi_1 + \pi_2 + \pi_3 = 1$) are weights that determine the intensity of each outbreak relative to the other two outbreaks.

This extension would help to identify the most intense outbreak in each district or sub-areal unit. By closely examining the most intense outbreak, we can determine the factors contributing to its severity. Understanding these factors is crucial for developing strategies to mitigate them and prevent future intense outbreaks.

Given that fewer than or more than three outbreaks have been detected in an area, our modelling framework can also be easily extended to include as many outbreak parameters as necessary. For instance, a model with n outbreaks can be specified as follows:

$$\lambda_t = \exp\left(\boldsymbol{d}_t\boldsymbol{\beta} + \sum_{i=1}^{n} \gamma_i f_{(i,t)}\right) \tag{7.8}$$

where n is the number of (suspected) outbreaks in the area under consideration.

This study used the squared exponential function to estimate the OIFs. A potential extension of the proposed model would, therefore, explore OIF curves that are not symmetric, such as those derived from the skew-normal distribution, particularly in studies that are not constrained by small sample sizes.

7.5 Conclusion

The work presented in this thesis shows how statistical methods can be used to contribute to achieving the 2030 Sustainable Development Goals (SDGs). Paper 1 contributes to the good health and well being of individuals in an urban setting in Blantye, Malawi, by showing the areas with the highest incidence of typhoid. Identifying areas with high typhoid incidence will also assist policymakers in determining if there is a statistical association between the sources of drinking water and the type of latrines used by people and the occurrence of fever. Consequently, these findings contribute to interventions and policy decisions related to clean water and sanitation (SDG 6, Ensure access to water and sanitation for all).

The malnutrition work among mother-child pairs in Malawi contributes to the

second goal in SDG 2 by contributing to ending all forms of malnutrition by 2030. By identifying the areas with the highest burdens of the double and triple burden of malnutrition among mother-child pairs, this work can assist policy makers to know which areas to target to reduce the prevalence of malnutrition. Similarly, the work on soil-transmitted helminths which produced fine-scale maps and subnational maps for the prevalence of STH are useful in contributing to the fourth target of SDG 3 of fighting communicable diseases.

A key finding in the fourth paper is the identification of multiple outbreaks occurring within the same district over time. The variation in each outbreak's peak and scale parameters across the 77 Nepalese districts underscores the importance of district-specific models rather compared to region-wide models that might fail to account for sub-unit differences. This district-level analysis provides more precise insights into dengue outbreak dynamics, which can be critical for contributing to the reduction of communicable diseases (SDG 3). By addressing local disparities, our proposed model contributes to more targeted and effective public health interventions, ultimately supporting the broader goal of improving health outcomes at the community level.

In conclusion, the work presented in this thesis represents a significant contribution to spatial epidemiology and public health, particularly in resource-constrained settings. The findings from this thesis contribute to the understanding of disease patterns and health outcomes in developing countries and also provide actionable insights for public health practitioners and policy makers. The overarching theme of this thesis is that the development and application of spatial and spatio-temporal methods can help to identify risk factors for health outcomes and the geographic distributions of the health outcomes. Although the methods used in this thesis focused on specific health outcomes, the methodologies discussed are broadly applicable and can be adapted to other health-related Sustainable Development Goals (SDGs). Therefore, this thesis argues and recommends developing and applying spatial and spatio-temporal statistical methods to existing datasets to guide the implementation of interventions.

References

[1] X. Li, H. H. Chang, Q. Cheng, P. A. Collender, et al. "A spatial hierarchical model for integrating and bias-correcting data from passive and active disease surveillance systems". In: *Spatial and Spatio-temporal Epidemiology* 35 (2020), p. 100341.

[2] W. H. Organization et al. *A toolkit for national dengue burden estimation*. Tech. rep. World Health Organization, 2018.

[3] J. D. Stanaway, R. C. Reiner, B. F. Blacker, E. M. Goldberg, et al. "The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet Infectious Diseases* 19.4 (2019), pp. 369–381.

[4] R. B. Dissanayake, E. Giorgi, M. Stevenson, R. Allavena, et al. "Estimating koala density from incidental koala sightings in South-East Queensland, Australia (1997–2013), using a self-exciting spatio-temporal point process model". In: *Ecology and Evolution* 11.20 (2021), pp. 13805–13814.

[5] D. Sagastume, J. L. Peñalvo, M. Ramírez-Zea, K. Polman, et al. "Dynamics of the double burden of malnutrition in Guatemala: a secondary data analysis of the demographic and health surveys from 1998–2015". In: *Public Health* 229 (2024), pp. 135–143.

[6] K. Deribe, A. Mbituyumuremyi, J. Cano, M. J. Bosco, et al. "Geographical distribution and prevalence of podoconiosis in Rwanda: a cross-sectional country-wide survey". In: *The Lancet Global Health* 7.5 (2019), e671–e680.

[7] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

APPENDICES

A Paper 1 Supplementary Material

A.1 Spatial covariates

A.1.1 Elevation

The elevation raster was downloaded from the Worldpop website [1]. Figure A.1 illustrates the elevation in meters in Ndirande township.

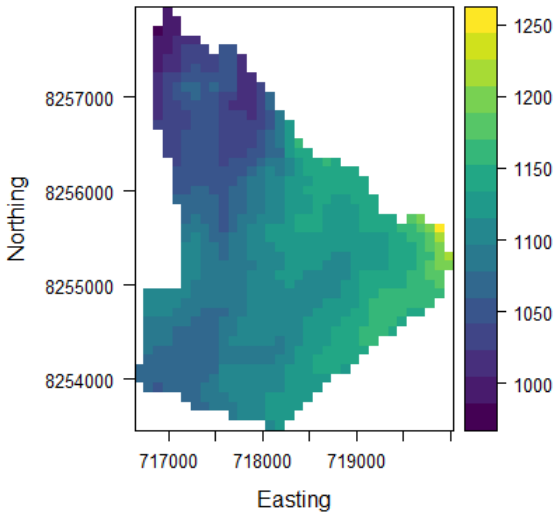

Figure A.1: Elevation (meters)

A.2 Distance to the health facility

We calculated the Euclidean distance from each location in Ndirande township to Ndirande health facility. Ndirande health facility is the largest government owned health facility in Ndirande township. The health facility is illustrated as a white star in Figure A.2

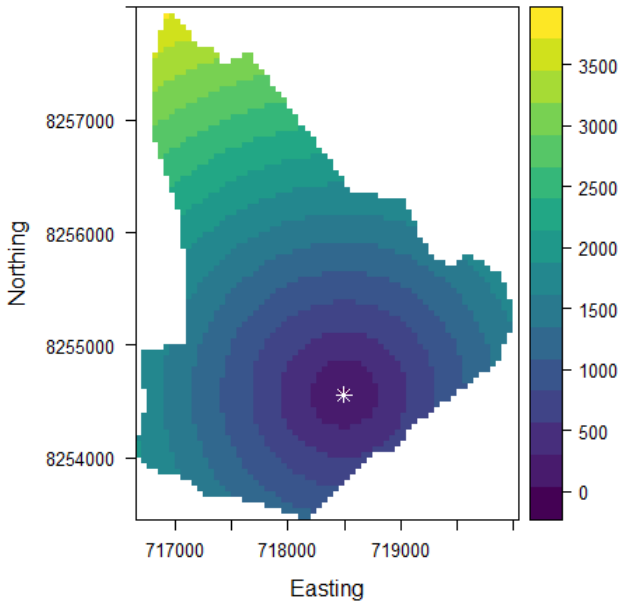

Figure A.2: Distance in meters from every location on the grid to Ndirande Health facility.

A.3 Water, sanitation and hygiene (WASH) score

A water, sanitation, and hygiene (WASH) study was carried out in Ndirande township in 2018 as part of the STRATAA study. A total of 14,136 households were sampled in the study. Households were asked several questions related to their WASH and economic levels. Some of the questions asked to these households included:

- The number of rooms a house has (continuous variable).
- The type of toilet used by the house (no toilet facility, toilet shared with other households (public), toilet shared with neighbours and household use only).
- Material of the toilet used by the household (open defecation, pit latrine with a wooden or soil floor, pit latrine with slab, flush or pour toilet).
- The main source of drinking water for a household (borehole and other unprotected sources, public standpipe, piped to the house, protected well or borehole, private tap located outside of the house, public standpipe and public tap outside the house).

A WASH score was derived from the above questions using Principal Components Analysis (PCA). Figure A.3 shows the percentage of variation that was explained by

the components. It is common practice in epidemiological studies that measure the socioeconomic status of a household to use the first component to derive a desired socioeconomic index or score [2, 3]. Our WASH score was, therefore, based on the first principal component.

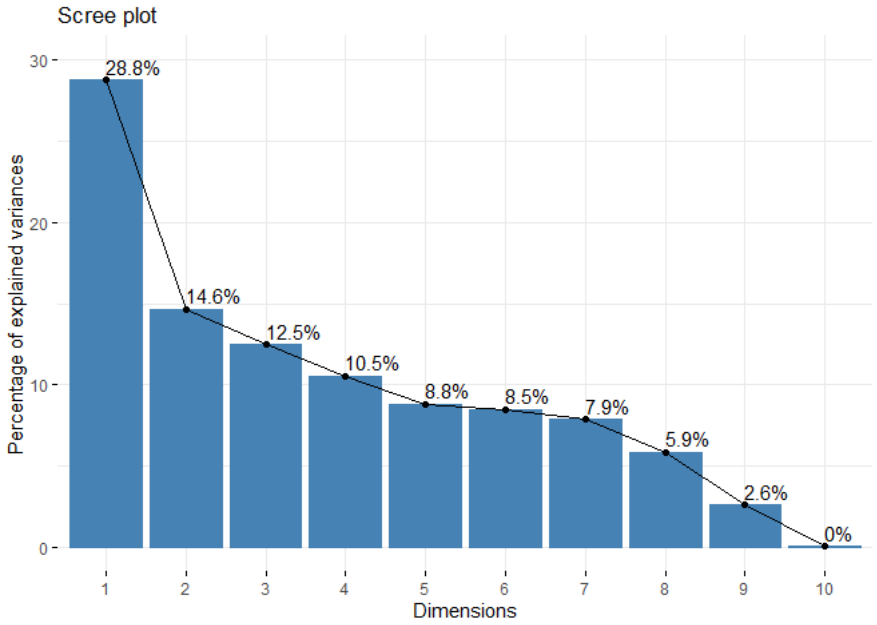

Figure A.3: Eigen values illustrating the variance percentage explained by each component

Figure A.4 illustrates that the main contributions to the first component of the PCA was from the type of toilet facility and the number of rooms in a house.

The PCA score was then fitted to a linear geostatistical model [4] given below using the PrevMap package [5].

$$Y(x_i) = \mu + Z_i + S(x_i) \tag{A.1}$$

where $Y(x_i)$ is the observed WASH score at location i, μ is the constant mean effect (intercept), Z_i ($\sim N(0, \tau^2)$) are independently distributed Gaussian variables, and $S(x_i)$ ($\sim N(0, \sigma^2)$ is a zero-mean stationary and isotropic Gaussian process.

After assessing the goodness of fit of the model using a semi-variogram, a linear prediction over the whole study area was carried out. This prediction was converted to a raster and used as a covariate in the model.

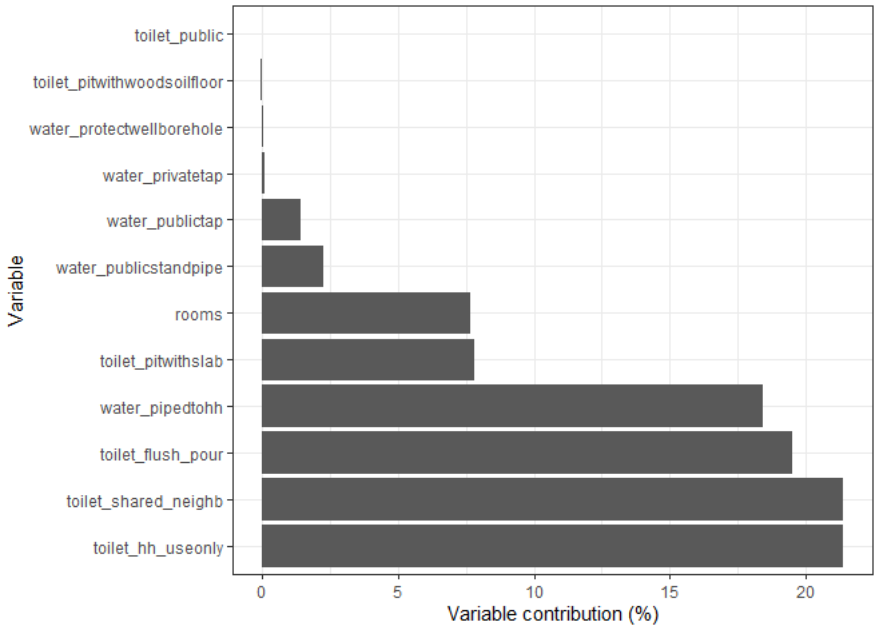

Figure A.4: Contribution of variables to the WASH score

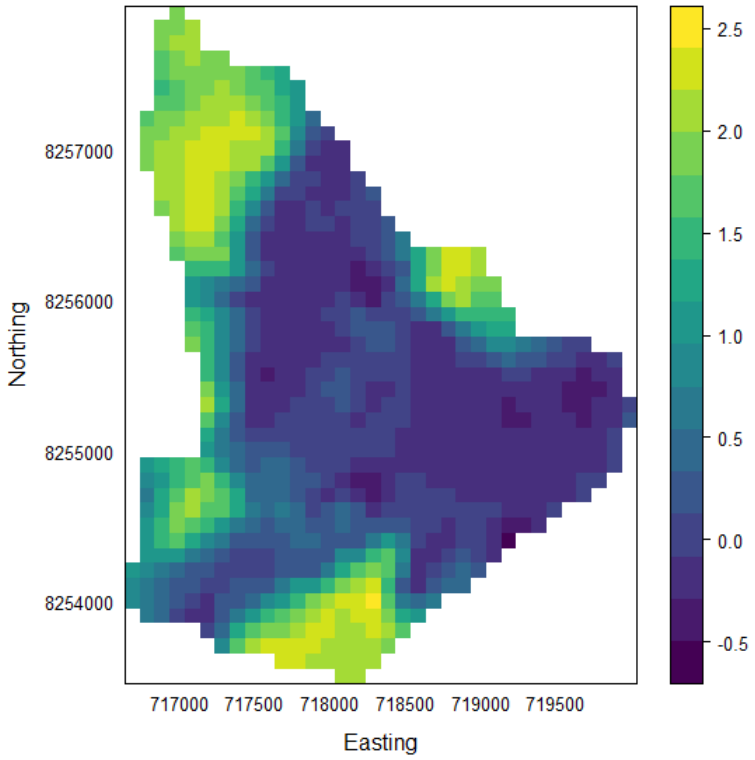

Figure A.5: Interpolated Water Sanitation and Hygiene (WASH) score

Figure A.6 below illustrates the model validation for the linear geostatistical model used to predict a WASH score throughout Ndirande. The empirical variogram (which shows the residual spatial correlation) is shown in red, whilst the black

dotted lines show the simulated envelope for the variogram. Since some parts of the empirical variogram fell outside the simulated envelope, we rejected the null hypothesis of no spatial correlation for the WASH data. We, therefore, concluded that there was some spatial correlation in the WASH survey.

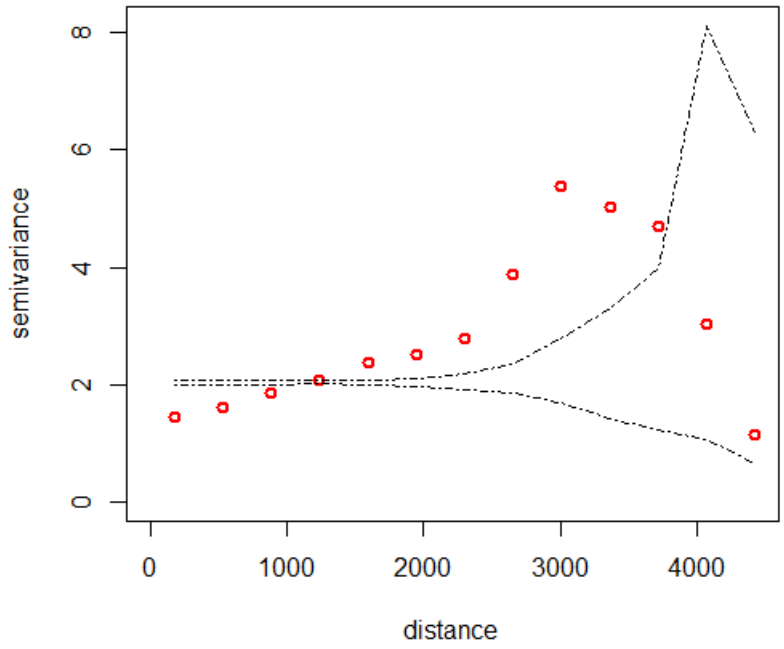

Figure A.6: Simulated envelope (black lines) and empirical variogram (red)

A.4 Ndirande population distribution plots

Figure A.7 illustrates the estimated total number of people per grid cell (population count) at 100 m resolution and the estimated population density per grid cell at 1km resolution in Ndirande in 2018. The age-gender specific population distribution plots are presented in Figure A.8.

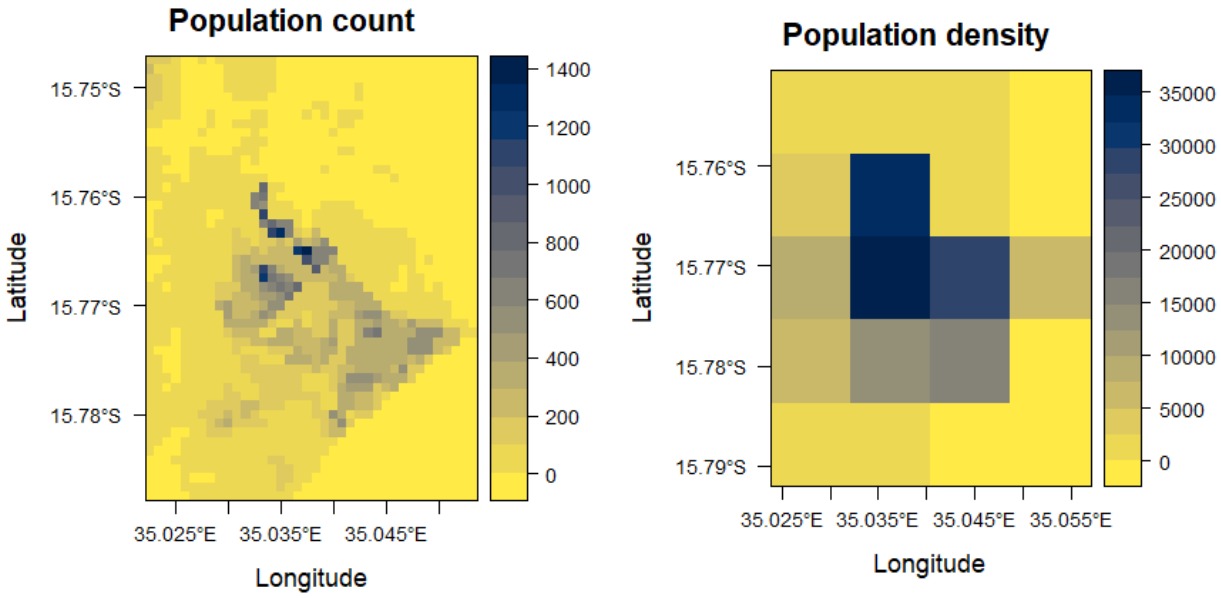

Figure A.7: Map of population distribution in Ndirande in 2018.

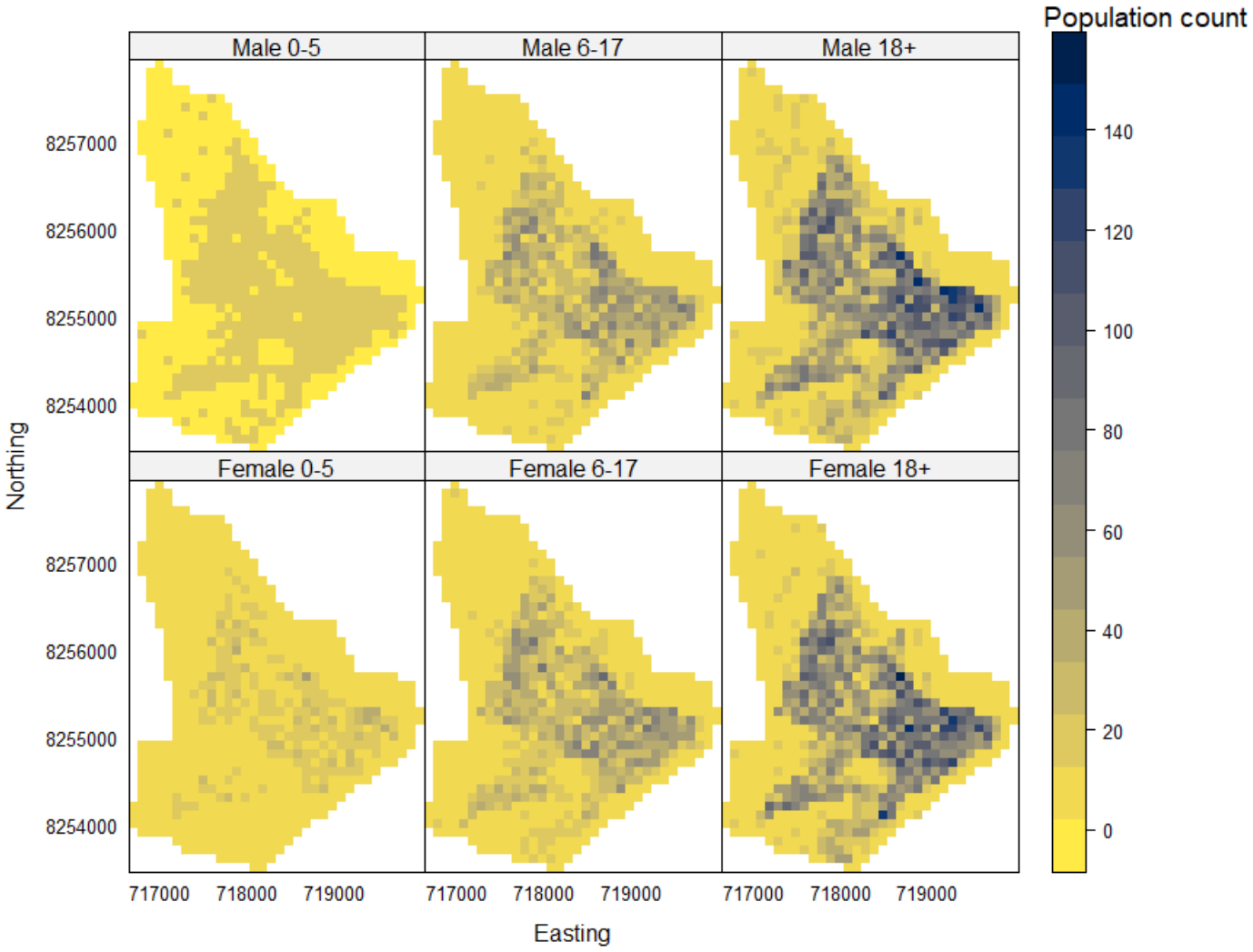

Figure A.8: Map of age and gender-specific population distribution in Ndirande in 2018.

A.5 Model validation plots

We fitted an inhomogeneous K-function to validate our spatial point pattern model. The list of figures below (Figures A.9, A.10, A.11, A.12, A.13 and A.14) show that the K-functions from the observed data mostly fell within the simulated envelope for most of the distances. This suggests that our model was a good fit for the data.

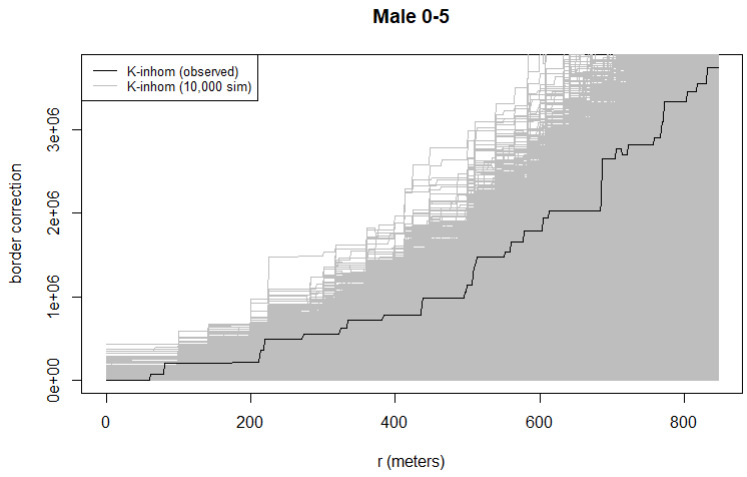

Figure A.9: Spatial inhomogeneous K-function for males aged between 0 and 5 years. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

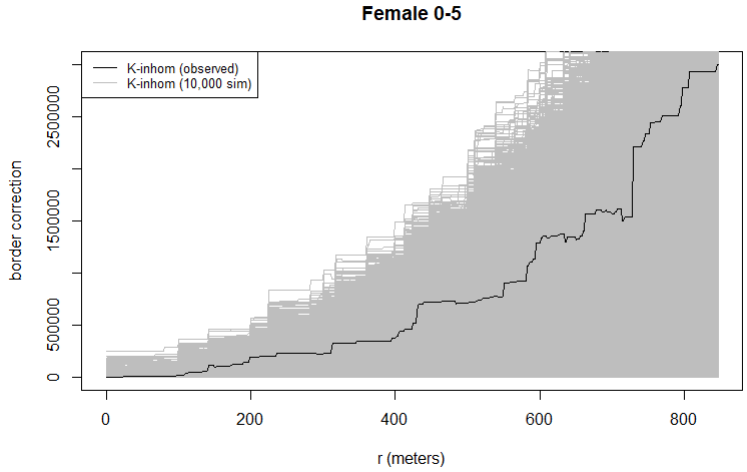

Figure A.10: Spatial inhomogeneous K-function for females aged between 0 and 5 years. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

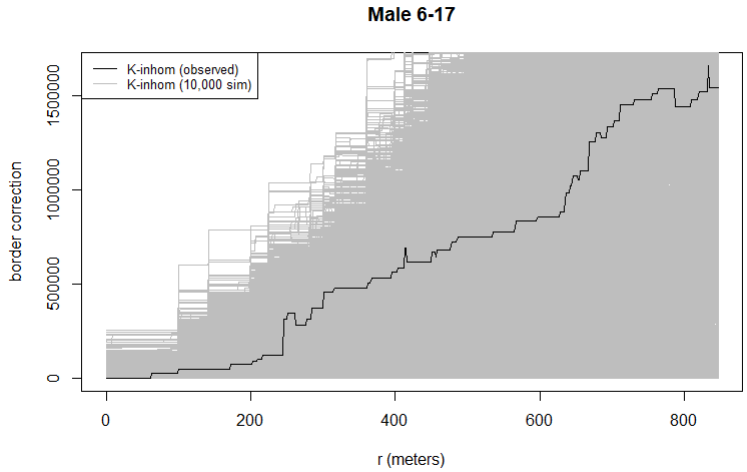

Figure A.11: Spatial inhomogeneous K-function for males aged between 6 and 17 years. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

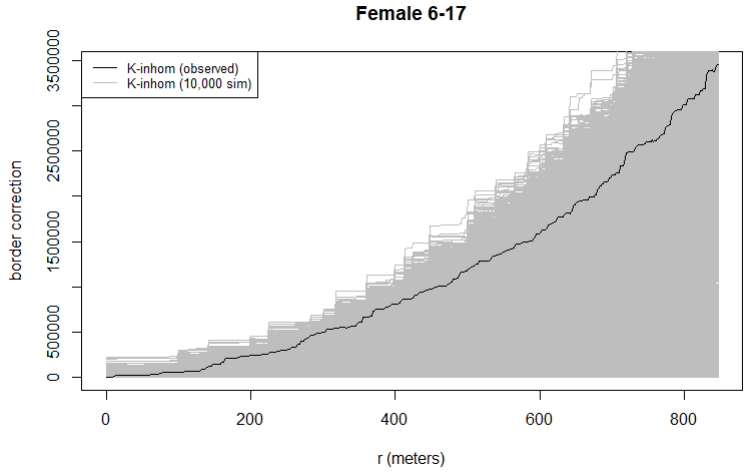

Figure A.12: Spatial inhomogeneous K-function for females aged between 6 and 17 years. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

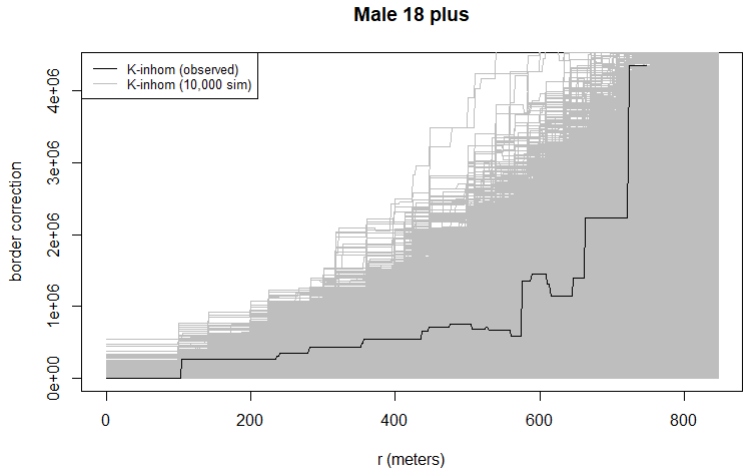

Figure A.13: Spatial inhomogeneous K-function for males aged 18 years and above. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

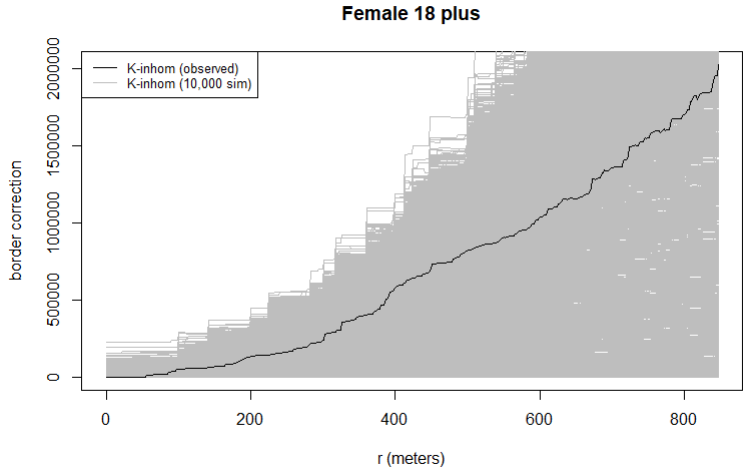

Figure A.14: Spatial inhomogeneous K-function for females aged 18 years and above. The black line represents the inhomogeneous K-functions from the observed data, whilst the grey areas represent the inhomogeneous K-functions from the 10,000 realised bootstrap samples

A.6 Spatio-temporal model

A.6.1 Model formulation

A spatio-temporal point pattern process can be defined as a realization of a stochastic process whose events are countable [6]. The set of events can be written as (x_k, t_k) where $x_k \in \Re^2$ is the location of an event and $t_k \in \Re^+$ is the time at

which event k occurred [7]. The log-likelihood of this process for a marked scenario is given as:

$$L_{ij}(\theta) = \sum_{i=1}^{2} \sum_{j=1}^{3} L_{ij}(\theta) \tag{A.2}$$

$$L_{ij}(\theta) = \sum_{k=1}^{n_{ij}} \log \lambda_{ij}\left(x_k, t_k\right) - \int_A \int_T \lambda_{ij}\left(x, t\right) dxdt \tag{A.3}$$

and $\lambda_{ij}\left(x, t\right) = \exp\left(\alpha_i + \gamma_j + d\left(x, t\right)' \beta + \log m_{ij}(x, t)\right)$ are intensities of the spatial and spatio-temporal point processes. In equation A.3:

- x_k for $k = 1, ..., n$ are locations for the observed typhoid cases at time t for a typhoid case with gender i (male or female) and age j (0-5 years, 6-17 years or 18+ years)
- A is the study region and T the temporal region
- $\lambda\left(x, t\right)$ is the intensity of the process
- α_i are the intercepts for typhoid case with gender i and γ_j the intercepts for a typhoid case with age j
- $d\left(x, t\right)$ is the matrix of spatial and temporal covariates (such as distance to Ndirande health clinic in meters, elevation in meters, WASH score, and season) with their associated coefficients β.
- $m_{ij}(x, t)$ is an offset corresponding to the population for an individual with gender i and age j at location x and time t.

Model A.3 uses the same bootstrap procedure for confidence intervals that was defined in the main paper for the purely spatial model.

A.6.2 Model validation

Similar to the purely spatial model defined in the main paper, the spatio-temporal model can be validated using a spatio-temporal inhomogeneous K-function. The inhomogeneous spatio-temporal K-function is given as:

The space-time inhomogeneous function is defined as [8] :

$$K_{AT}(u,v) = 2\pi \int_0^v \int_0^u g\left(u',v'\right) u' du' dv' \tag{A.4}$$

where

- u is the change in space ($\|x - x'\|$) and v the change in time ($|t - t'|$)
- (u, v) is a vector representing differences in the spatio-temporal domain
- $g(u,v) = \frac{\lambda_2(u,v)}{\lambda(x,t)\lambda(x',t')}$

A non-parametric version of equation A.4 can be implemented in the *stpp* software. The non-parametric spatio-temporal inhomogeneous K-function for an infectious disease such as typhoid is mathematically defined as follows [7]:

$$\widehat{K}_{AT}(u,v) = \frac{1}{|A \times T|} \frac{n}{n_v} \sum_{k=1}^{n_v} \sum_{h=1;h>k}^{n_v} \frac{1}{w_{kh}} \frac{1}{\lambda\left(x_k,t_k\right)\lambda\left(x_h,t_h\right)} \mathbf{1}_{\{\|x_k - x_h\| \leq u; t_h - t_k \leq v\}} \tag{A.5}$$

The parameter w_{kh} in equation A.5 denotes the spatial edge correction factor whilst n_v denotes the number of (typhoid) occurrences for which $t_k \leq T_1 - v, T = [T_0, T_1]$ [7].

B Paper 2 Supplementary Material

B.1 DBM and TBM samples flow chart

Figure B.1 below shows illustrates how the DBM and TBM samples were derived from the 2015-16 Malawi Demographic and Health Survey data.

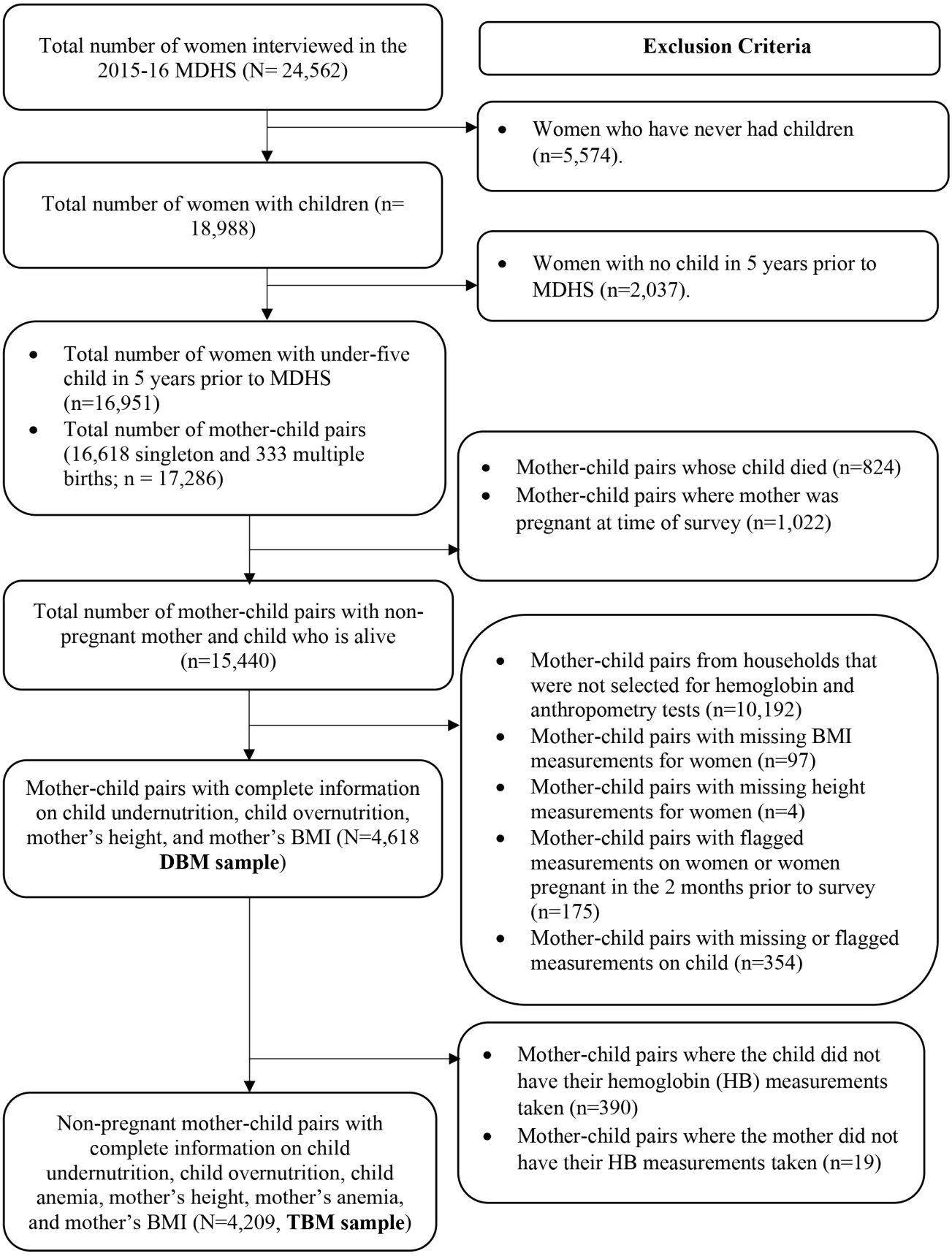

Figure B.1: Flowchart of the sample included in the analysis from the 2015-16 Malawi Demographic and Health Survey (MDHS) (numbers are not weighted).

B.2 Sources of spatial covariates

A gridded continuous map of the 2015 nightlight data for Malawi was downloaded from the Worldpop website [1]. The Worldpop is a research centre that models and develops gridded continuous maps, otherwise known as raster datasets, to aid in mapping and geospatial modelling. The elevation raster for Malawi was also downloaded from Worldpop [1]. Both the nightlight and elevation rasters had spatial resolutions of $100m^2$. Climate data, including temperature, precipitation and evapotranspiration were extracted from the TerraClimate website for 2015 [9]. We computed an aridity index by taking the proportion of the precipitation to the evapotranspiration. The climate data are characterized by a monthly time interval and a spatial resolution of approximately 4 km (equivalent to 1/24th of a degree).

We also used spatial covariates generated by the Demographic and Health Survey (DHS) Program. We, specifically, obtained data on the percentage of women aged 15-49 who are literate, percentage of children 12-23 months who had received all 8 basic vaccinations and percentage of women who had a live birth in the five years preceding the survey who had 4 or more antenatal care visits [10–12]. All the DHS rasters were modelled using the 2015-16 Malawi DHS and had a resolution of 5km $\times$ 5km.

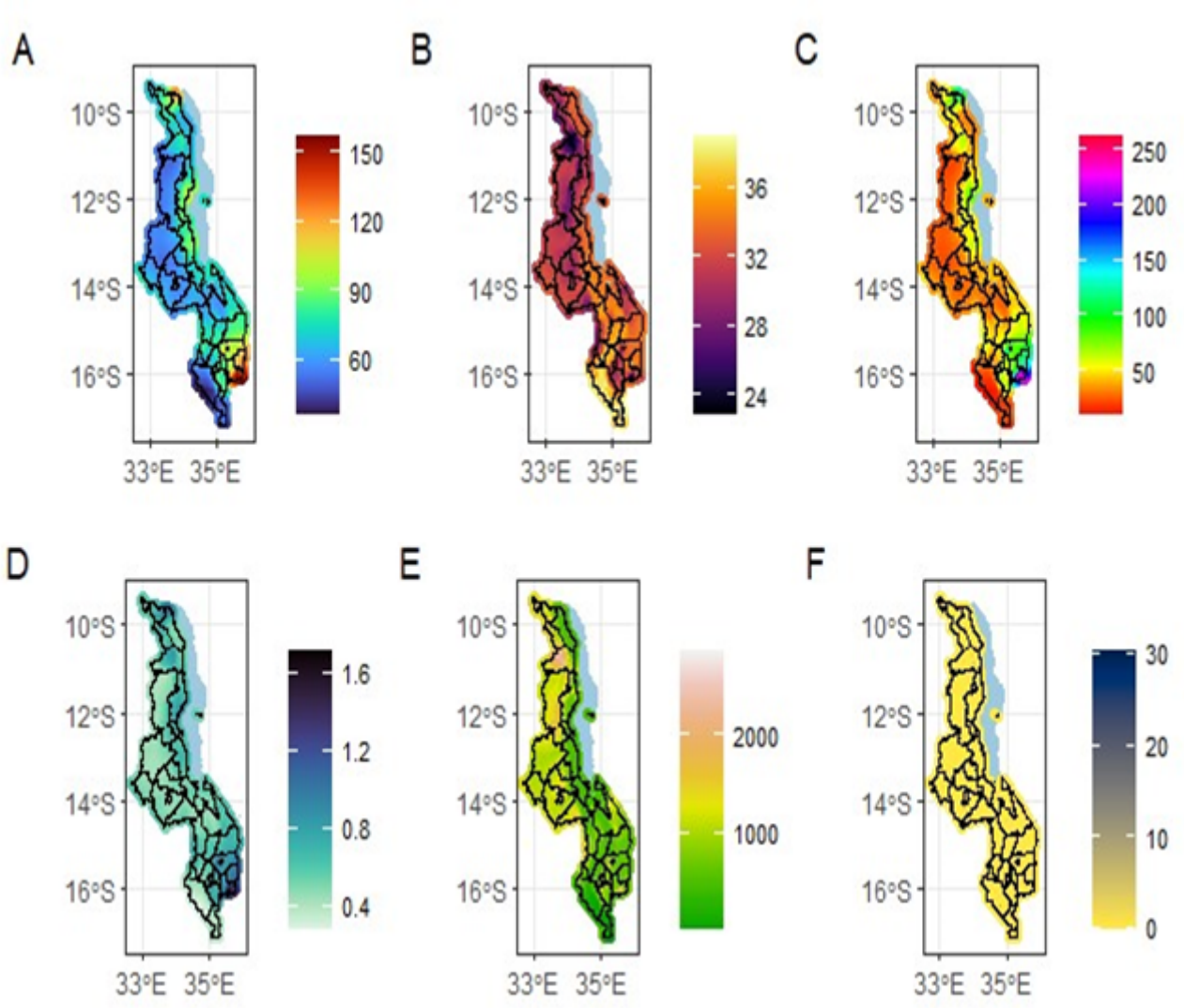

Figure B.2: Maps of spatial covariates used to model and map the prevalence of DBM and TBM. A = Precipitation (mm); B = Maximum temperature (∘C); C = Evapotranspiration (mm); D= Aridity index (ratio of Precipitation to Evapotranspiration); E = Elevation (m); F = Nightlight (nanoWatts/cm2/sr).

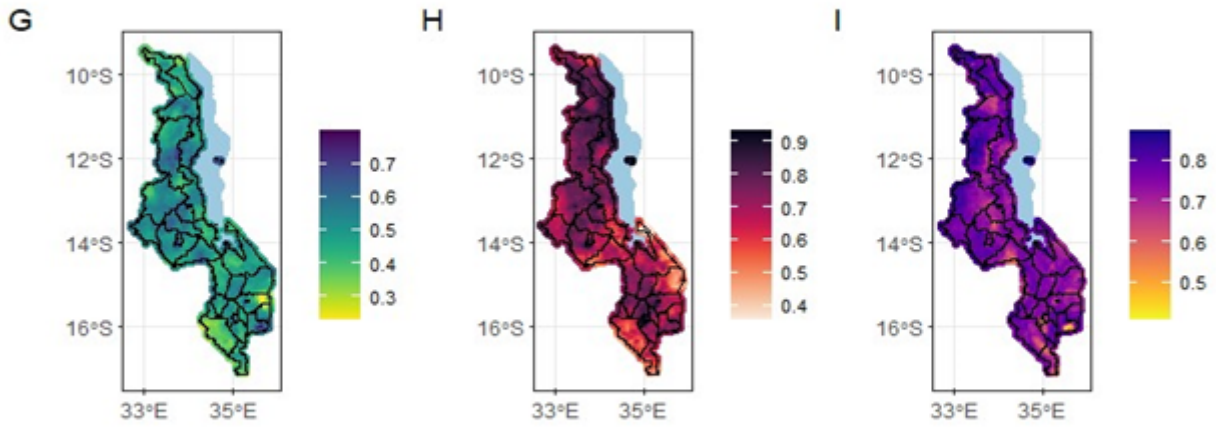

Figure B.3: Maps of spatial covariates used to model and map the prevalence of DBM and TBM. G = Percentage of women who had a live birth in the five years preceding the survey who had 4+ antenatal care visits; H = Percentage of women aged 15-49 who are literate; I = Percentage of children 12-23 months who had received all 8 basic vaccinations.

B.3 WHO Conceptual Framework for the Double Burden of Malnutrition

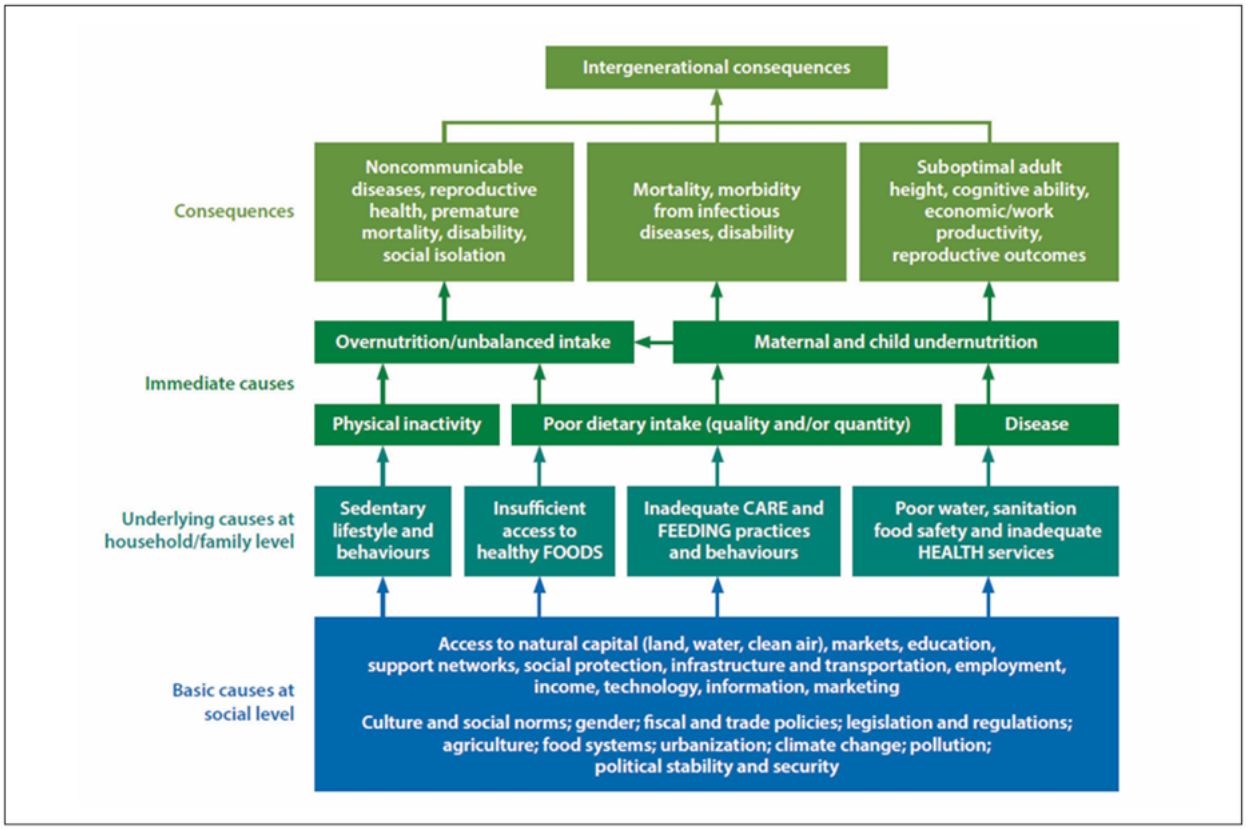

Figure B.4: The World Health Organization Conceptual Framework for DBM.

B.4 Definitions of variables used in the multilevel analyses

Table B.1: Definitions of variables used in the analysis for DBM and TBM.

Variable name	Definition	Variable code(s)
Mother-child pair variables		
Age group of child in months	<12, 12-23, 24-35, 36-47, 48-59 months	B8
Sex of child	Male, Female	B4
Age group of mother at the time of the survey	15-19, 20-34, 35+ years	V013
Mother's employment status	Not working, working	V714
Parity	<2, 2+	V201
Fewer than 4 ANC visits during pregnancy with most recently born child	Yes (<4 ANC visits), No ($\geq$ 4 visits)	M14
Mother's highest completed education level	None, Primary (incomplete and complete), secondary (incomplete and complete), and higher	V149
Mother's marital status at the time of the survey	In a union (married or cohabiting) and not in a union (never been in a union, single, divorced, or widowed)	V502
Household-level variables		
Household wealth index	Low, middle, high	V191
Community-level variables		
Area of residence	Urban, Rural	V025
Proportion of households in middle or rich wealth quintiles in a cluster	Total number of households with middle or higher wealth quintile in a cluster, divided by total number of households in a cluster	V190, V001
Proportion of women with fewer than 4 ANC visits during pregnancy of child in mother-child pair in a cluster	Total number of mothers who attended less than 4 ANC visits in a cluster, divided by the total number of mothers in that cluster	V190, V001

B.5 Multilevel binomial mixed effects model

Let Y_{jk} be the random variable denoting the number of mother-child pairs with DBM or TBM out of the total number of mother-child pairs in a household j that is nested within a cluster k. We then modelled the outcome variable Y_{jk} using a three-level binomial mixed effects model with probability ($p(x_{jk})$) of having DBM or TBM such that:

$$\log\left(\frac{p(x_{jk})}{1-p(x_{jk})}\right) = d(x)^T\beta + \varphi_k + \delta_{jk} \tag{B.1}$$

where $d(x)$ is a vector of variables associated with the regression coefficients β, and φ_k is the effect of cluster/community k (k=1,2,...,K) and δ_{jk} is the effect of household j (j=1,2,...,J) within cluster k. The random effects ($\varphi_k + \delta_{jk}$) are assumed to be independent of one another and are normally distributed with zero means and constant variances ($\varphi_k \sim N(0, \sigma_k), \delta_{jk} \sim N(0, \sigma_{jk})$ [13, 14]. The individual, household, and cluster-level independent variables that were included in the analysis of DBM and TBM are described in the main manuscript.

B.6 Exploratory analysis for spatial correlation using the theoretical variogram

To assess whether there was any spatial correlation in DBM and TBM in Malawi, we fitted the binomial mixed model defined above and we included the georeferenced covariates described in section 2.1 above. We fitted this model to all the 9 outcomes (child wasting, child stunting, child underweight, child overweight, child anaemia, maternal short height, maternal underweight, maternal overnutrition, and maternal anaemia) in R using the lme4 package. To test for spatial correlation in the data, we extracted the random effects ($\varphi_k + \delta_{jk}$) from the mixed effects model for each of the 9 models and fitted them to a variogram [15–17]. We then generated 95% confidence intervals of the random effects variograms under the assumption of spatial independence. We concluded that there was no evidence of spatial correlation in the data since the variograms for all the 9 models fell within the 95% confidence bounds (Figure B.5 and Figure B.6). We, therefore, fitted non-spatial mixed effects models for all the 9 outcomes and not geostatistical models since the data did not show any evidence of residual spatial correlation [17, 18]. These models were used to compute continuous predictions of the 9 outcomes at 3 Km^2 grids and at district level.

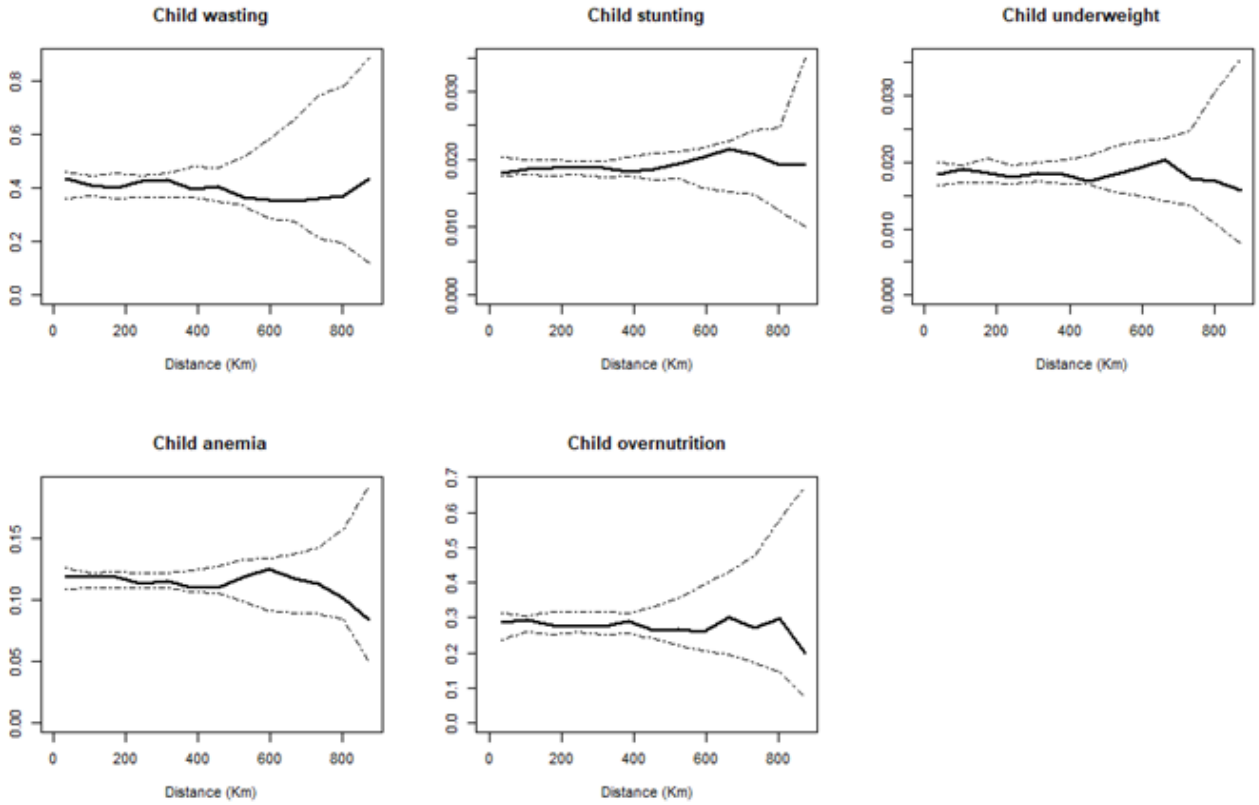

Figure B.5: Plot of the empirical variogram (represented by the black solid line) based on the random effects from Binomial mixed models for the child-level outcomes. The dotted lines correspond to 95% confidence intervals generated under the assumption of spatial independence [4].

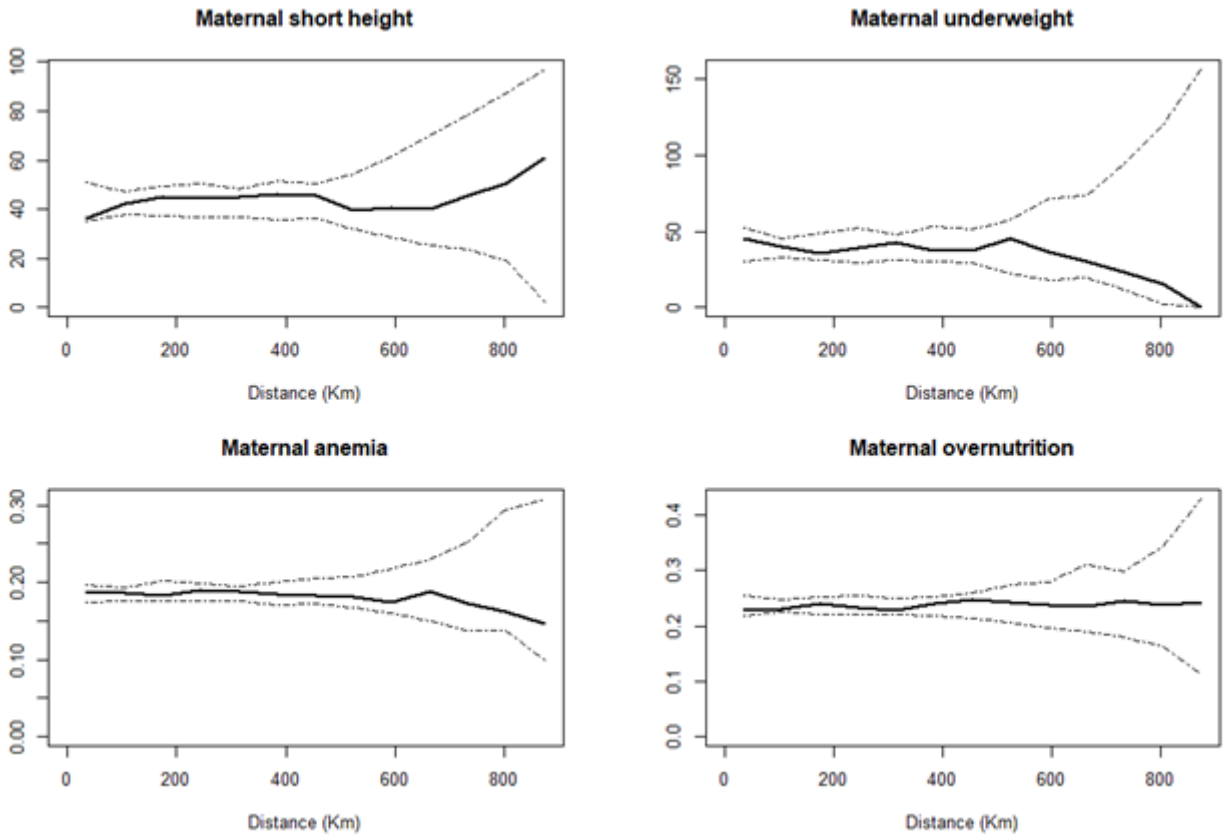

Figure B.6: Plot of the empirical variogram (represented by the black solid line) based on the random effects from Binomial mixed models for the maternal-level outcomes The dotted lines correspond to 95% confidence intervals generated under assumption of spatial independence [4].

B.7 Model estimates

Table B.2: Model parameter estimates and associated 95% confidence intervals (CI) for for child-level outcomes.

Variable	**Child underweight**	**Child wasting**	**Child stunting**	**Child anaemia**	**Child overnutrition**
	Estimate (95% CI)	**Estimate (95% CI)**	**Estimate (95% CI)**	**Estimate (95% CI)**	**Estimate (95% CI)**
Intercept	-1.26 (-2.03, -0.49)	-6.87 (-11.52, -2.22)	-0.01 (-0.54, 0.53)	2.17 (1.50, 2.93)	-6.48 (-9.64, -3.31)
Nightlight	-0.06 (-0.12, 0.001)	NA	-0.05 (-0.09, -0.01)	-0.02 (-0.06, 0.04)	NA
Maximum temperature	NA	0.08 (-0.03, 0.18)	NA	NA	NA
Percentage of literate women	-1.14 (-1.14, -0.08)	-0.39 (-2.92, 2.14)	-0.79 (-1.53, -0.05)	-0.21 (-3.03, -1.06)	-0.61 (-2.56, 1.33)
Proportion of children fully vaccinated	NA	NA	NA	NA	3.90 (0.06, 7.74)

Note: NA corresponds to a situation where the spatial covariate is not included in the model.

Table B.3: Model parameter estimates and associated 95% confidence intervals (CI) for maternal-level outcomes.

Variable	**Maternal underweight**	**Maternal short height**	**Maternal anaemia**	**Maternal overnutrition**
	Estimate (95% CI)	**Estimate (95% CI)**	**Estimate (95% CI)**	**Estimate (95% CI)**
Intercept	-6.88 (-10.43, -3.32)	-4.08 (-5.60, -2.57)	-4.68 (-5.90, -3.46)	-4.55 (-5.45, -3.65)
Nightlight	0.01 (-0.13, 0.17)	-0.05 (-0.16, 0.05)	NA	0.08 (0.04, 0.12)
Percentage of literate women	-0.33 (-3.90, -3.25)	0.26 (-1.80, 2.32)	0.11 (0.07, 0.15)	3.74 (2.54, 4.94)

Note: NA corresponds to a situation where the spatial covariate is not included in the model.

B.8 Additional results

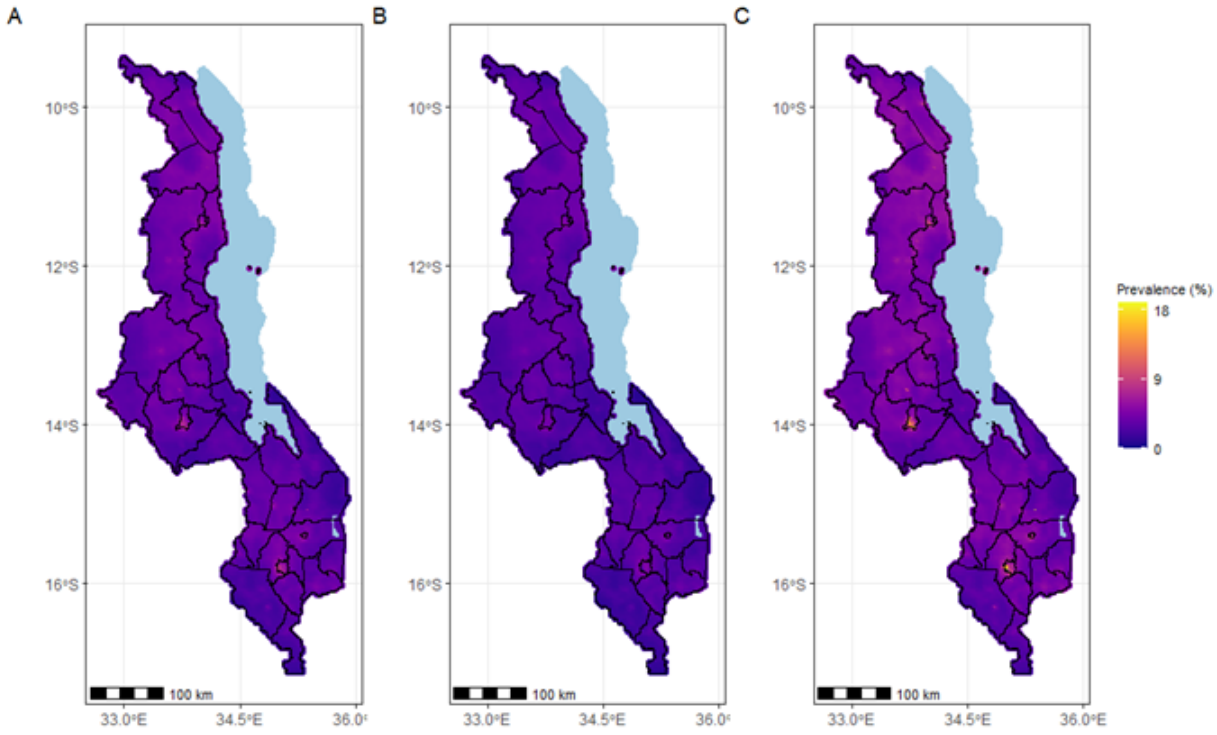

Figure B.7: Predicted mother-child pair DBM prevalence maps of Malawi; mean predicted prevalence (A) and, lower (B) and upper 95% CI bounds (C).

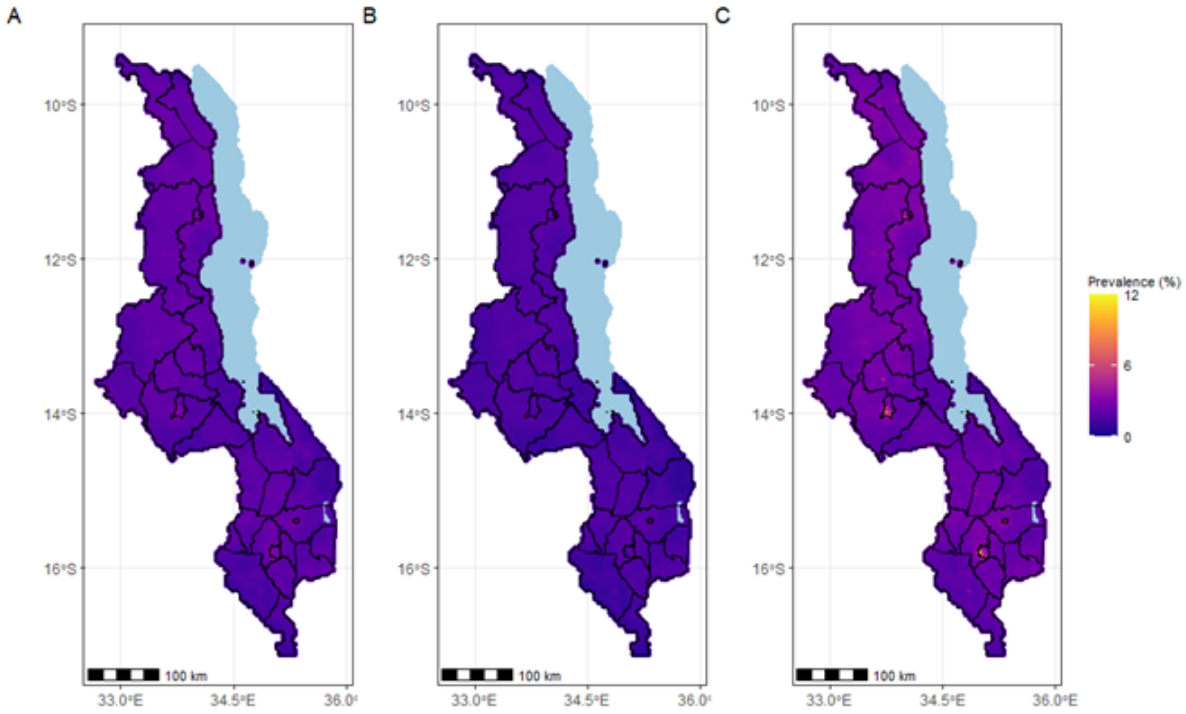

Figure B.8: Predicted mother-child pair TBM prevalence maps of Malawi; mean predicted prevalence (A) and, lower (B) and upper 95% CI bounds (C).

B.9 Model predictions

We predicted the prevalence of DBM and TBM across Malawi at point-level. We defined our predictive target as T^*. We divided the continuous surface of Malawi into a regular grid of $3km^2$. The following equation was used to generate prevalence surfaces for DBM and TBM: $T^* = p(x) : x \in A$ where $T = d(x)^T\beta + \varphi_k + \delta_j k$ for each of the 9 outcomes and A denotes the regular grid for Malawi for the point estimates [4].

To generate maps for DBM and TBM from the non-spatial mixed effects models, we used statistical probability rules that are also employed in other health research that use outcomes that are derived from multiple indicators [19] as follows. Let:

- $P(A)$ be the probability that a child is stunted
- $P(B)$ probability that a child is wasted
- $P(C)$ probability that a child is underweight
- $P(D)$ be the probability that a child is overweight or obese

- $P(E)$ be the probability that a child is anaemic
- $P(F)$ be the probability that a mother has a short height
- $P(G)$ be the probability that a mother is underweight
- $P(H)$ be the probability that a mother is overweight or obese
- $P(I)$ be the probability that a mother is anaemic.

Then the probability that a child has undernutrition (is stunted or wasted or underweight) is defined as:

$$P(\text{child undernutrition}) = 1 - \{(1 - P(A)) \times (1 - P(B)) \times (1 - P(C))\} \quad \text{(B.2)}$$

And the probability that a mother has undernutrition (short height or underweight) is defined as:

$$P(\text{maternal undernutrition}) = 1 - \{(1 - P(F)) \times (1 - P(G))\} \quad \text{(B.3)}$$

Consequently, the probability that a mother-child pair has the double burden of malnutrition (DBM1 = child undernutrition and maternal overnutrition OR DBM2 = child overnutrition and maternal undernutrition) is defined as:

$$P(\text{DBM1}) = P(\text{child undernutrition}) \times P(\text{maternal overnutrition}) \quad \text{(B.4)}$$

$$P(\text{DBM2}) = P(\text{child overnutrition}) \times P(\text{maternal undernutrition}) \quad \text{(B.5)}$$

Then the overall double burden of malnutrition (child undernutrition and maternal overnutrition OR child overnutrition and maternal undernutrition) is given as:

$$P(\text{Any DBM}) = 1 - \{(1 - P(\text{DBM1})) \times (1 - P(\text{DBM2}))\} \quad \text{(B.6)}$$

The probability that a mother-child pair has the triple burden of malnutrition (TBM1 = child undernutrition, maternal overnutrition and child anaemia OR TBM2 = child overnutrition, maternal undernutrition and maternal anaemia) is

calculated as follows:

$$P(\text{TBM1}) = P(\text{child undernutrition}) \times P(\text{maternal overnutrition}) \times P(\text{child anaemia}) \tag{B.7}$$

$$P(\text{TBM2}) = P(\text{child overnutrition}) \times P(\text{maternal undernutrition}) \times P(\text{maternal anaemia}) \tag{B.8}$$

Similar to the overall double burden of malnutrition, the overall triple burden of malnutrition (child undernutrition, child anaemia and maternal overnutrition OR child overnutrition, maternal undernutrition, and maternal anaemia) is given as:

$$P(\text{Any TBM}) = 1 - \{(1 - P(\text{TBM1})) \times (1 - P(\text{TBM2}))\} \tag{B.9}$$

C Paper 3 Supplementary Material

C.1 Summary of Data Characteristics: Countries, Dates, and Sample Sizes

Table C.1 outlines the countries included in the study, the sample sizes, and the years the data were collected. The sample sizes are the raw data obtained from the ESPEN website, prior to merging the geolocated STH data with the spatial covariates.

Table C.1: Countries included in the analysis, data collection dates, and sample sizes per country.

Country	Year	Sample size	Mean (Std. dev)	Median (IQR)
Southern Africa				
Botswana	2015	128	46 (7)	48 (43, 50)
South Africa	2017	152	41 (11)	47 (35, 49)
Swaziland	2015	262	50 (3)	50 (50, 50)
Central Africa				
Angola	2014	121	29 (3)	30 (30, 30)
Cameroon	2012	184	50 (0)	50 (50, 50)
Chad	2015	281	49 (4)	50 (50, 50)
DRC	2015	112	57 (22)	50 (50, 50)
Gabon	2015	182	49 (25)	48 (34, 53)
Eastern Africa				
Burundi	2014	209	50 (0)	50 (50, 50)
Eritrea	2015	162	51 (8)	52 (50, 56)
Ethiopia	2009	102	105 (32)	105 (97, 107)
Kenya	2015	63	58 (4)	58 (56, 60)
Madagascar	2015	305	51 (10)	50 (50, 50)
Malawi	2018	277	29 (2)	30 (29, 30)
Mozambique	2007	134	50 (0)	50 (50, 50)
Rwanda	2014	183	50 (1)	50 (50, 50)
South Sudan	2018	103	48 (4)	50 (49, 50)
Tanzania (Mainland)	2018	301	34 (10)	30 (30, 30)
Uganda	2013	83	58 (14)	60 (60, 63)
Zambia	2005	57	60 (4)	60 (60, 61)
Zimbabwe	2010	126	42 (10)	45 (37, 49)
Western Africa				
Benin	2017	66	70 (31)	50 (49, 80)
Burkina Faso	2004	87	59 (6)	60 (59, 61)
Cote d'Ivoire	2014	529	31 (2)	30 (30, 32)
The Gambia	2015	206	50 (5)	50 (49, 50)
Ghana	2008	76	59 (4)	60 (60, 60)
Guinea-Bissau	2018	55	49 (2)	50 (50, 50)
Liberia	2015	320	49 (4)	50 (50, 50)
Mali	2004	187	71 (18)	69 (66, 70)
Mauritania	2015	72	57 (39)	50 (47, 62)
Niger	2006	73	65 (31)	60 (60, 60)
Nigeria	2014	705	50 (5)	50 (50, 52)
Senegal	2013	117	50 (5)	50 (50, 50)
Sierra Leone	2008	114	83 (32)	99 (55, 110)
Togo	2015	1077	15 (0)	15 (15, 15)

Std. dev = Standard deviation.
IQR = Interquartile range (25^{th} quartile, 75^{th} quartile).

Table C.2 outlines the countries excluded from the study, the sample sizes, and the years the data were collected.

Table C.2: Countries excluded from the analysis, data collection dates, and sample sizes per country.

Country	**Year**	**Sample size**	**Comment**
Algeria	NA	NA	No data (zero data points) for STH
Cape Verde	2012	9	All data points do not have location data
Central African Republic	1983	1	None
Comoros	NA	NA	No data (zero data points) for STH
Equatorial Guinea	NA	NA	No data (zero data points) for STH
Guinea	2013	40	None
Mauritius	2015	47	None
Republic of Congo	1985	1	None
Seychelles	2014	6	None
Tanzania (Zanzibar)	2011	40	None

NA = Not available.

C.2 Non-randomized probability integral transform for binomial geostatistical models

This supplementary material gives detailed information on the non-randomized probability integral transform (nrPIT) as outlined by Czado et al. and modified by Giorgi et al. for binomial geostatistical models [17, 20].

Let $Y = \{Y_i; 1 =, \ldots, n\}$ denote the vector of random variables of the number of STH (any STH or species-specific) positive cases, Y_i out of n_i tested individuals at location x_i, for $i = 1, \ldots, n$. We assume that Y_i follows a Binomial distribution with probability $p(\mathbf{x}_i)$ and linear predictor

$$\log\left\{\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right\} = d(\mathbf{x}_i)^\top \beta + S(\mathbf{x}_i) + Z_i,$$

where β is the vector of regression coefficients associated with the matrix of covariates $d(\mathbf{x}_i)$. $S(\mathbf{x}_i)$ and Z_i are the Gaussian process and Gaussian noise, respectively, that have a mean of zero and variance of σ^2 and τ^2.

To outline the nrPIT we let $Q(Z)$ denote the cumulative density function of a random variable Z and Y_i^* denote the random variable of the positive tested STH (any STH or species-specific) cases at a set of hold-out locations, say $\mathbf{x}_j^*$ for $j =, \ldots, q$. Conditional on $Y = y$, the conditional cumulative probability distribution (CPD) of Y_i^* is given as:

$$Q(y_j^*|y) = P(Y_j^* \leq y_j^*|y_1, \ldots, y_n). \tag{C.1}$$

To compute Equation C.1, we first define $W = \{S(\mathbf{x}_i) + Z_i : i = 1, \ldots, n\}$ and $W_j^* = S(\mathbf{x}_j^*) + Z_j$ for $j = 1, \ldots, q$. Since it follows from the model assumptions that $Q(y_j^*|w_j, y) = Q(y_j^*|w_j)$, Equation C.1 is expressed as:

$$\begin{aligned} Q(y_j^* \mid y) &= \int_{-\infty}^{+\infty} f(w_j \mid y) Q(y_j^*|w_j, y)\, dw_j \\ &= \int_{-\infty}^{+\infty} f(w_j \mid y) Q(y_j^* \mid w_j)\, dw_j \end{aligned} \tag{C.2}$$

where $Q(y_j^*|w_j)$ is the CPD of a Binomial distribution with number of trials n_j and probability $p(x_j^*)$ and $f(w_j|y)$ is the density function of the predictive distribution of W_j. To compute the integral in Equation C.2, we then simulated $10,000$ samples from $f(w_j|y)$ as follows: 1) simulate $10,000$ samples from W conditionally on y; 2) use the resulting samples from the previous step to simulate from W_j given W, which corresponds to a multivariate Gaussian distribution. More details on this can be found in other texts [5].

Consequently, the nrPIT is defined as

$$\text{nrPIT}(u \mid y_j^*, y) = \begin{cases} 0 & \text{if } u \leq Q(y_j^* - 1 \mid y) \\ \frac{[u - Q(y_j^* - 1 \mid y)]}{[Q(y_j^* \mid y) - Q(y_j^* - 1 \mid y)]} & \text{if } Q(y_j^* - 1 \mid y) \leq u \leq Q(y_j^* \mid y) \\ 1 & \text{if } u \geq Q(y_j^* \mid y) \end{cases}. \tag{C.3}$$

We evaluate the calibration of the model by computing the average nrPIT across all counts. This is expressed as:

$$\text{nrPIT}(u) = \frac{1}{q} \sum_{j=1}^{q} \text{nrPIT}(u \mid y_j^*, y). \tag{C.4}$$

Czado et al demonstrated, assuming a well-calibrated model, that the expected value of $nrPIT(u)$ is u under the assumption of a well-calibrated model [20].

To perform the diagnostic assessment for our models, we randomly choose three subsets of locations, representing 30%, 40%, and 50% of the data-sets under consideration. The count of positive cases in these hold-out sets is denoted as y_j in the aforementioned equations.

We generated a plot of nrPIT$(j/10)$ against $j/10$ for $j = 1, \ldots, 10$ to evaluate the calibration of the models. We also generated a 95% confidence envelope using the following steps: 1) simulating 10,000 Binomial observations from the distribution of y_j^*, given $W_j = w_j$, where w_j is determined as outlined in our approximation of Equation C.4; 2) for each simulated Binomial dataset, calculate the nrPIT as defined in Equation C.4; 3) utilizing the resulting 10,000 nrPITs to compute 95%

confidence intervals for nrPIT($j/10$) at $j = 1, \ldots, 10$. This process ensures that the 95% confidence intervals are generated under the "null hypothesis" of a well-calibrated model.

C.3 Spatial covariate parameter estimates for country-level models

C.4 Estimates of log scale of spatial correlation from geostatistical models

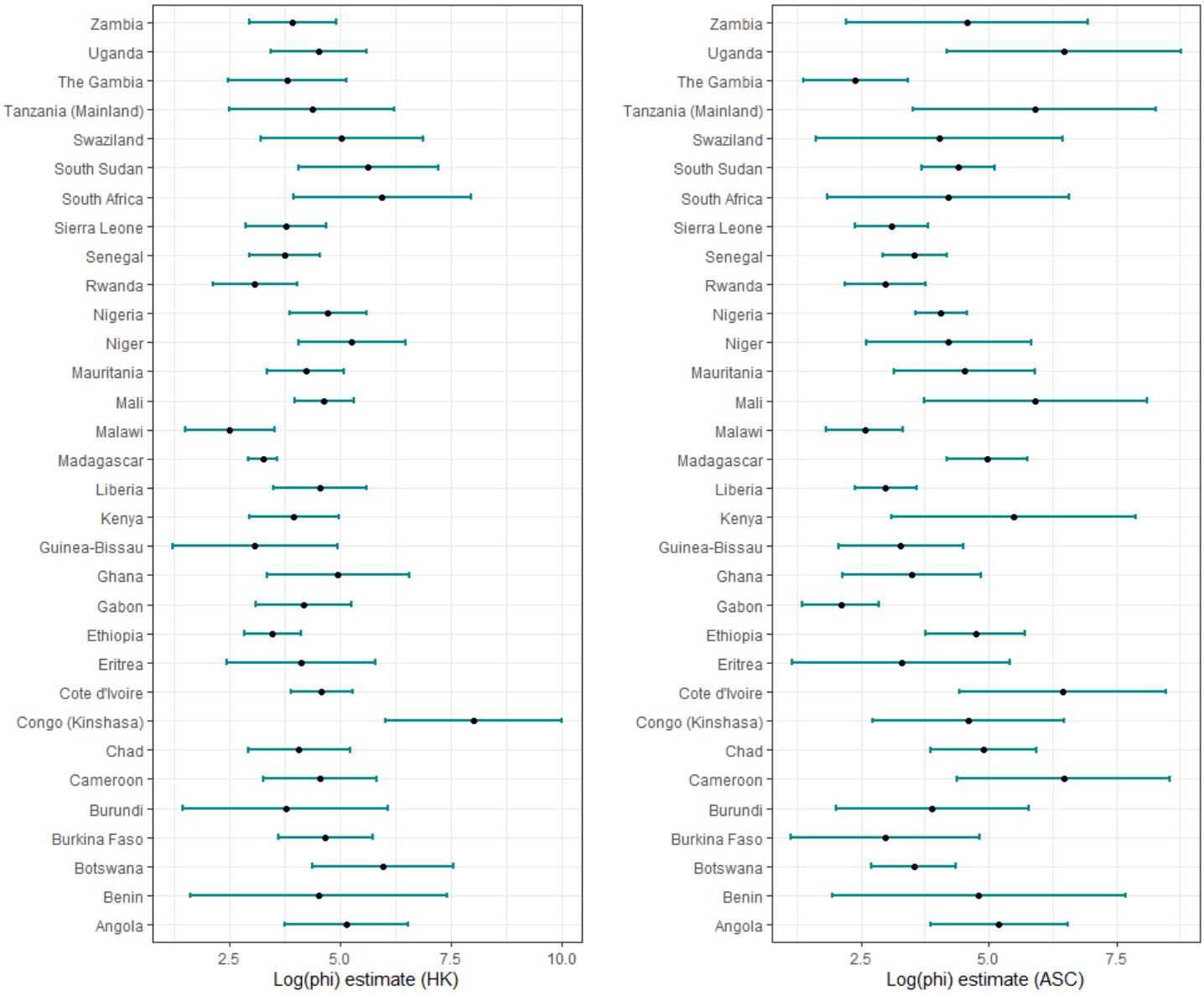

Figure C.1: Graph showing the estimated log of the scale of the spatial correlation per country for Hookworm (HK) and *Ascaris* (ASC).

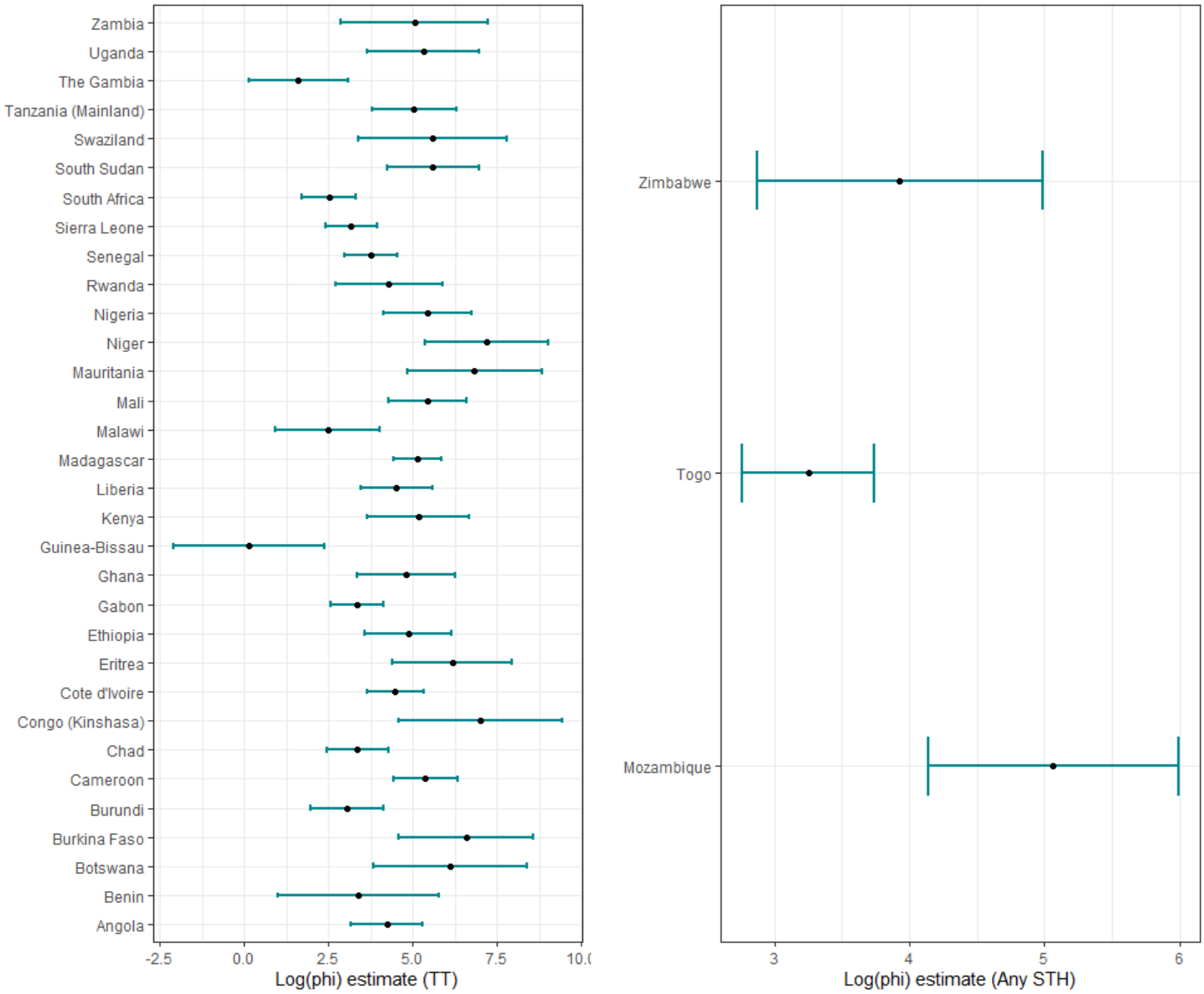

Figure C.2: Graph showing the estimated log of the scale of the spatial correlation per country for *Trichiura* (TT) and any STH (STH).

Table C.3: Summary of Monte Carlo maximum likelihood estimates of geostatistical models for all country models.

Country	**Variable parameter estimate direction**															
	Hookworm				**Ascaris**				**Trichiura**				**Any STH**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Southern Africa																
Botswana	-	N	+	N	N	N	+	N	-	N	+	N				
South Africa	N	N	N	N	N	N	N	N	N	N	N	N				
Swaziland	N	N	+	N	N	N	+	N	N	N	N	N				
Central Africa																
Angola	-	N	+	N	N	N	N	-	N	N	N	N				
Cameroon	-	N	+	-	N	N	N	N	N	N	N	-				
Chad	N	N	+	-	N	N	N	-	N	N	N	N				
DRC	N	N	N	N	N	N	+	-	N	N	+	-				
Gabon	-	N	+	-	-	N	+	N	-	N	+	-				
Eastern Africa																
Burundi	N	N	N		N	N	N	-	N	N	N	-				
Eritrea	N	N	+	N	N	N	N	N	N	N	N	N				
Ethiopia	-	N	-	N	N	N	N	N	N	N	N	-				
Kenya	N	+	N	N	N	N	+	N	N	N	N	N				
Madagascar	-	N	+	N	N	N	+	N	N	N	+	N				
Malawi	N	N	N	-	N	N	N	N	N	N	N	N				
Mozambique													N	N	+	N
Rwanda	N	N	N	-	-	N	+	-	-	N	+	-				
South Sudan	N	N	+	-	N	N	+	-	N	N	N	N				
Tanzania Mainland	N	N	+	N	N	N	+	N	N	N	+	N				
Uganda	N	N	N	N	N	N	N	N	N	N	N	N				
Zambia	N	N	+	N	N	N	N	N	N	N	N	N				
Zimbabwe													N	N	+	N
Western Africa																
Benin	-	N	N	N	N	N	N	N	N	N	N	N				
Burkina Faso	N	N	+	N	N	N	N	N	N	N	N	N				
Cote d'Ivoire	-	N	N	-	N	N	N	N	-	N	+	-				
Ghana	N	N	N	N	N	N	N	N	N	N	N	N				
Guinea-Bissau	-	N	+	N	N	N	N	N	N	N	N	N				
Liberia	N	N	N	N	N	N	N	N	-	N	N	N				
Mali	-	N	+	N	N	N	N	N	N	N	N	N				
Mauritania	N	N	+	N	N	N	N	N	N	N	N	N				
Niger	N	N	+	N	N	N	N	N	N	N	N	N				
Nigeria	N	+	+	-	N	N	+	N	N	N	+	-				
Senegal	N	N	+	N	-	N	N	N	N	N	N					
Sierra Leone	N	N	+	N	N	N	+	N	N	N	+	-				
Togo													-	N	N	N

Covariate 1 = Precipitation; 2 = Poverty index; 3 = Precipitation or Aridity index; 4 = Soil type (clay, sand, or silt) or soil PH.
+ = Positive association; - = Negative association; N = Not included in the model.
DRC = Democratic Republic of the Congo (Congo Kinshasa).

C.5 Estimates of variance of spatial correlation from geostatistical models

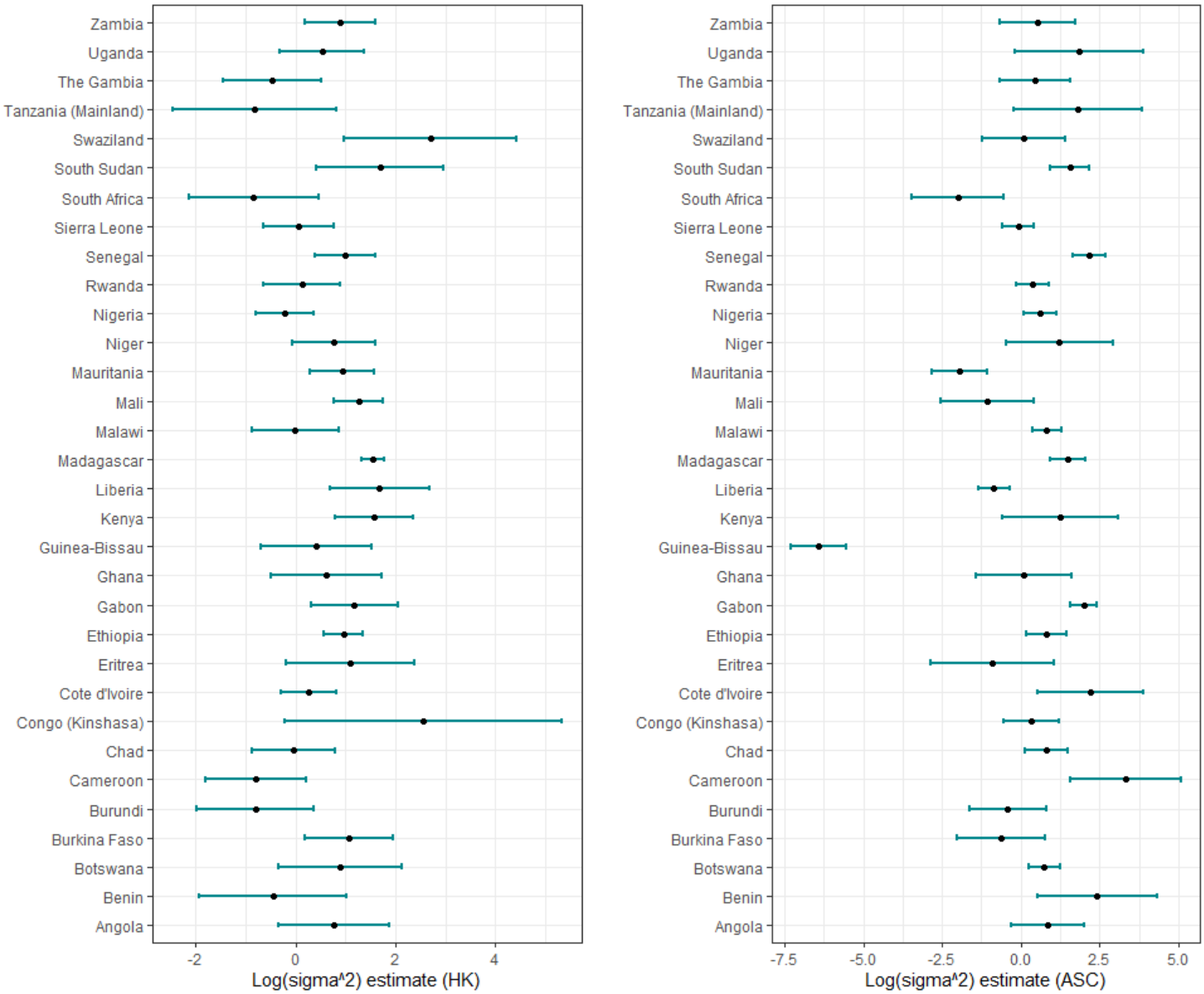

Figure C.3: Graph showing the estimated log of variance of spatial correlation per country for Hookworm (HK) and Ascaris (ASC).

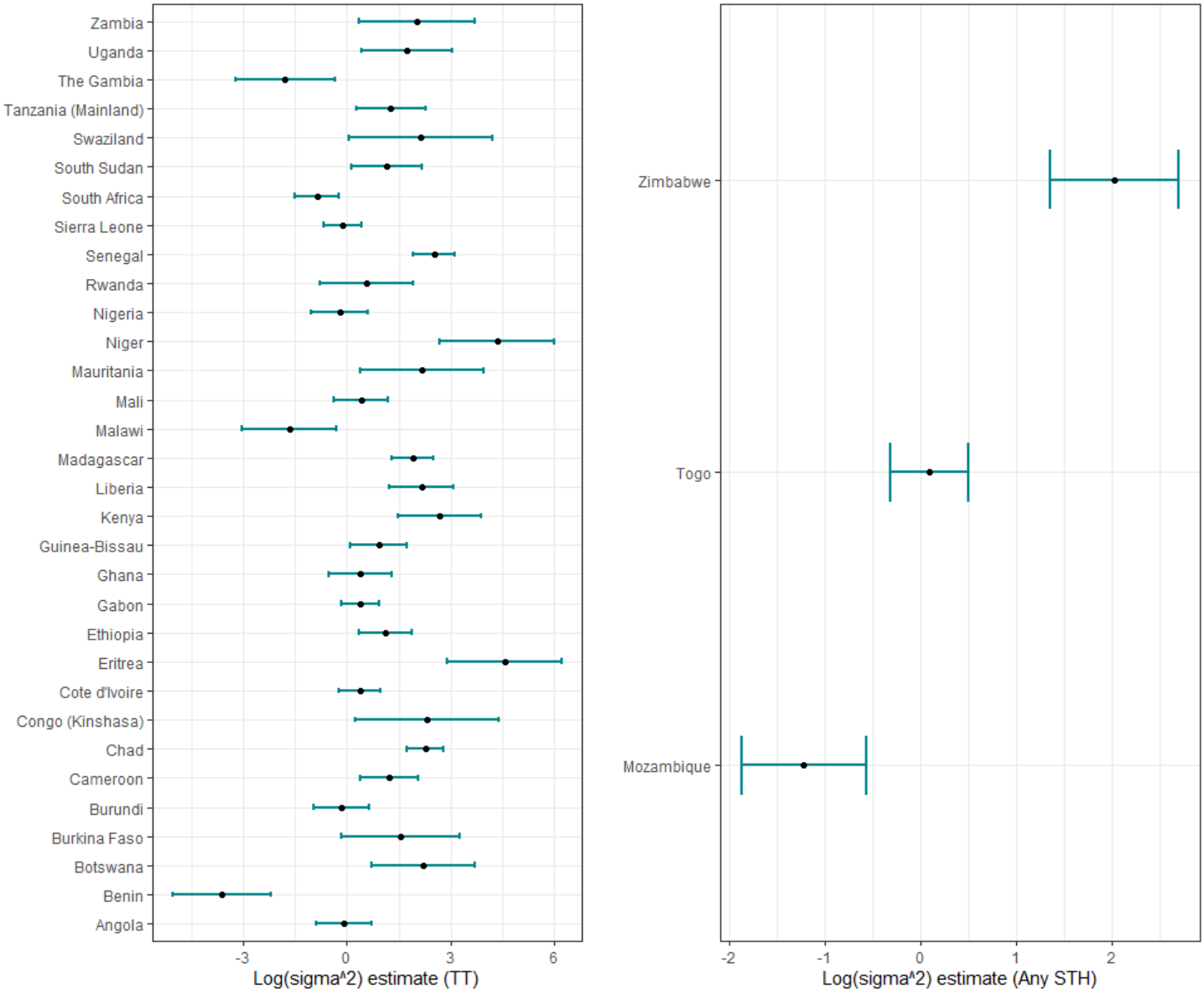

Figure C.4: Graph showing the estimated log of variance of spatial correlation per country for Trichiura (TT) and any STH (STH).

D Paper 4 Supplementary Material

D.1 Exploratory analysis and variable processing

Figure D.1 shows scatter plots of the log of dengue incidence against several spatial covariates. The plot shows that most of the covariates had a linear relationship with the incidence of dengue.

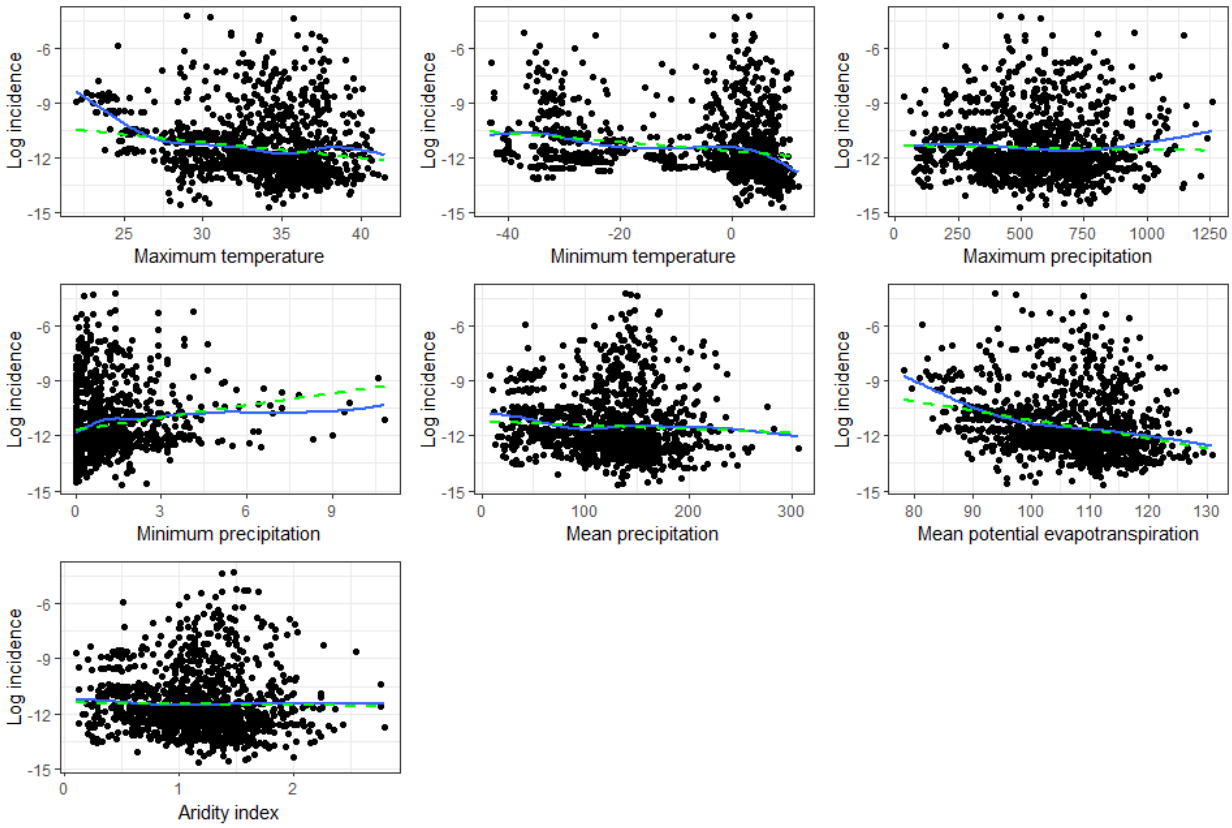

Figure D.1: Scatter plots of the log incidence against maximum temperature, minimum temperature, maximum precipitation, minimum precipitation, mean precipitation, mean evapotranspiration, and aridity index. The dashed green lines are regression lines from a linear model, whilst the blue solid lines are natural splines from a generalized additive model.

D.2 Principal components analysis further results

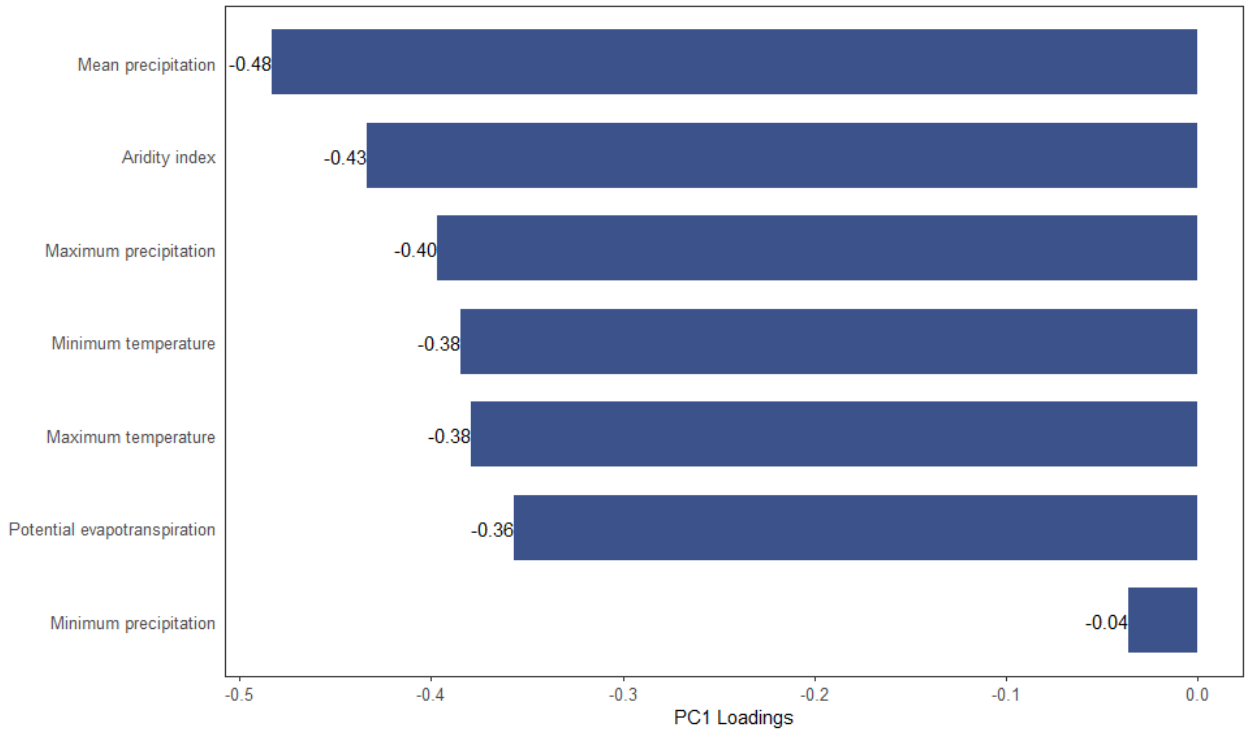

Figure D.2: Loadings of the environmental exposure index (PC1)

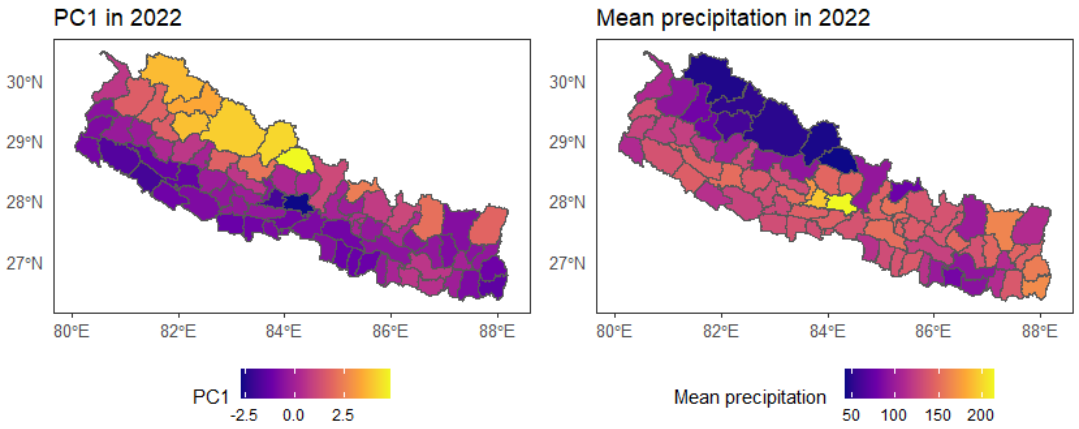

Figure D.3: Maps of the environmental exposure index (PC1) and the mean precipitation

D.3 Model coefficients

Table D.1: Maximum likelihood estimates and 95% confidence intervals (CI) for the model intercept

District	Coefficient	95% CI
Achham	-26.392	(-71.047, -22.371)
Arghakhanchi	-24.923	(-31.303, -16.090)
Baglung	-36.457	(-71.001, -27.781)
Baitadi	-27.962	(-71.892, -20.275)
Bajhang	-26.285	(-167.286, -13.336)
Bajura	-56.729	(-102.784, -19.323)
Banke	-22.959	(-32.592, -16.340)
Bara	-12.697	(-26.281, 1.688)
Bardiya	-12.664	(-22.008, -11.539)
Bhaktapur	-14.554	(-15.469, -13.556)
Bhojpur	-22.840	(-29.013, -14.345)
Chitawan	-22.018	(-34.056, -20.396)
Dadeldhura	-21.429	(-33.035, -17.199)
Dailekh	-14.926	(-19.450, -13.927)
Dang	-12.518	(-16.152, -11.123)
Darchula	-28.625	(-47.768, -17.642)
Dhading	-13.247	(-13.599, -13.084)
Dhankuta	-19.164	(-26.134, -12.863)
Dhanusha	-19.007	(-29.295, -5.354)
Dolakha	-13.081	(-17.572, 19.743)
Dolpa	-1.855	(-3.005, -0.394)
Doti	-34.106	(-87.332, -28.268)
Gorkha	-13.010	(-18.319, 15.282)
Gulmi	-14.492	(-20.189, -14.051)
Humla	-3.011	(-9.244, -0.561)
Ilam	-19.930	(-37.304, -3.101)
Jajarkot	-21.528	(-25.277, 32.679)
Jhapa	-14.343	(-84.094, 18.870)
Jumla	-1.352	(-2.082, 30.528)

Continued on next page

Table D.1 – continued from previous page

District	Coefficient	95% CI
Kabhrepalanchok	-27.492	(-45.998, -16.414)
Kailali	-22.412	(-41.168, -13.322)
Kalikot	-5.655	(-9.571, 80.214)
Kanchanpur	-21.452	(-70.935, -12.999)
Kapilbastu	-13.745	(-16.502, -10.140)
Kaski	-32.152	(-103.581, -17.967)
Kathmandu	-14.744	(-14.767, -14.299)
Khotang	-32.853	(-149.562, -25.717)
Lalitpur	-43.326	(-61.470, -27.795)
Lamjung	-20.175	(-20.684, -19.176)
Mahottari	-15.759	(-26.577, 0.702)
Makawanpur	-13.441	(-53.896, -11.295)
Manang	-0.817	(-1.431, -0.418)
Morang	-12.633	(-23.798, -11.503)
Mugu	-4.440	(-9.940, -2.010)
Mustang	-1.214	(-1.833, 0.530)
Myagdi	-5.338	(-30.846, 91.048)
Nawalparasi east	-12.756	(-13.194, -12.170)
Nawalparasi west	-27.594	(-101.242, -12.256)
Nuwakot	-18.278	(-19.277, -17.625)
Okhaldhunga	-21.315	(-26.206, -19.128)
Palpa	-47.780	(-48.779, -41.718)
Panchthar	-27.078	(-32.684, -17.951)
Parbat	-21.951	(-25.983, -19.871)
Parsa	-10.758	(-12.892, -2.452)
Pyuthan	-15.376	(-17.312, -13.409)
Ramechhap	-29.053	(-46.061, -17.717)
Rasuwa	-8.343	(-193.609, 14.709)
Rautahat	-94.858	(-95.857, -88.943)
Rolpa	-20.056	(-20.350, -19.715)
Rukum east	-5.874	(-45.025, -1.945)

Continued on next page

Table D.1 – continued from previous page

District	Coefficient	95% CI
Rukum west	-19.459	(-22.246, -7.492)
Rupandehi	-13.012	(-27.539, -10.175)
Salyan	-41.266	(-100.122, -25.066)
Sankhuwasabha	-22.294	(-23.293, -21.881)
Saptari	-20.884	(-23.439, -16.816)
Sarlahi	-17.917	(-33.761, 11.605)
Sindhuli	-36.405	(-98.222, -27.759)
Sindhupalchok	-22.565	(-26.402, -20.855)
Siraha	-29.085	(-38.468, -20.677)
Solukhumbu	-6.841	(-11.462, 17.704)
Sunsari	-13.855	(-19.845, -6.911)
Surkhet	-30.796	(-57.925, -17.654)
Syangja	-19.976	(-25.855, -8.631)
Tanahu	-22.164	(-31.211, -8.496)
Taplejung	-43.563	(-147.470, -10.099)
Terhathum	-21.978	(-24.589, -13.708)
Udayapur	-29.071	(-47.358, -18.067)

Table D.2: Maximum likelihood estimates and 95% confidence intervals (CI) for the environmental exposure index

District	Coefficient	95% CI
Achham	6.090	(-3.092, 23.019)
Arghakhanchi	4.005	(-0.537, 25.220)
Baglung	3.165	(-4.419, 31.380)
Baitadi	4.173	(-3.500, 15.464)
Bajhang	-6.502	(-18.452, 46.743)
Bajura	11.351	(8.692, 12.349)
Banke	-5.605	(-10.122, -1.467)
Bara	0.630	(-15.060, 15.913)

Continued on next page

Table D.2 – continued from previous page

District	Coefficient	95% CI
Bardiya	0.067	(-6.033, 0.910)
Bhaktapur	0.031	(-46.820, 9.798)
Bhojpur	-2.791	(-11.547, 2.644)
Chitawan	-0.256	(-0.654, 0.218)
Dadeldhura	5.180	(1.908, 13.620)
Dailekh	-1.722	(-17.886, 7.058)
Dang	-0.495	(-1.862, 1.038)
Darchula	-0.630	(-10.311, 23.257)
Dhading	0.157	(-9.625, 1.769)
Dhankuta	1.262	(-33.149, 18.600)
Dhanusha	-2.811	(-9.699, 2.919)
Dolakha	-7.921	(-32.482, 28.303)
Dolpa	-7.904	(-10.420, -6.046)
Doti	11.897	(7.387, 14.283)
Gorkha	-0.739	(-17.191, 19.846)
Gulmi	-0.171	(-8.858, 7.310)
Humla	-9.404	(-16.300, -5.989)
Ilam	-0.198	(-8.792, 6.945)
Jajarkot	-2.822	(-9.709, 27.599)
Jhapa	0.412	(-2.648, 19.271)
Jumla	-6.252	(-13.175, 51.717)
Kabhrepalanchok	-3.612	(-15.205, 4.498)
Kailali	0.688	(-2.244, 5.430)
Kalikot	-10.319	(-13.276, -7.989)
Kanchanpur	0.681	(-3.655, 1.832)
Kapilbastu	0.870	(-0.933, 10.756)
Kaski	-0.972	(-9.832, 6.812)
Kathmandu	0.556	(-9.881, 13.769)
Khotang	1.135	(-6.306, 4.734)
Lalitpur	-10.496	(-25.884, 4.987)
Lamjung	-6.684	(-39.037, 4.073)

Continued on next page

Table D.2 – continued from previous page

District	Coefficient	95% CI
Mahottari	-0.026	(-10.434, 2.170)
Makawanpur	-0.092	(-6.853, 0.092)
Manang	-5.296	(-10.281, 26.249)
Morang	2.759	(-9.434, 11.937)
Mugu	-12.214	(-15.065, -7.575)
Mustang	-4.237	(-5.062, -3.985)
Myagdi	-10.290	(-68.550, 2.108)
Nawalparasi east	0.799	(-14.606, 2.930)
Nawalparasi west	-0.124	(-3.189, 2.412)
Nuwakot	2.057	(-19.362, 4.625)
Okhaldhunga	0.508	(-5.440, 12.609)
Palpa	2.536	(0.520, 5.069)
Panchthar	0.047	(-21.495, 25.083)
Parbat	-0.577	(-5.815, 2.529)
Parsa	1.301	(-1.358, 8.354)
Pyuthan	-0.588	(-9.580, 6.195)
Ramechhap	-13.087	(-45.541, 4.442)
Rasuwa	-9.834	(-15.485, 70.465)
Rautahat	3.486	(-2.233, 34.317)
Rolpa	-0.805	(-7.724, 4.560)
Rukum east	-8.539	(-8.952, -7.915)
Rukum west	-0.986	(-8.012, 3.504)
Rupandehi	0.485	(-0.090, 7.376)
Salyan	-10.822	(-31.339, -0.862)
Sankhuwasabha	-3.671	(-26.161, 5.387)
Saptari	-0.867	(-24.084, 41.372)
Sarlahi	-1.368	(-20.824, 39.783)
Sindhuli	-2.897	(-9.315, 8.738)
Sindhupalchok	-9.178	(-35.067, 10.357)
Siraha	-8.216	(-14.846, 1.140)
Solukhumbu	-11.888	(-20.356, 31.769)

Continued on next page

Table D.2 – continued from previous page

District	Coefficient	95% CI
Sunsari	0.389	(-6.024, 25.843)
Surkhet	-10.952	(-23.017, 4.812)
Syangja	0.121	(0.095, 0.176)
Tanahu	-1.007	(-5.255, -0.377)
Taplejung	6.255	(-10.130, 36.060)
Terhathum	2.071	(-11.263, 48.722)
Udayapur	-13.073	(-31.939, 1.972)

Figures D.4, D.5, and D.6 illustrate the predicted timing (μ) of the three outbreaks and their confidence intervals (CIs).

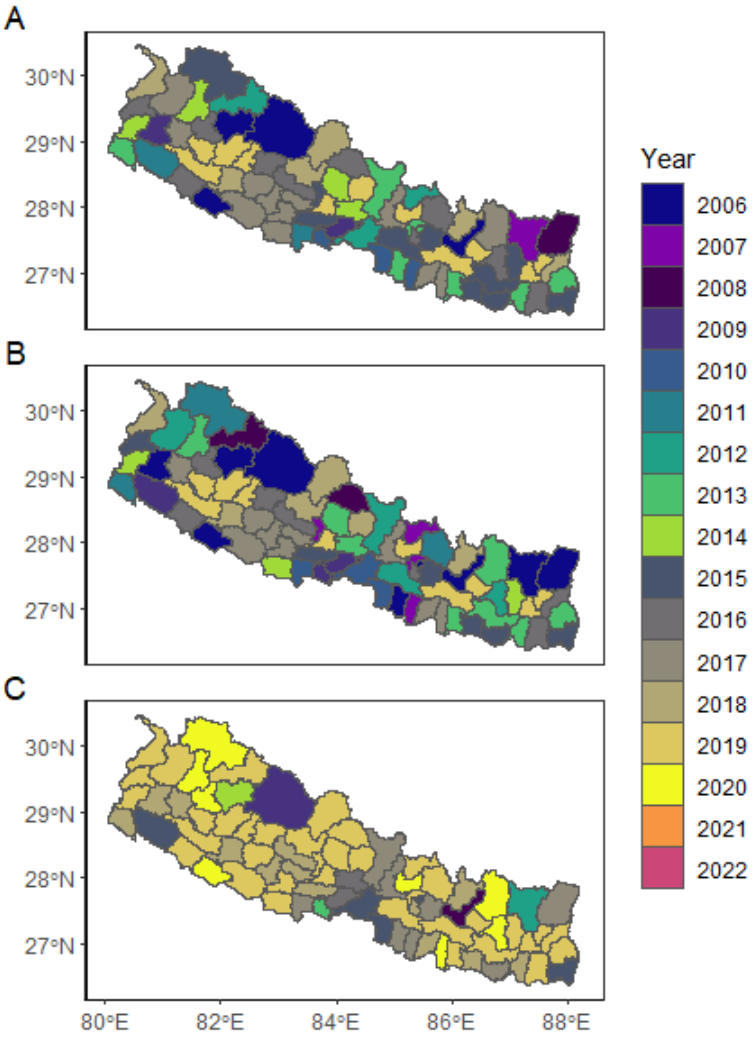

Figure D.4: Maps showing the timing of the first outbreak (μ_1, A), and the lower (B) and upper (C) bounds of the 95% CIs.

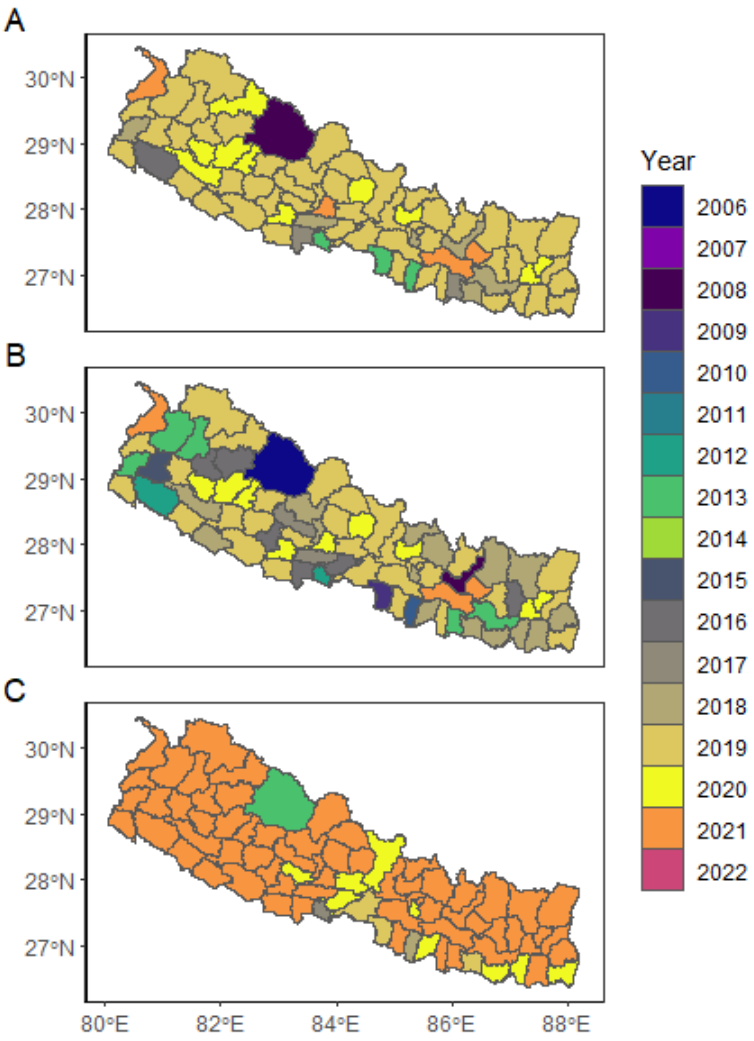

Figure D.5: Maps showing the timing of the second outbreak (μ_2, A), and the lower (B) and upper (C) bounds of the 95% CIs.

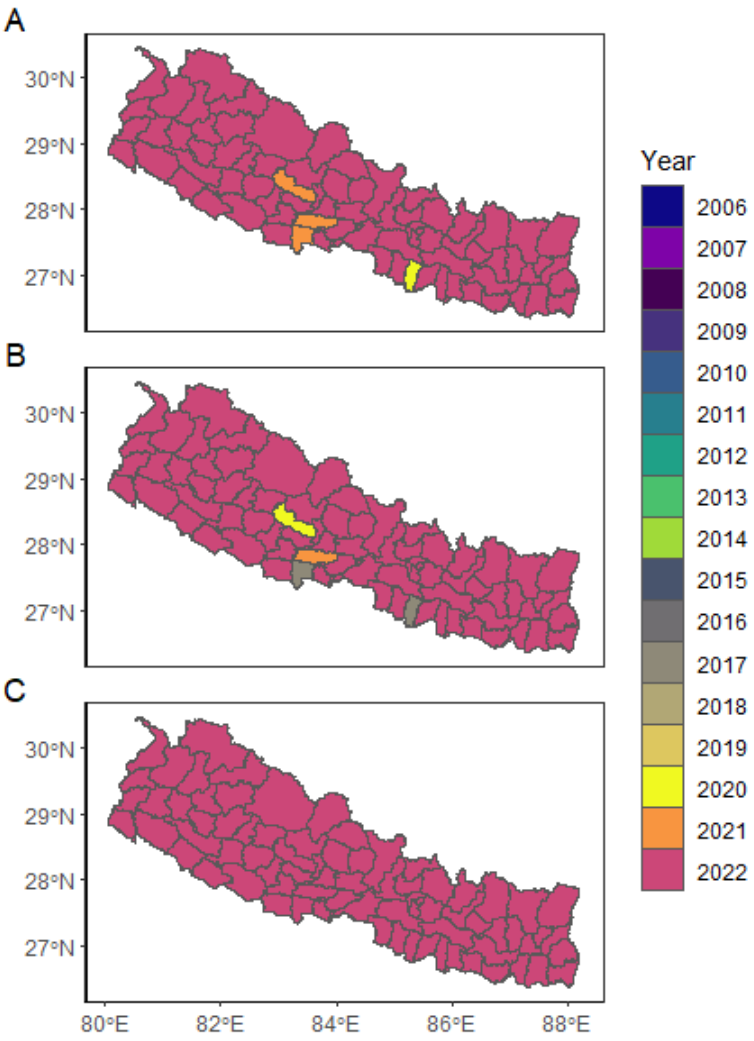

Figure D.6: Maps showing the timing of the third outbreak (μ_3, A), and the lower (B) and upper (C) bounds of the 95% CIs.

Figures D.7, D.8, and D.9 illustrate the predicted scale parameters (ω) of the three outbreak intensity functions (OIFs) and their confidence intervals (CIs).

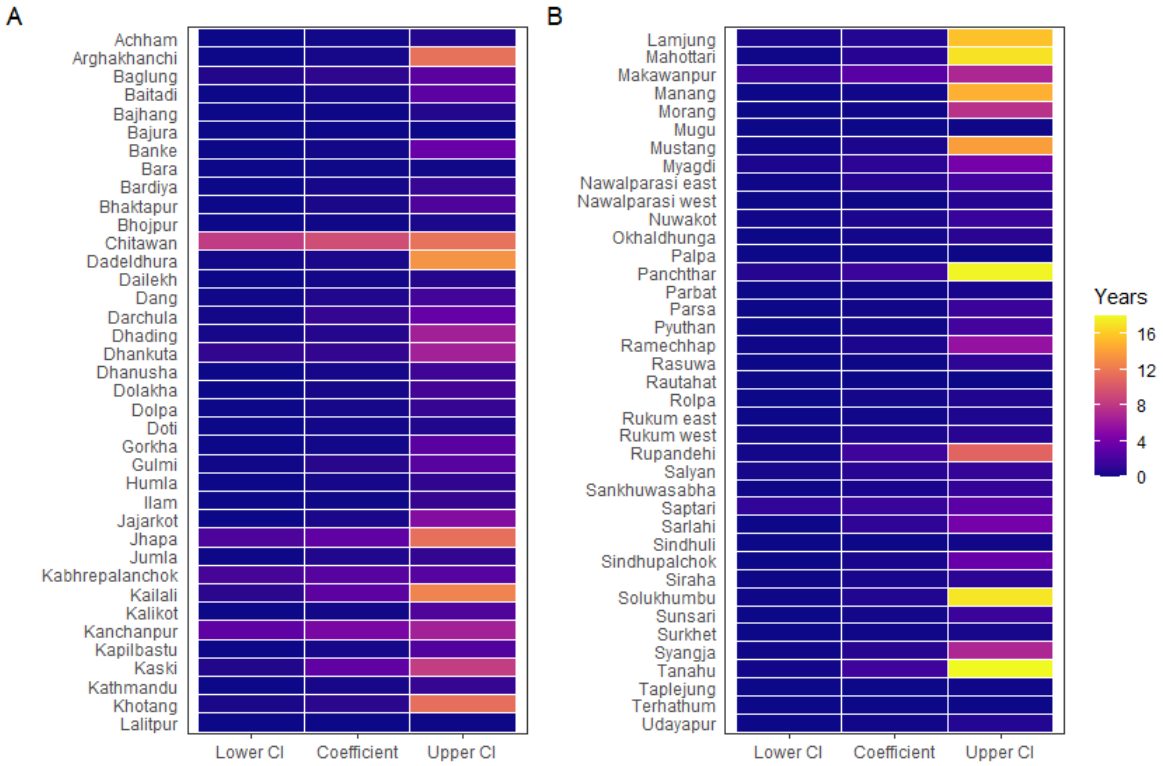

Figure D.7: Heat maps showing the scale parameter of the first OIF (ω_1), and the lower and upper CIs.

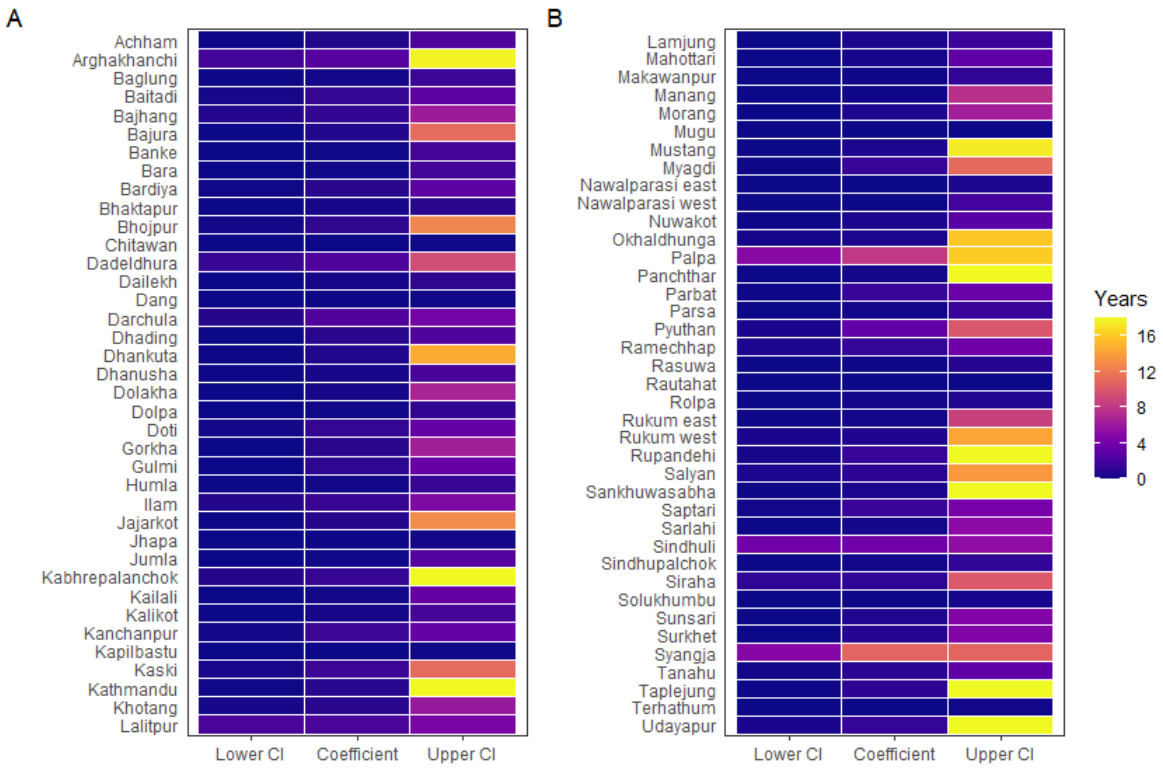

Figure D.8: Heat maps showing the scale parameter of the second OIF (ω_2), and the lower and upper CIs.

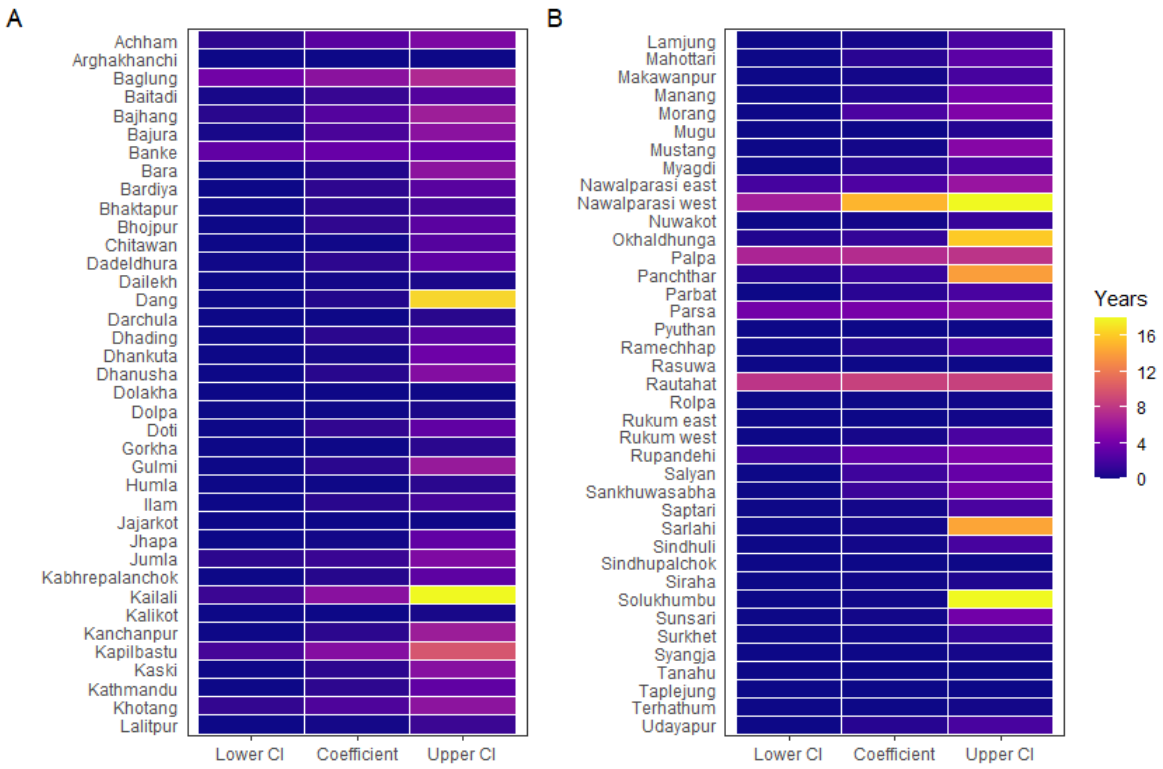

Figure D.9: Heat maps showing the scale parameter of the third OIF (ω_3), and the lower and upper CIs.

Figures D.10, D.11, and D.12 illustrate the estimated coefficients (γ) of the three outbreak intensity functions (OIFs) and their confidence intervals (CIs).

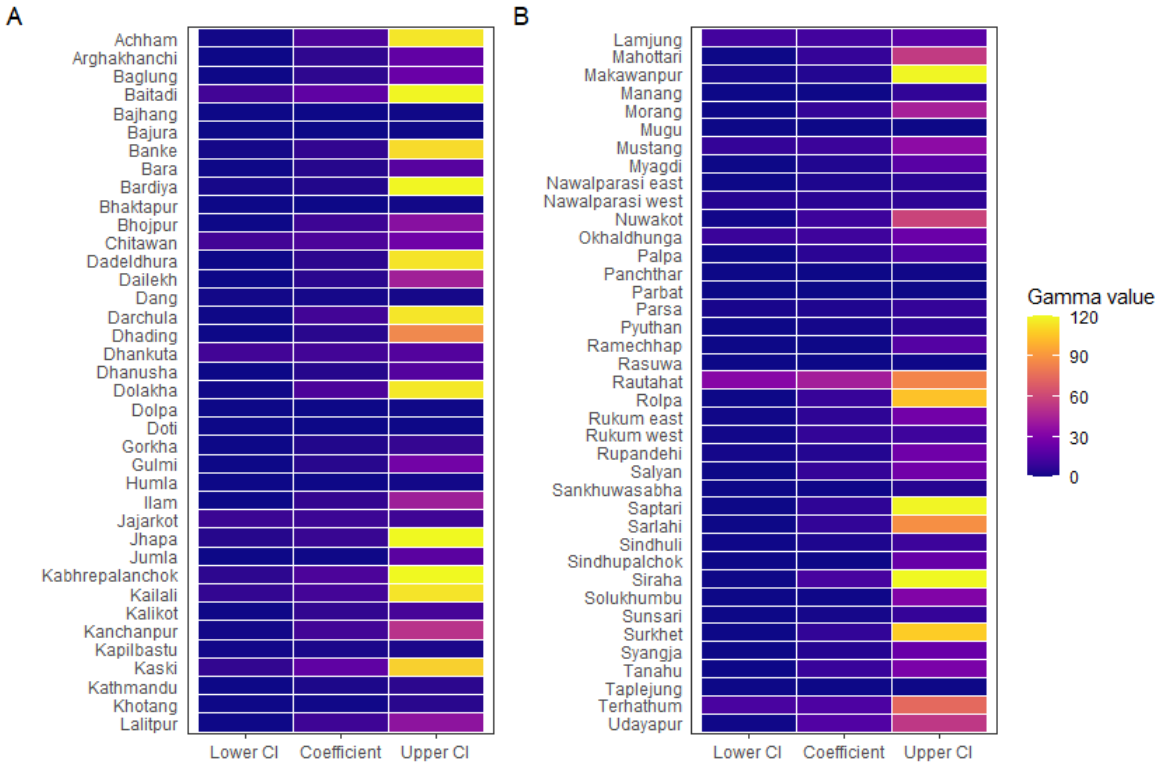

Figure D.10: Heat maps showing the estimated coefficient of the first OIF (γ_1), and the lower and upper CIs.

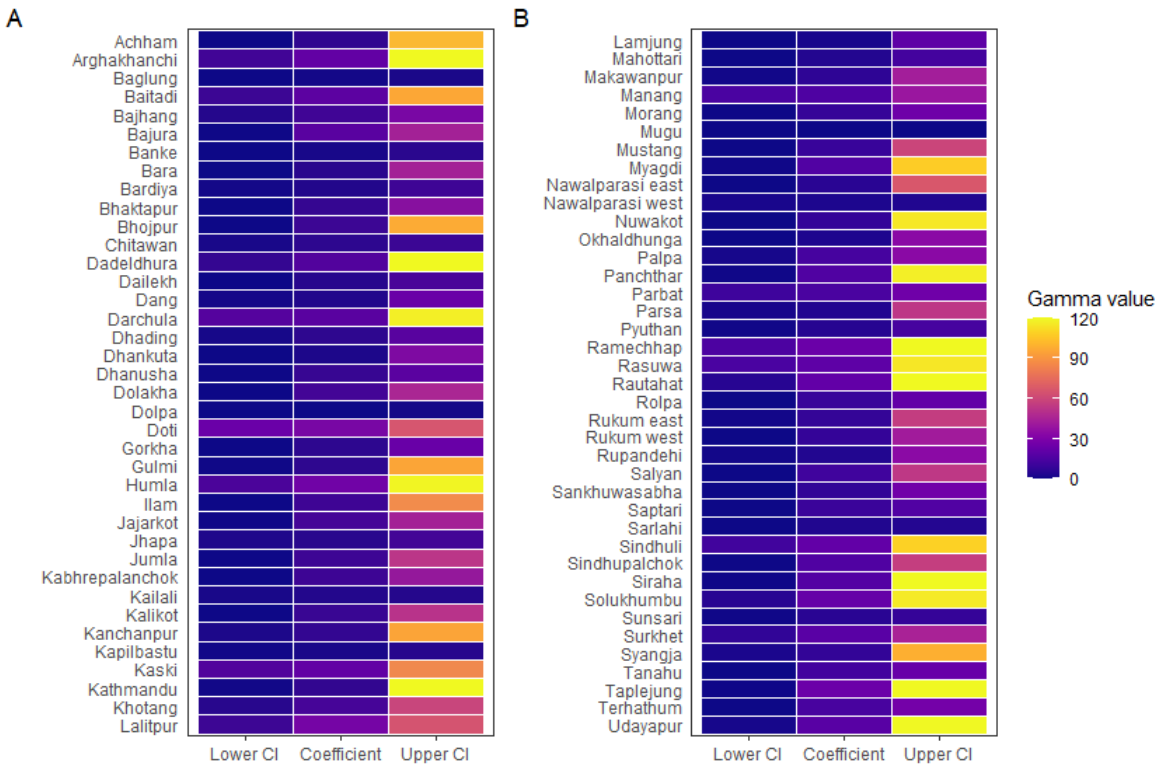

Figure D.11: Heat maps showing the estimated coefficient of the second OIF (γ_2), and the lower and upper CIs.

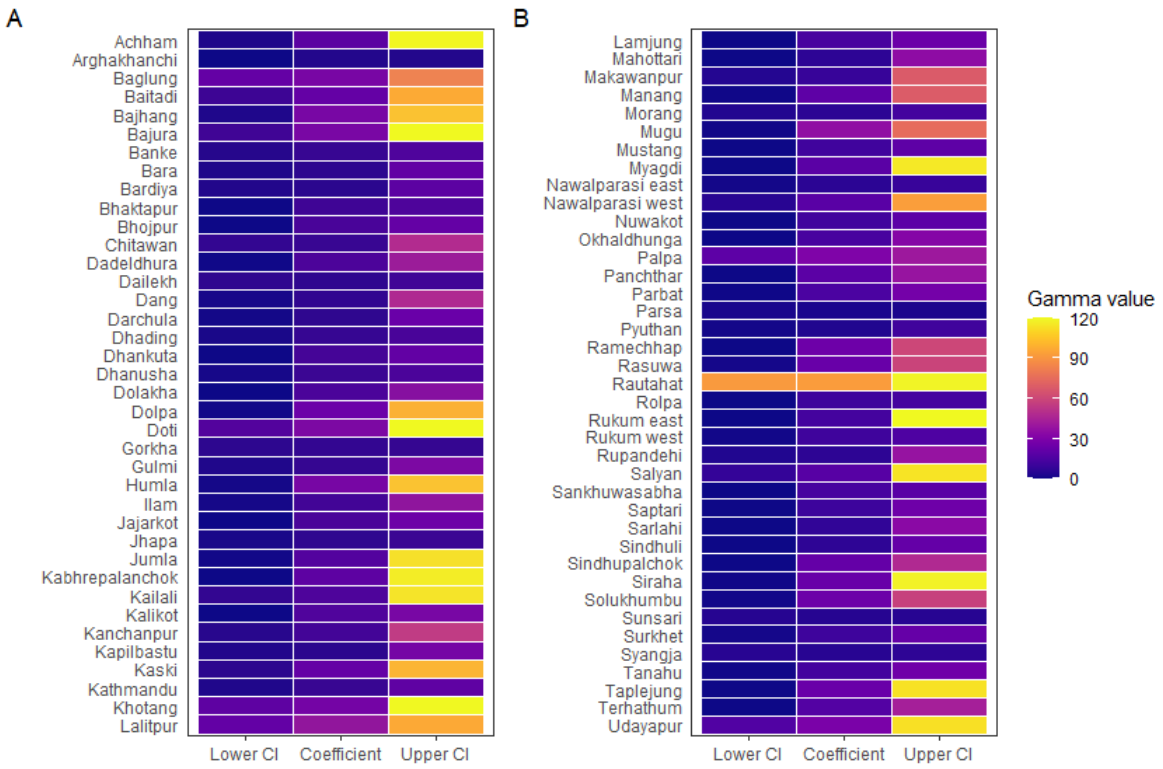

Figure D.12: Heat maps showing the estimated coefficient of the third OIF (γ_3), and the lower and upper CIs.

D.4 Model validation

We carried out a Chi-square (χ^2) goodness of fit test for each of the models fitted to each district. The goodness of fit test was carried out under the null hypothesis that there was no statistically significant difference between the expected and predicted dengue counts. Each of the tests had 16 degrees of freedom (df = n-1 = 17-1) Table D.3 shows the Chi-square statistic and *p-values* for each district.

Table D.3: Summary of Chi-square (χ^2) goodness of fit test results for the 77 Nepalese districts

District	**χ^2 Statistic**	**P value**
Achham	< 0.001	1.000
Arghakhanchi	0.0303	1.000
Baglung	0.0429	1.000
Baitadi	< 0.001	1.000
Bajhang	< 0.001	1.000
Bajura	< 0.001	1.000
Banke	48.0402	< 0.001
Bara	2.8748	0.942
Bardiya	2.7853	0.986
Bhaktapur	1.1521	0.8859
Bhojpur	< 0.001	1.000
Chitawan	1569.247	< 0.001
Dadeldhura	0.0031	1.000
Dailekh	0.7167	0.8693
Dang	10.2822	0.6707
Darchula	< 0.001	1.000
Dhading	3.3669	0.948
Dhankuta	< 0.001	1.000
Dhanusha	0.5548	0.968
Dolakha	< 0.001	1.000
Dolpa	< 0.001	1.000
Doti	< 0.001	1.000
Gorkha	1.7128	0.974

Continued on next page

Table D.3 – continued from previous page

District	χ^2 Statistic	P value
Gulmi	1.2785	0.989
Humla	< 0.001	1.000
Ilam	< 0.001	1.000
Jajarkot	< 0.001	1.000
Jhapa	1605.370	< 0.001
Jumla	< 0.001	1.000
Kabhrepalanchok	0.3878	0.9989
Kailali	241.808	< 0.001
Kalikot	< 0.001	1.000
Kanchanpur	2.6436	0.9886
Kapilbastu	57.0414	< 0.001
Kaski	1.7006	0.9889
Kathmandu	2.9773	0.9652
Khotang	< 0.001	1.000
Lalitpur	0.0888	0.9999
Lamjung	< 0.001	1.000
Mahottari	0.7551	0.9932
Makawanpur	72.3325	< 0.001
Manang	< 0.001	1.000
Morang	6.3734	0.4969
Mugu	< 0.001	0.9995
Mustang	< 0.001	1.000
Myagdi	< 0.001	1.000
Nawalparasi east	3.3502	0.8508
Nawalparasi west	166.9342	< 0.001
Nuwakot	< 0.001	1.000
Okhaldhunga	< 0.001	1.000
Palpa	10.3832	0.1679
Panchthar	< 0.001	1.000
Parbat	< 0.001	1.000
Parsa	115.3268	< 0.001

Continued on next page

Table D.3 – continued from previous page

District	χ^2 **Statistic**	**P value**
Pyuthan	5.5307	0.6996
Ramechhap	0.0019	1.000
Rasuwa	< 0.001	1.000
Rautahat	69.1356	< 0.001
Rolpa	< 0.001	1.000
Rukum east	< 0.001	1.000
Rukum west	< 0.001	1.000
Rupandehi	1986.736	< 0.001
Salyan	< 0.001	1.000
Sankhuwasabha	< 0.001	1.000
Saptari	0.1445	0.9996
Sarlahi	10.536	0.1602
Sindhuli	< 0.001	1.000
Sindhupalchok	< 0.001	1.000
Siraha	< 0.001	1.000
Solukhumbu	< 0.001	1.000
Sunsari	1.9479	0.9244
Surkhet	1.9622	0.8544
Syangja	1.1601	0.9918
Tanahu	0.3106	0.9974
Taplejung	< 0.001	1.000
Terhathum	< 0.001	1.000
Udayapur	< 0.001	1.000

The figures below show the plots of the observed vs predicted counts of dengue cases in each district.

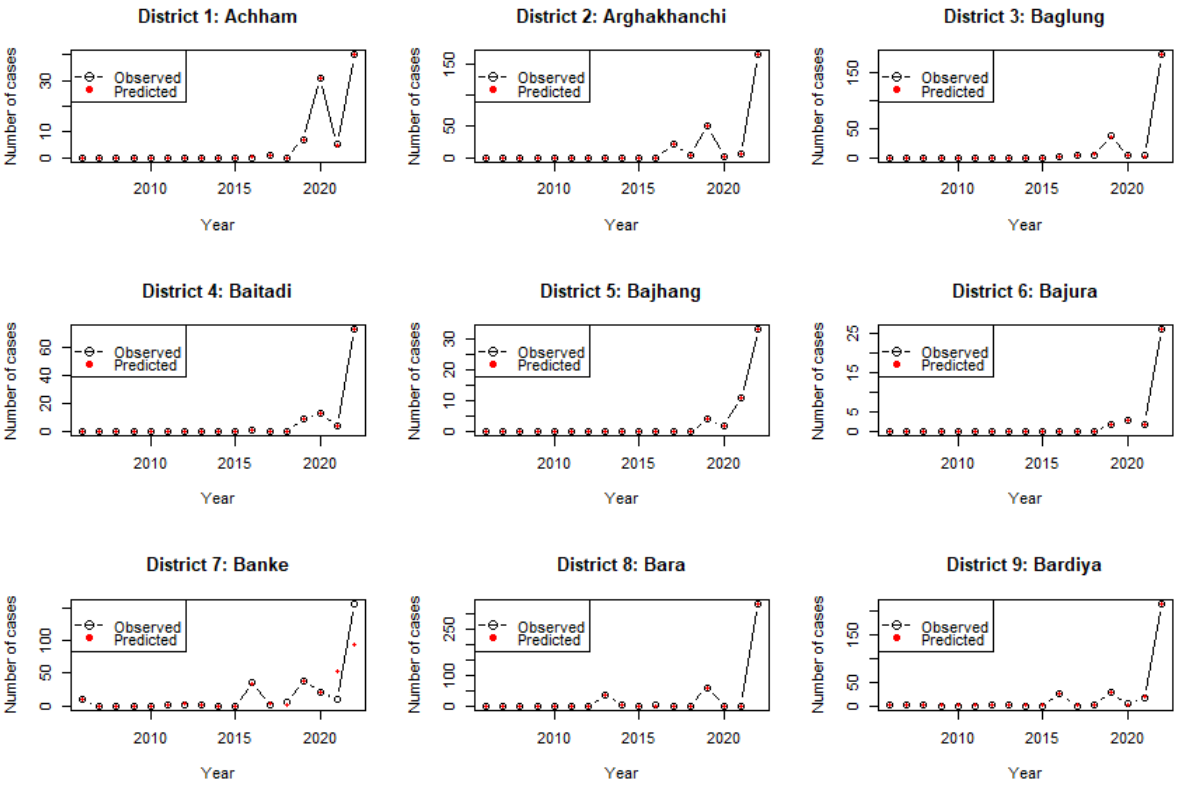

Figure D.13: Plots of the observed vs predicted counts of dengue in districts 1 to 9

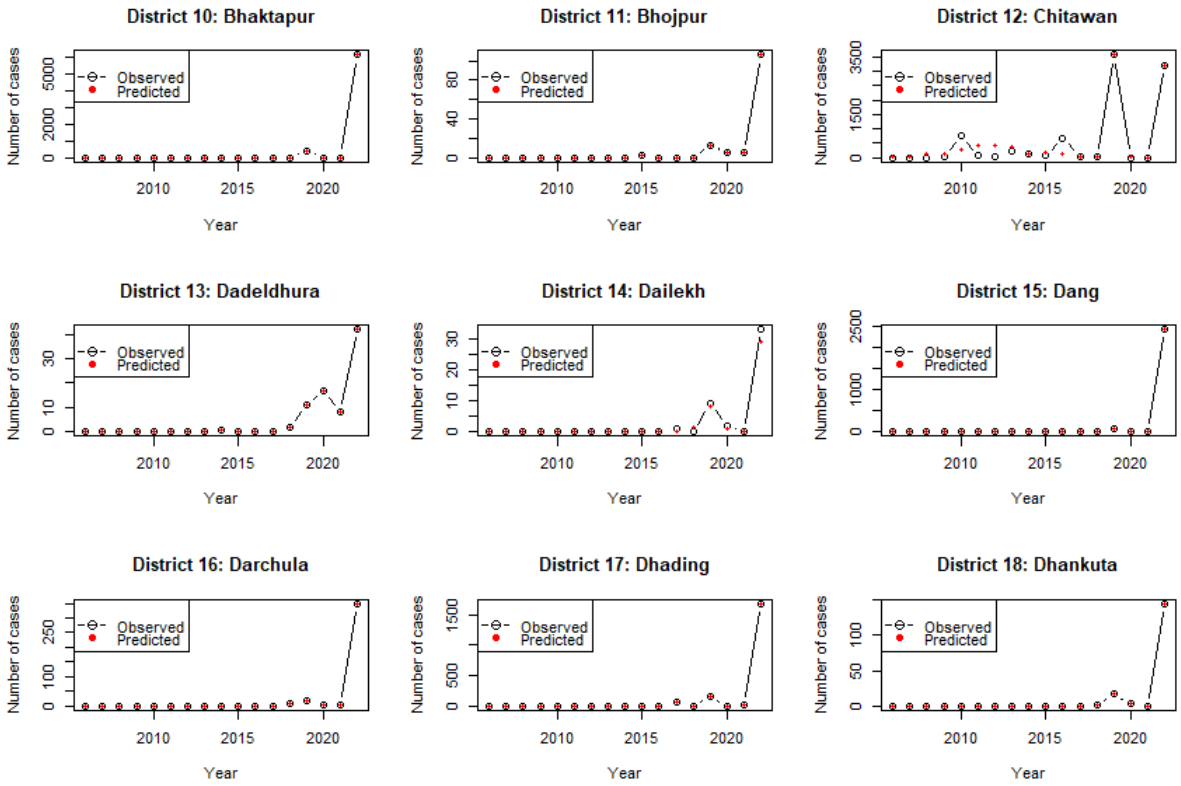

Figure D.14: Plots of the observed vs predicted counts of dengue in districts 10 to 18

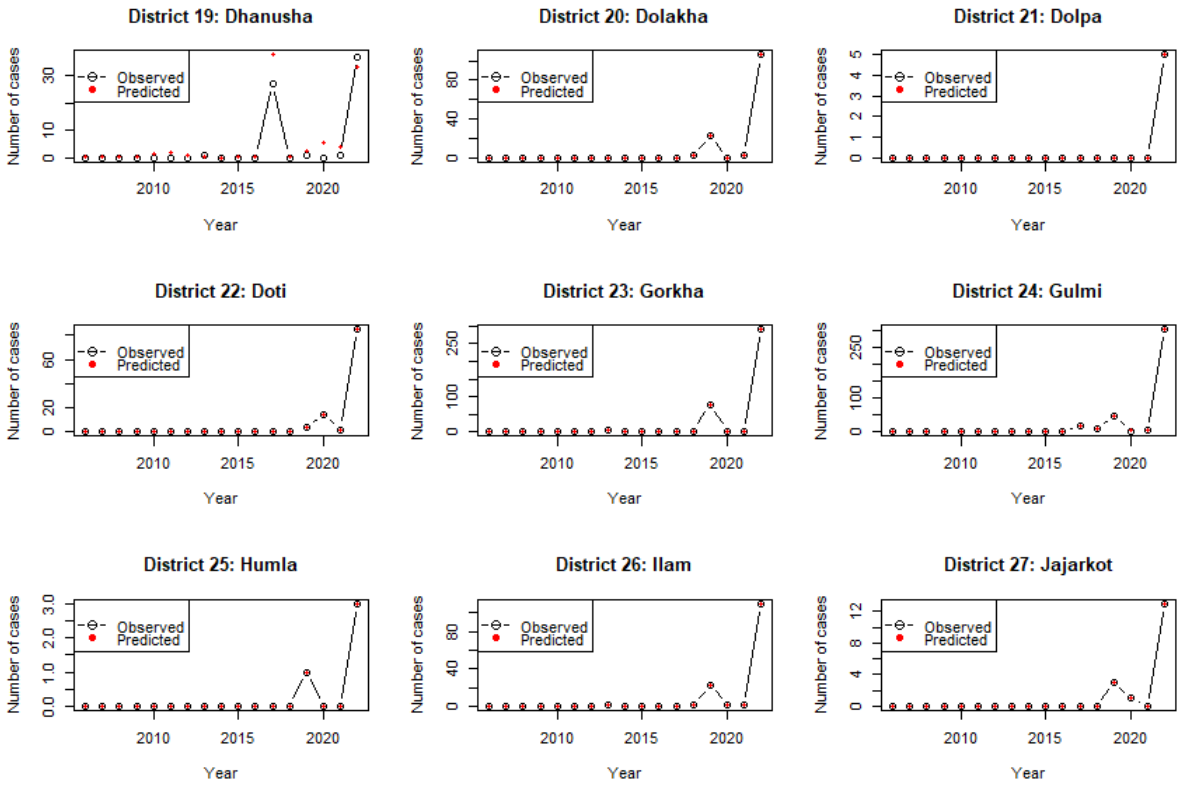

Figure D.15: Plots of the observed vs predicted counts of dengue in districts 19 to 27

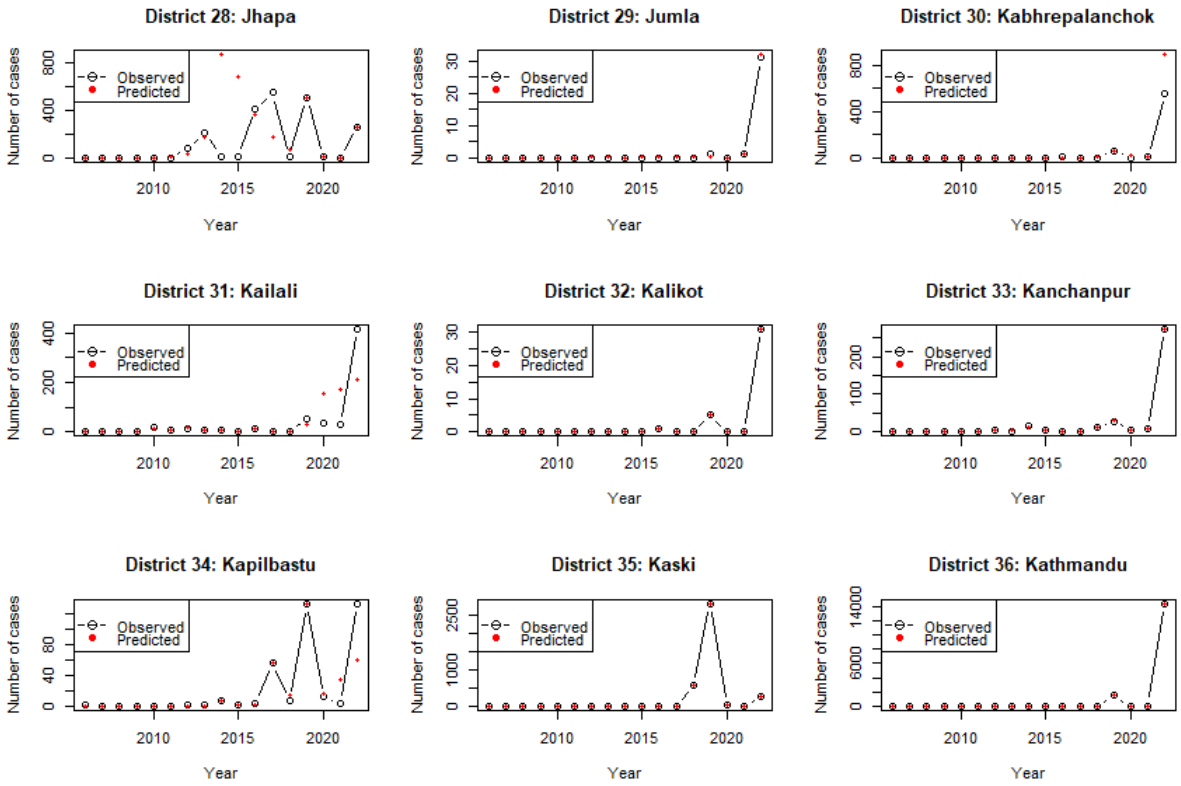

Figure D.16: Plots of the observed vs predicted counts of dengue in districts 28 to 36

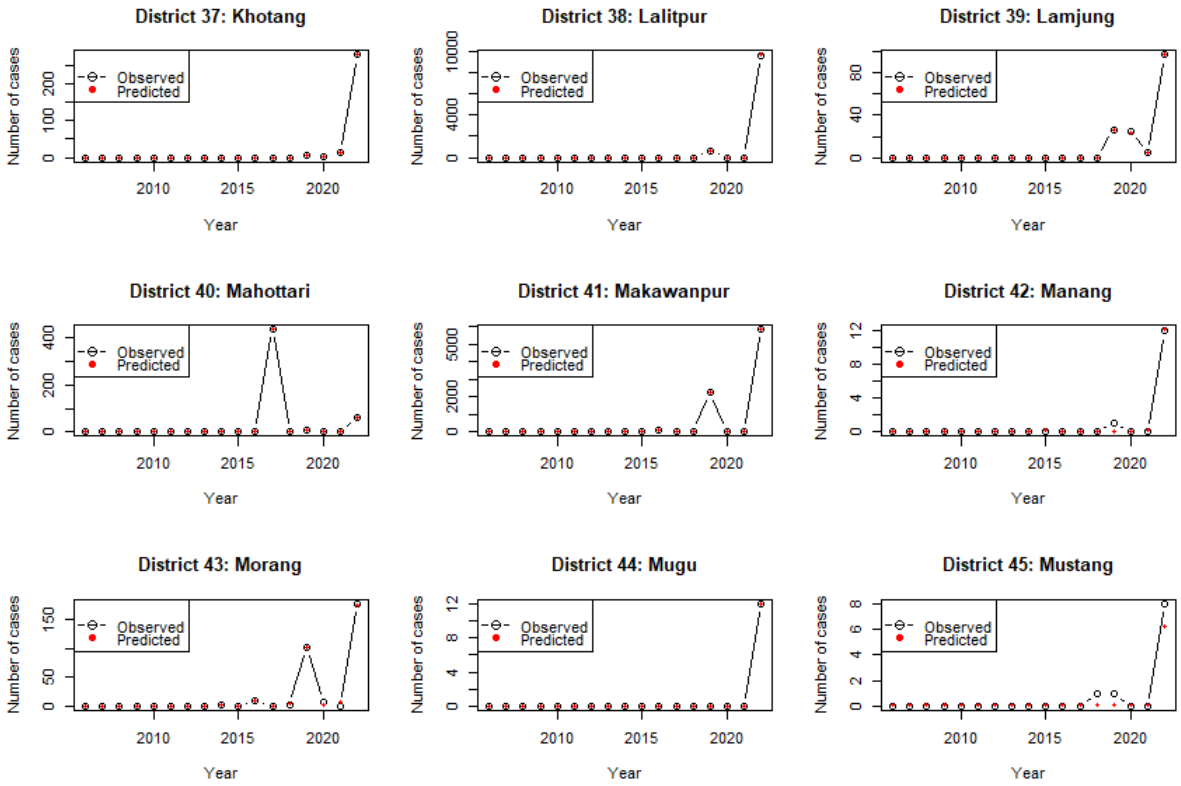

Figure D.17: Plots of the observed vs predicted counts of dengue in districts 37 to 45

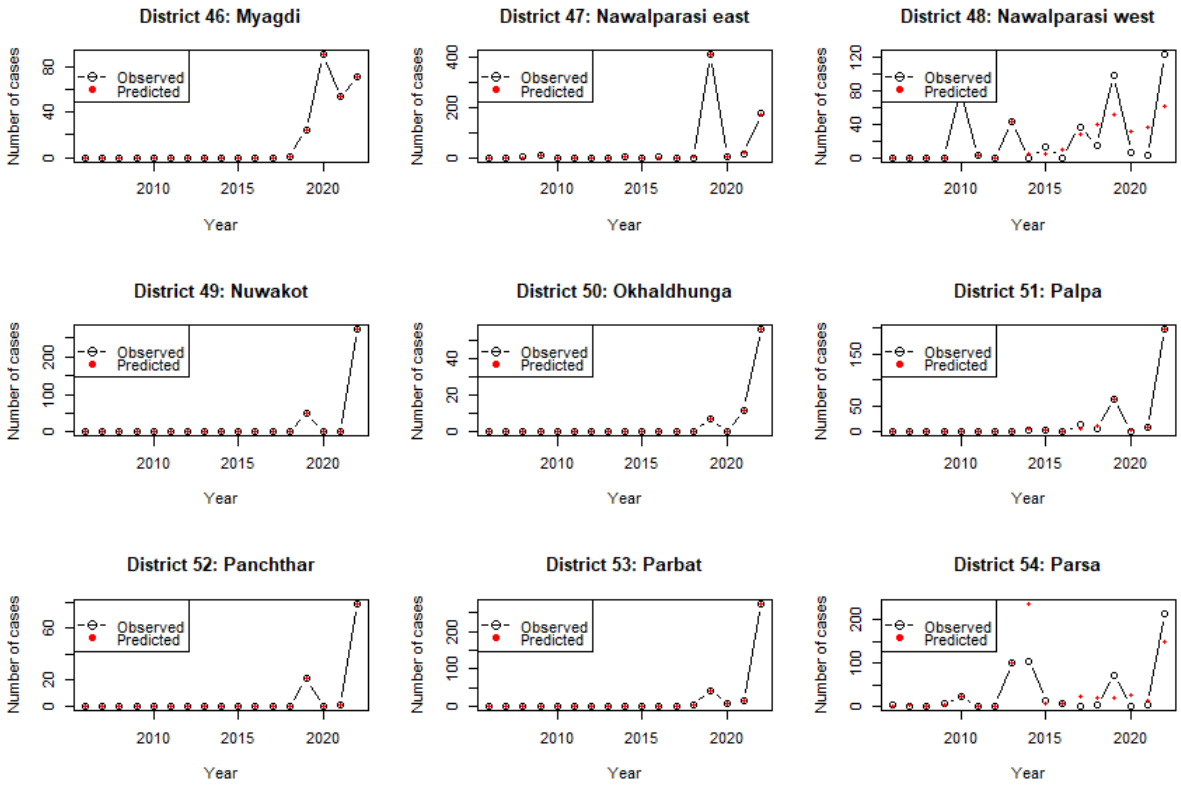

Figure D.18: Plots of the observed vs predicted counts of dengue in districts 46 to 54

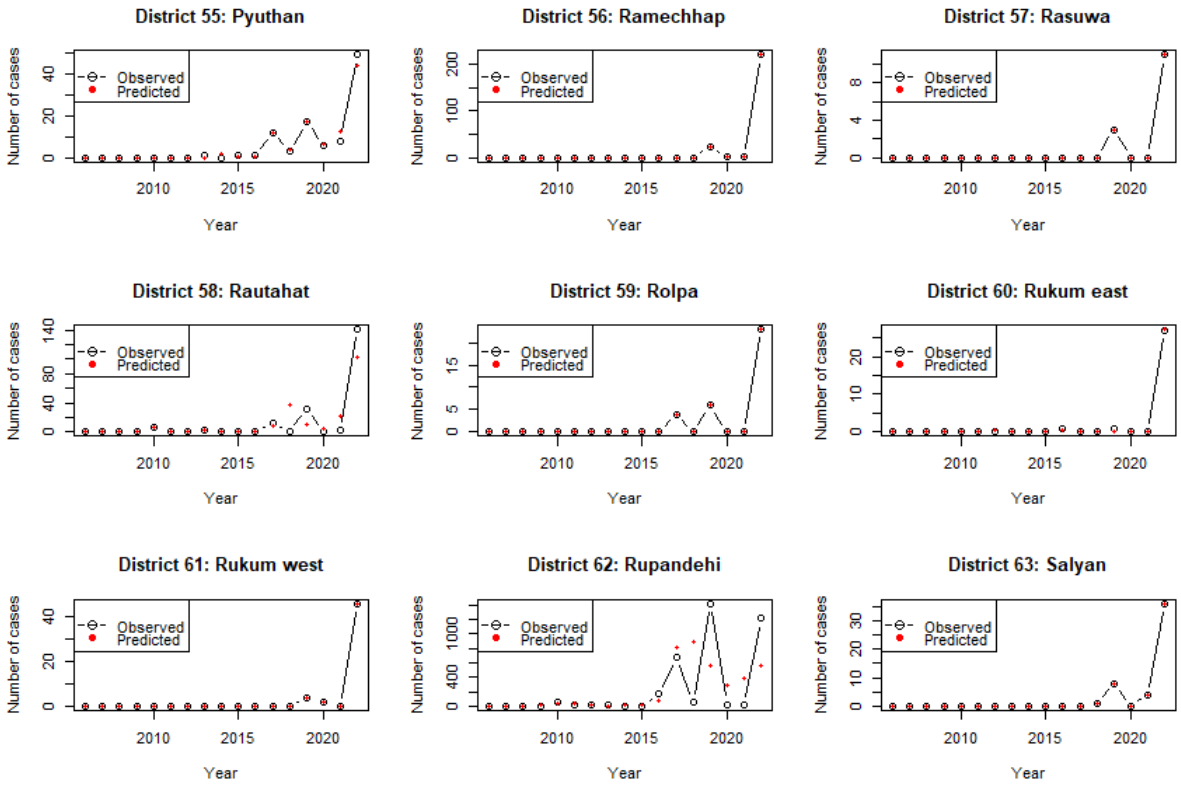

Figure D.19: Plots of the observed vs predicted counts of dengue in districts 55 to 63

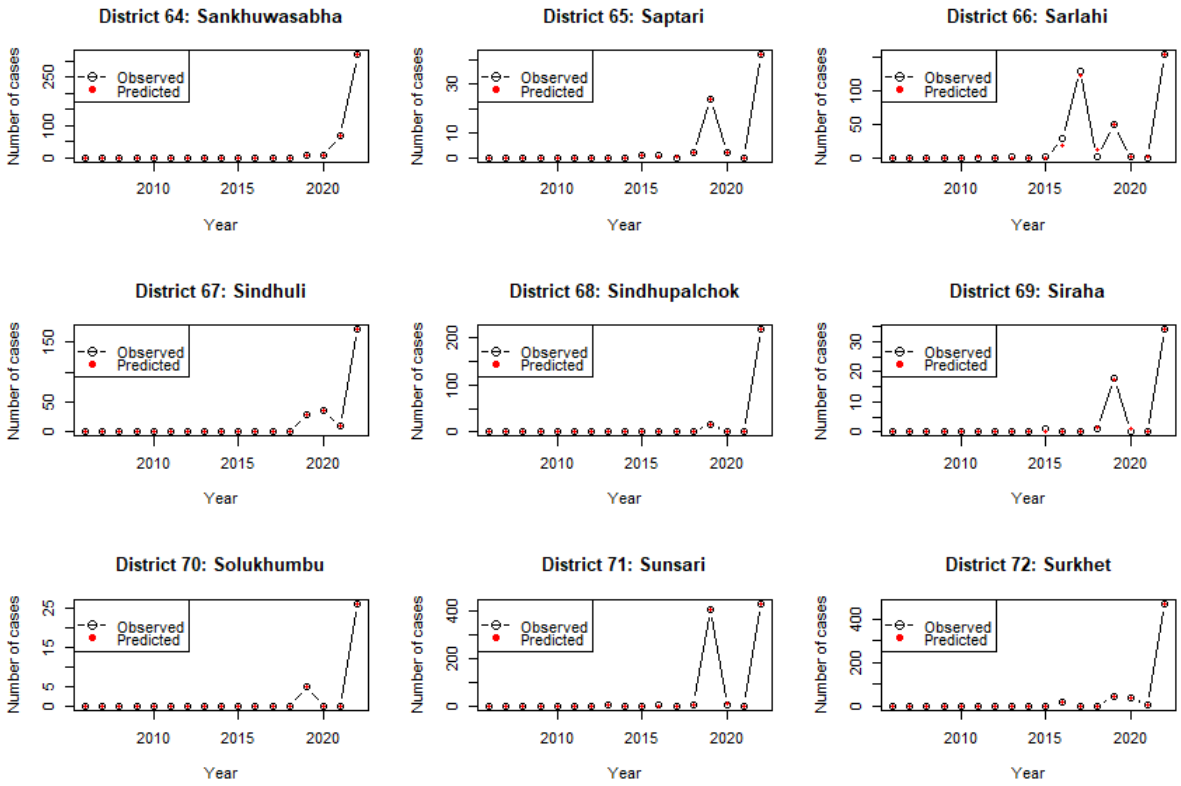

Figure D.20: Plots of the observed vs predicted counts of dengue in districts 64 to 72

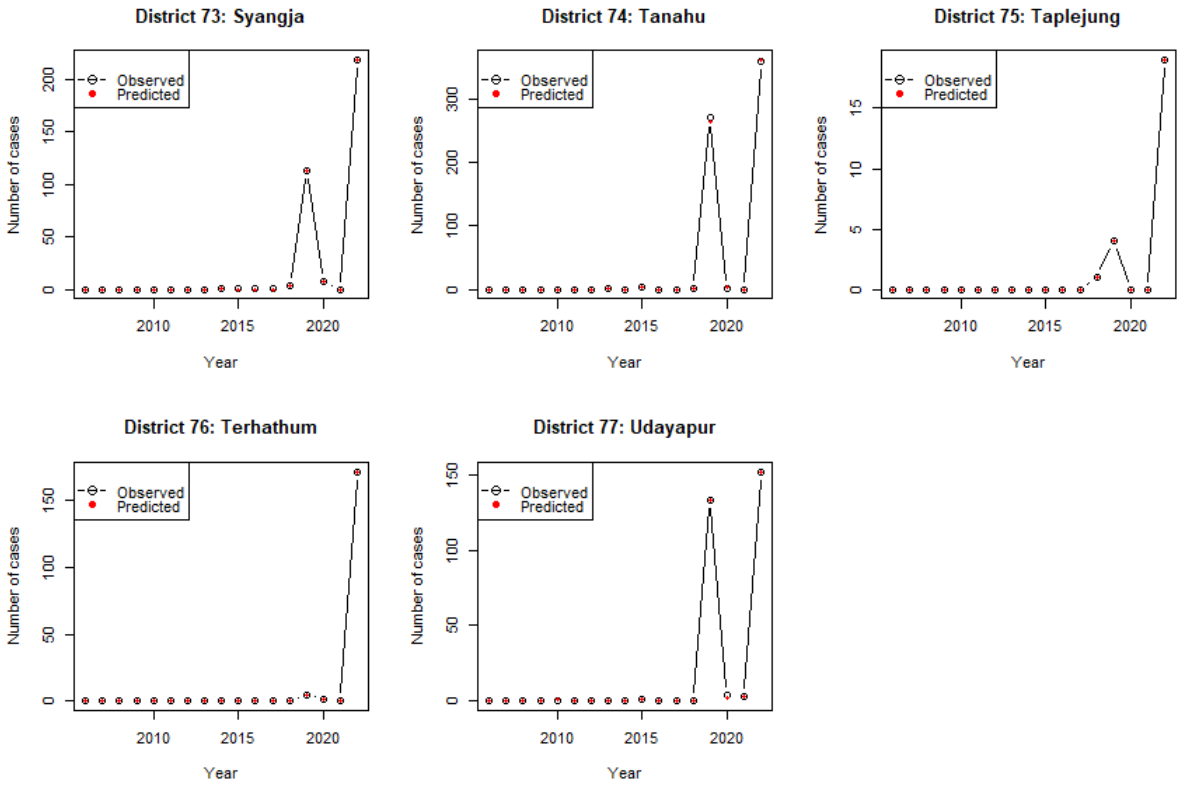

Figure D.21: Plots of the observed vs predicted counts of dengue in districts 73 to 77

The figures below show that, in districts where model validation was unsatisfactory, the differences between the observed and predicted dengue counts decreased as the number of outbreaks in the models increased.

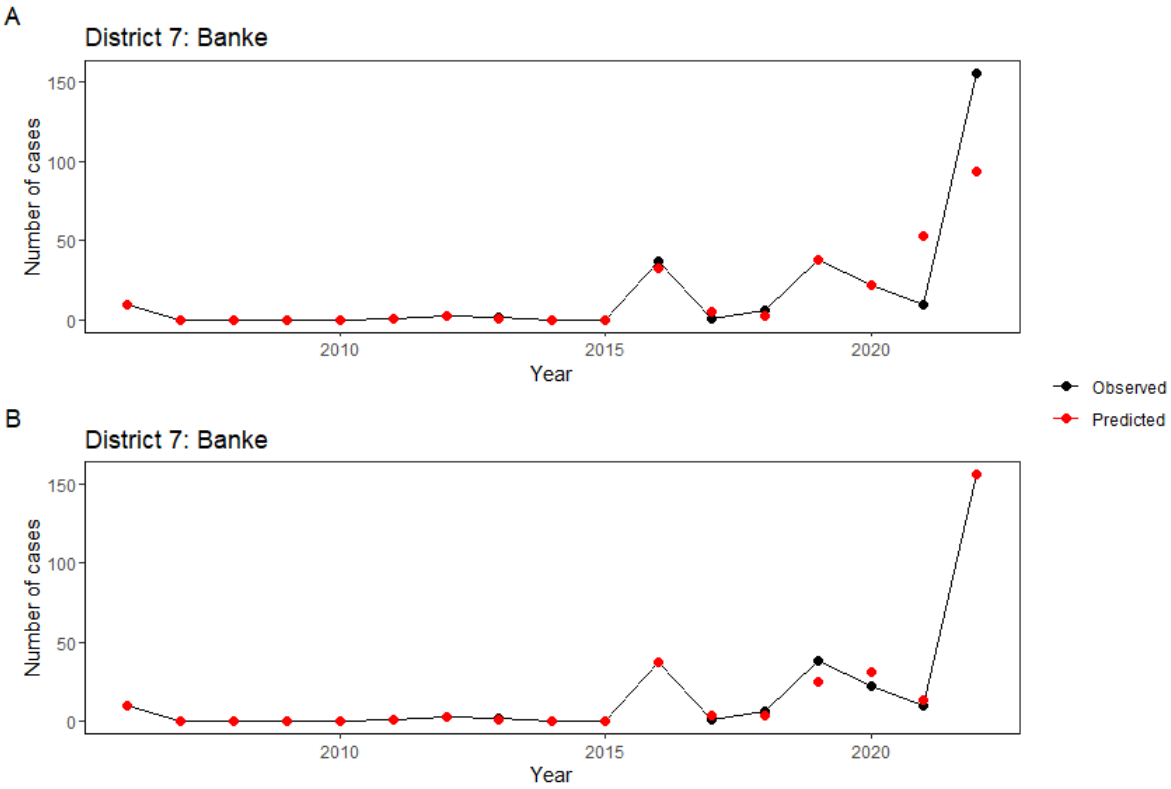

Figure D.22: Plot of observed versus predicted dengue counts for district number 7 (Banke), illustrating models with three outbreaks (Figure A) and four outbreaks (Figure B).

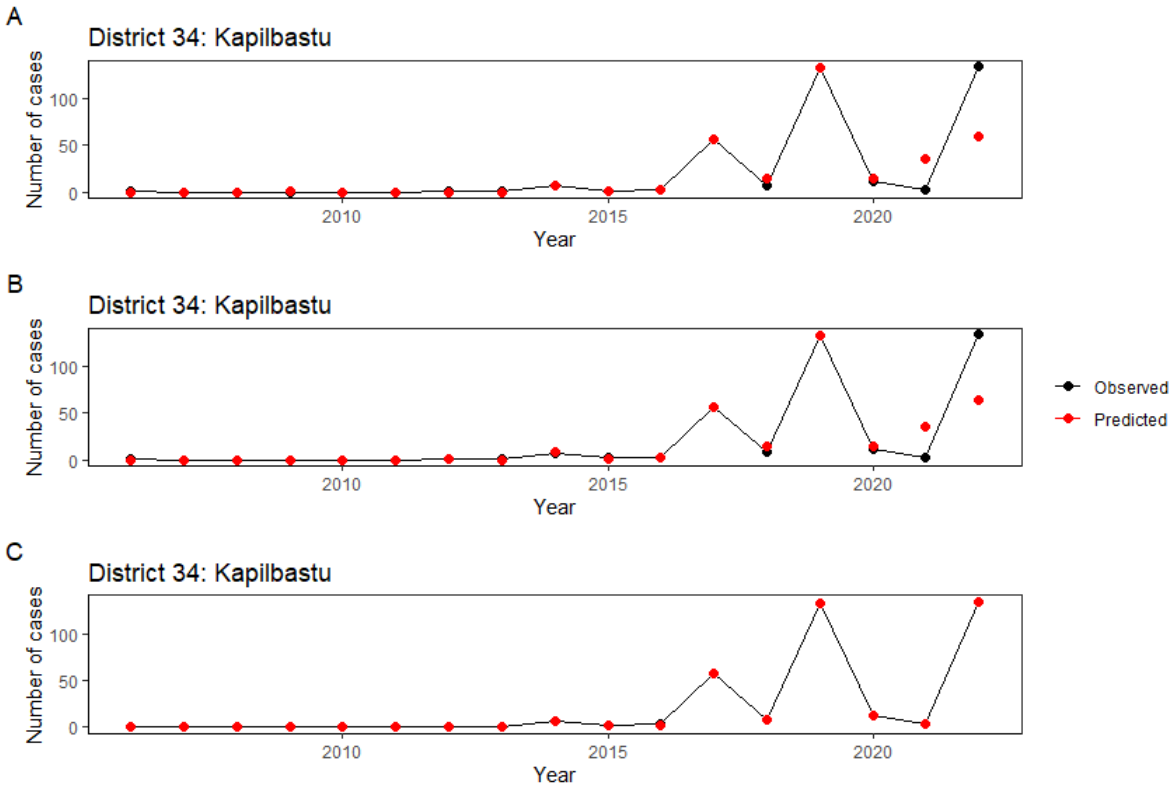

Figure D.23: Plot of observed versus predicted dengue counts for district number 34 (Kapilbatsu), illustrating models with three outbreaks (Figure A), four outbreaks (Figure B), and five outbreaks (Figure C).

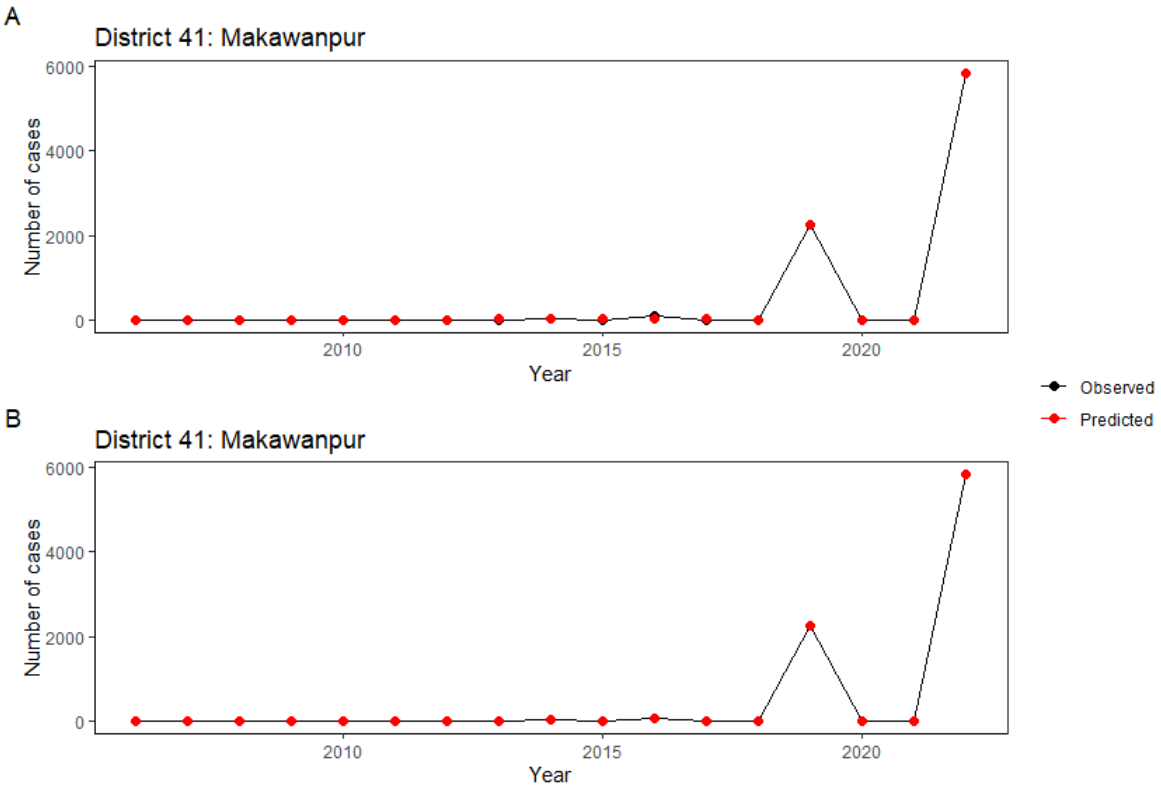

Figure D.24: Plot of observed versus predicted dengue counts for district number 41 (Makawanpur), illustrating models with three outbreaks (Figure A), and four outbreaks (Figure B).

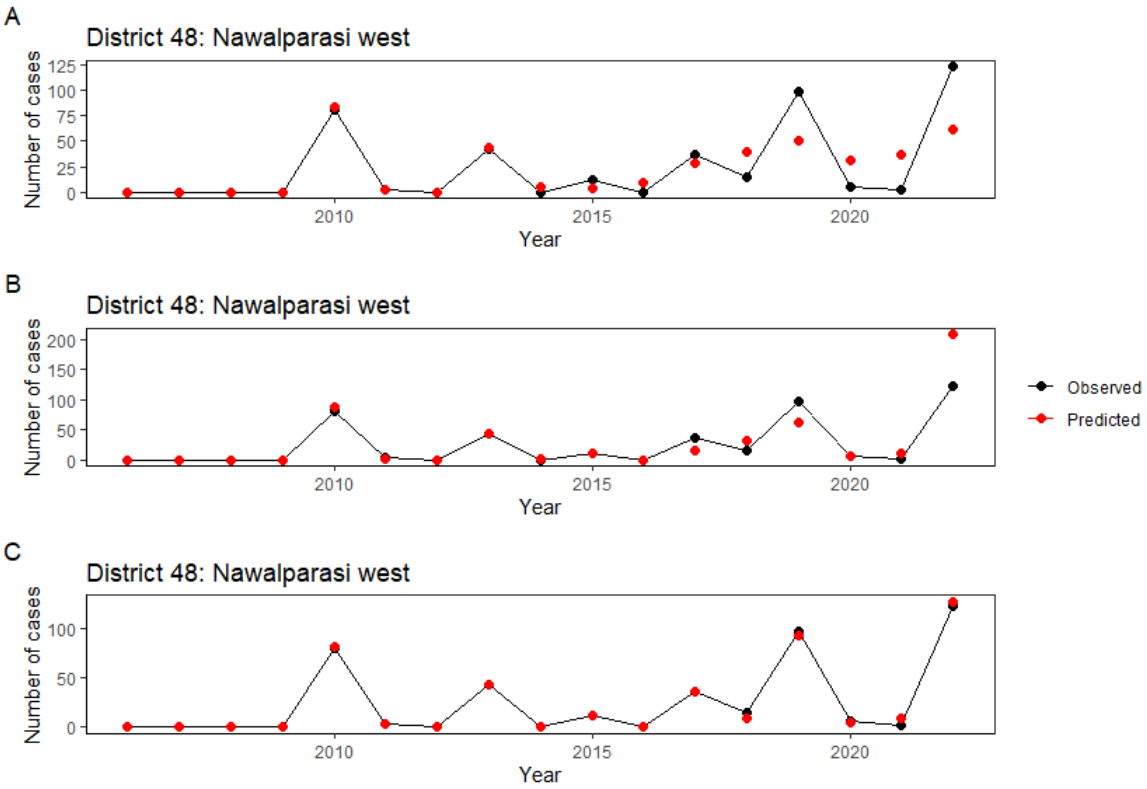

Figure D.25: Plot of observed versus predicted dengue counts for district number 48 (Nawalparasi West), illustrating models with three outbreaks (Figure A), four outbreaks (Figure B), and five outbreaks (Figure C).

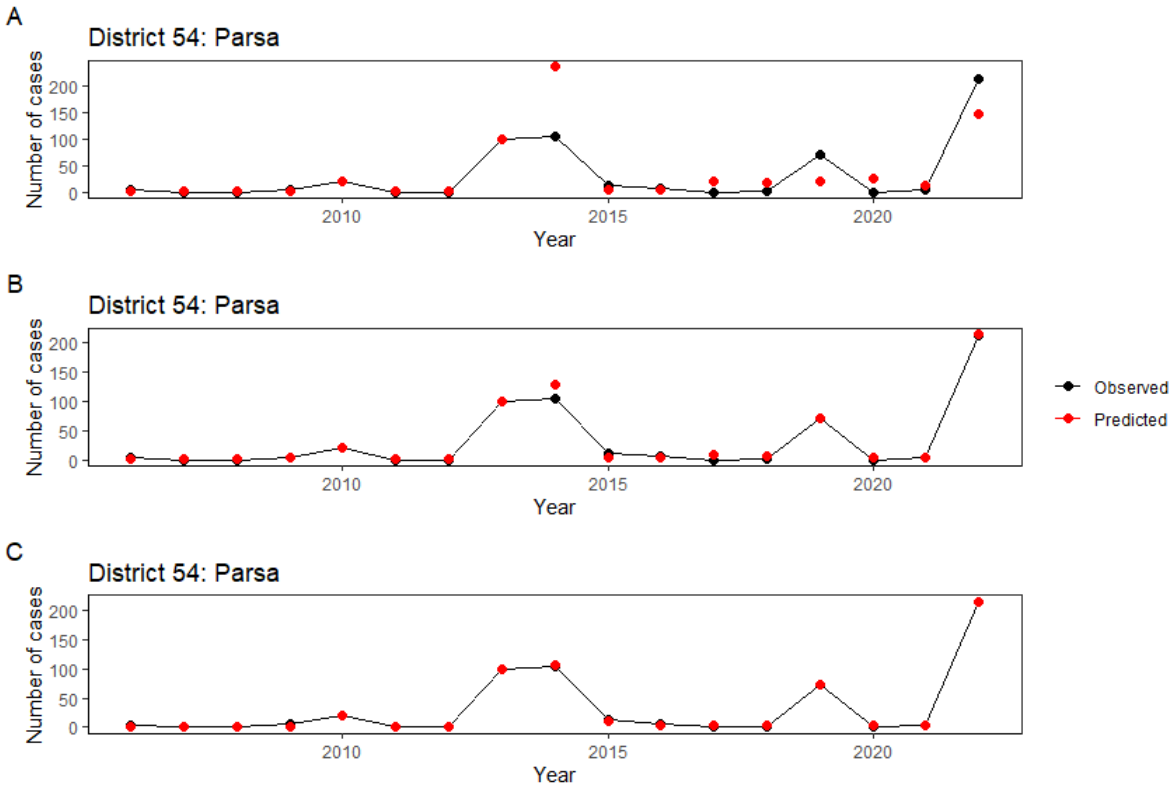

Figure D.26: Plot of observed versus predicted dengue counts for district number 54 (Parsa), illustrating models with three outbreaks (Figure A), four outbreaks (Figure B), and five outbreaks (Figure C).

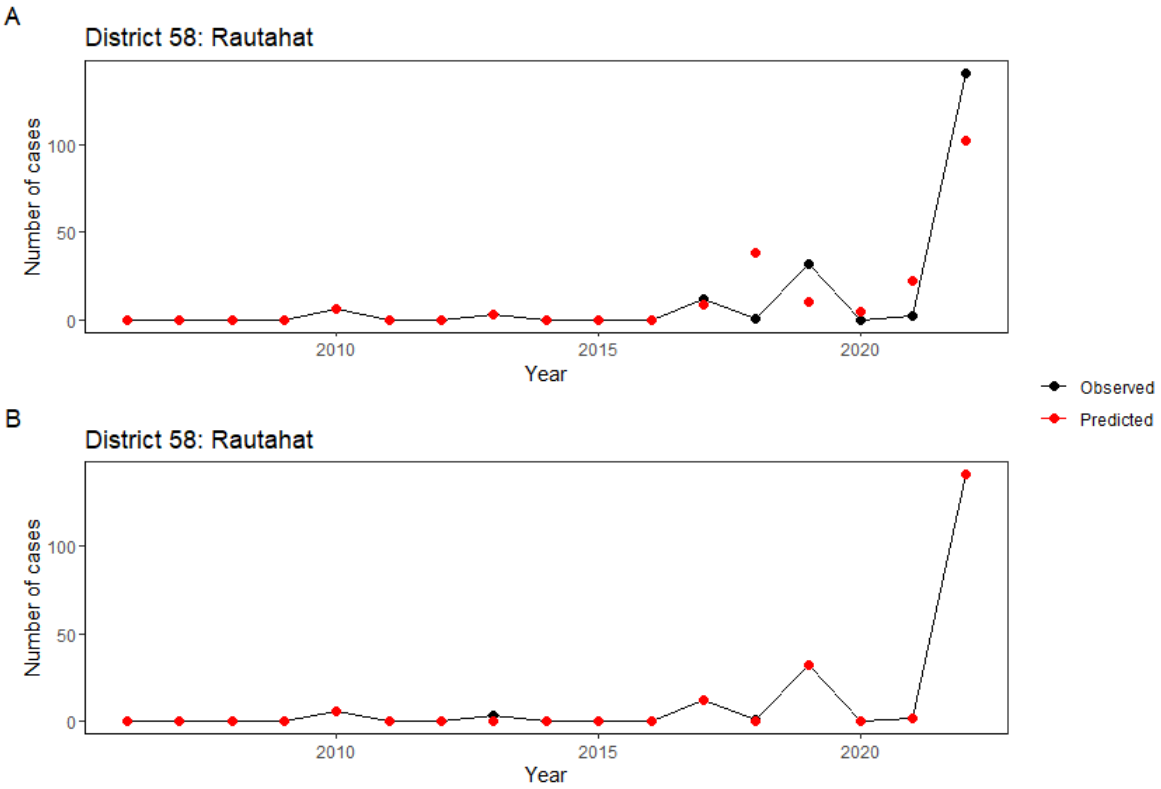

Figure D.27: Plot of observed versus predicted dengue counts for district number 58 (Rautahat), illustrating models with three outbreaks (Figure A), and four outbreaks (Figure B).

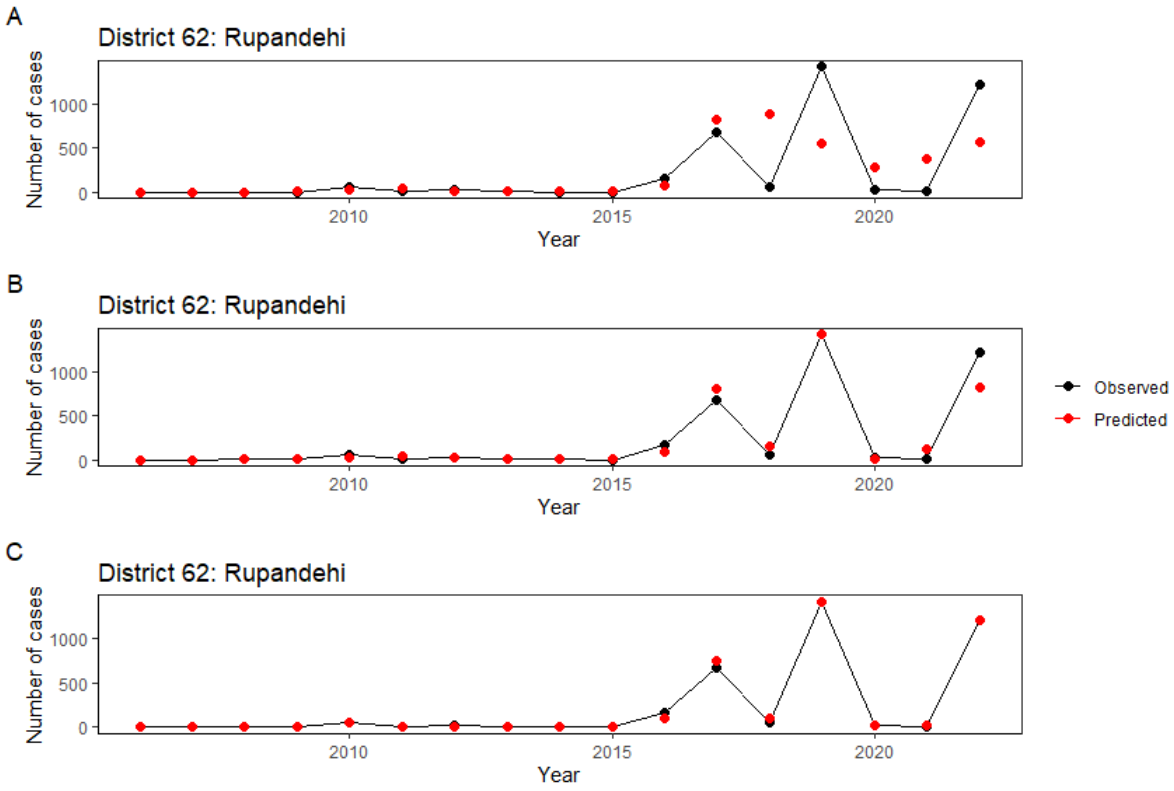

Figure D.28: Plot of observed versus predicted dengue counts for district number 62 (Rupandehi), illustrating models with three outbreaks (Figure A), four outbreaks (Figure B), and five outbreaks (Figure C).

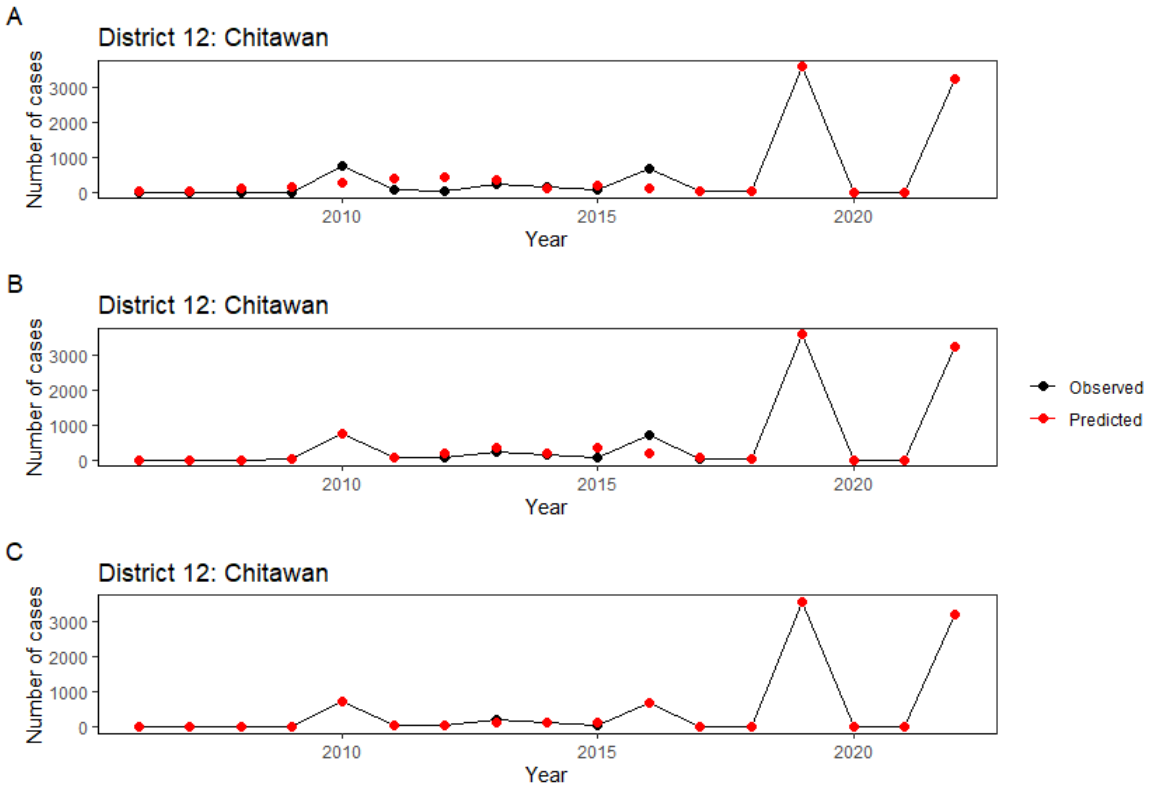

Figure D.29: Plot of observed versus predicted dengue counts for district number 12 (Chitawan), illustrating models with three outbreaks (Figure A), four outbreaks (Figure B), and five outbreaks (Figure C).

Supplementary Material References

[1] A. J. Tatem. "WorldPop, open data for spatial demography". In: *Scientific data* 4.1 (2017), pp. 1–4.

[2] L. D. Howe, B. Galobardes, A. Matijasevich, D. Gordon, et al. "Measuring socio-economic position for epidemiological studies in low-and middle-income countries: a methods of measurement in epidemiology paper". In: *International journal of epidemiology* 41.3 (2012), pp. 871–886.

[3] L. Hjelm, A. Mathiassen, D. Miller, and A. Wadhwa. "Creation of a wealth index". In: *United Nations World Food Programme* (2017).

[4] P. J. Diggle and E. Giorgi. *Model-based geostatistics for global public health: methods and applications*. Chapman and Hall/CRC, 2019.

[5] E. Giorgi and P. J. Diggle. "PrevMap: an R package for prevalence mapping". In: *Journal of Statistical Software* 78 (2017), pp. 1–29.

[6] P. J. Diggle, I. Kaimi, and R. Abellana. "Partial-likelihood analysis of spatio-temporal point-process data". In: *Biometrics* 66.2 (2010), pp. 347–354.

[7] E. Gabriel, B. S. Rowlingson, and P. J. Diggle. "stpp: an R package for plotting, simulating and analyzing Spatio-Temporal Point Patterns". In: *Journal of Statistical Software* 53 (2013), pp. 1–29.

[8] E. Gabriel and P. J. Diggle. "Second-order analysis of inhomogeneous spatio-temporal point process data". In: *Statistica Neerlandica* 63.1 (2009), pp. 43–51.

[9] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". In: *Nature Scientific Data* 5.1 (2018), pp. 1–12.

[10] ICF International. *Spatial Data Repository, The Demographic and Health Surveys Program. Modeled Surfaces*. 2015. URL: https://spatialdata.dhsprogram.com/modeled-surfaces/#survey=MW%7C2015%7CDHS.

[11] ICF International. *Spatial Data Repository, The Demographic and Health Surveys Program. Modeled Surfaces*. Funded by the United States Agency for International Development (USAID). 2015. URL: https://spatialdata.dhsprogram.com/modeled-surfaces/#survey=MW%7C2015%7CDHS.

[12] ICF International. *Spatial Data Repository, The Demographic and Health Surveys Program. Modeled Surfaces.* 2015. URL: https://spatialdata.dhsprogram.com/modeled-surfaces/#survey=MW%7C2015%7CDHS.

[13] N. Sommet and D. Morselli. "Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS." In: *International Review of Social Psychology* 30 (2017), pp. 203–218.

[14] G. Leckie. *Three-level multilevel models–concepts. LEMMA VLE Module 11, 1–47.* 2013.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker. "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823* (2014).

[16] K. Deribe, A. Mbituyumuremyi, J. Cano, M. J. Bosco, et al. "Geographical distribution and prevalence of podoconiosis in Rwanda: a cross-sectional country-wide survey". In: *The Lancet Global Health* 7.5 (2019), e671–e680.

[17] E. Giorgi, C. Fronterrè, P. M. Macharia, V. A. Alegana, et al. "Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict". In: *Journal of the Royal Society Interface* 18.179 (2021), p. 20210104.

[18] P. J. Diggle, J. A. Tawn, and R. A. Moyeed. "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3 (1998), pp. 299–350.

[19] N. de Silva and A. Hall. "Using the prevalence of individual species of intestinal nematode worms to estimate the combined prevalence of any species". In: *PLoS Negl Trop Dis* 4.4 (2010), e655.

[20] C. Czado, T. Gneiting, and L. Held. "Predictive model assessment for count data". In: *Biometrics* 65.4 (2009), pp. 1254–1261.