Spatial Modelling of Emerging Infectious Diseases: Quantifying the Role of Climate, Cities and Connectivity on Dengue Expansion in Brazil

Sophie Alice Lee

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London

March 2023

Department of Infectious Disease Epidemiology

Faculty of Epidemiology and Population Health

London School of Hygiene & Tropical Medicine

Funded by The Royal Society

Declaration

I Sophie Alice Lee, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date: 30/09/22

Abstract

Over the past 50 years, dengue has been expanding globally into previously unaffected areas. This has been attributed to climate change, urbanisation, and increased connectivity driven by human movement. In Brazil, the rapid expansion of dengue has led to outbreaks occurring in previously unaffected areas, including the temperate South region and remote areas of the Amazon rainforest. In this thesis, I use spatial models to explore the drivers of dengue re-emergence and expansion in 21st century Brazil. Spatial modelling techniques are used to disentangle the effect of increasing temperatures in South Brazil and the contribution of human movement around the Brazilian urban network to the expansion of the dengue transmission zone.

First, using Bayesian spatiotemporal models, I found an increased odds of dengue outbreaks in highly urbanised, well-connected cities which had already experienced an outbreak and had year-round temperatures suitable for dengue transmission (Chapter 2). Although these models were able to capture the significant impact of temperature on the expansion of dengue into South Brazil, they were unable to quantify the role of human movement in dengue expansion. I conducted a systematic review to identify how spatial connectivity had been accounted for in models of mosquito-borne disease transmission and the assumptions made about how spatial connectivity arises (e.g., human movement between regions) (Chapter 3). Although the number of spatial modelling papers had increased rapidly over the past 5 years, very few statistical models considered connectivity arising due to human movement and there were no models identified capable of accounting for multiple sources of spatial connectivity. Expanding current state-of-the-art statistical frameworks using ideas from network-based mechanistic models identified in this systematic review became the focus of the remainder of the thesis.

I developed a novel statistical modelling approach which can include multiple sources of spatial connectivity, such as similarities between close areas and human movement, and quantify the relative contribution of each source to the overall spatial structure of the model outcome (Chapter 4). This framework was applied to dengue outbreak data for the whole of Brazil between 2001 – 2020 (Chapter 5). Model results showed that human movement based on commuting for work or education contributed very little to the overall spatial structure of the number of outbreaks per municipality in Brazil, but this contribution was significantly higher in North and Northeast Brazil compared to South Brazil.

In this thesis, I have explored the complex, interacting drivers of dengue expansion in Brazil since 2001, including increasing temperatures in South Brazil and connections between cities arising from human movement around the Brazilian urban network. Although this thesis focuses on dengue expansion in Brazil, the methods presented here are flexible enough to be applied in any Bayesian hierarchical model where spatial connectivity exists within the data. Given the increasing risk of future pandemic pathogens due to increasing climate and globalisation, robust modelling tools are essential to gain better understanding of infectious disease emergence and identify areas at future risk of expansion.

Acknowledgements

Firstly, I would like to thank my supervisors Rachel, Theo and John for their invaluable support and feedback throughout my PhD. To Rachel for being so generous with your time and guidance – your enthusiasm and moral support dragged me out of many lockdown lows and made me feel like completing this PhD was possible. To Theo for reigniting my love of statistics (and mezze) during my visit to Cyprus. I would also like to thank the Royal Society for funding this project and making it possible.

Thank you to all my collaborators and co-authors for their input and suggestions. Muito obrigada to Brazilian colleagues, especially Rafa, Claudia and Christovam for helping me better understand dengue in Brazil, even if March 2020 was not the ideal time for a visit! Thank you to Bruna for making me feel so welcome in Rio and introducing me to samba. Thank you to Raquel and Chloe for making me feel so welcome in Barcelona and always being up for tapas and vermut!

I was very lucky to be part of a wonderful community of PhD students at LSHTM. Thank you to everyone in LG22 for making it such a fun place to work and for all the much needed coffee and Pumphandle breaks. In particular, thank you Ellie for our virtual tea breaks and chats, you made me feel less alone and I don't think I would have finished this thesis without them.

This thesis would not have been possible without the endless support and patience of friends and family who have had to listen to me ramble for 4 very long years. I promise to never make any of you read this. Most importantly, thank you to my Mum and Dad for always believing in me, giving me every opportunity possible and for being behind me whatever I choose to do.

Table of contents

List of figures

Acronyms

ACP - Population concentration areas [“*Áreas de Concentração de População*”]

aOR - Adjusted odds ratio

CAR - Conditional autoregressive

CI - Credible interval

DALYs - disability-adjusted life-years

DENV - dengue virus

DHF - dengue haemorrhagic fever

GAM - Generalised additive model

GLM - Generalised linear model

GMRF - Gaussian Markov random field

GWR - Geographically weighted regression

IBGE - Brazilian Institute of Geography and Statistics [“*Instituto Brasileiro de Geografia e Estatística*”]

INLA - Integrated nested Laplace approximation

LGM - Latent Gaussian model

MCMC - Markov chain Monte Carlo

REGIC - Regions of Influence of Cities [“*Regiões de Influência das Cidades*”]

REML - Restricted maximum likelihood

ROC - Receiver operating characteristic

scPDSI - Self-calibrating Palmer drought severity index

SINAN - Notifiable Diseases Information System [“*Sistema de Informação de Agravos de Notificação*”]

STAR - Structured additive regression

WHO - World Health Organisation

1. Introduction

1.1 Motivation

Recent public health emergencies such as the Zika epidemic in the Americas [1,2], the western Africa Ebola outbreak [3,4] and the ongoing COVID-19 pandemic [5] have brought emerging infectious diseases to the forefront of global health research. Changes in climate and land use have led to increased cross-species contact and therefore increased the risk of spillover from infectious zoonotic pathogens [6]. This, coupled with an increasingly connected world means emerging diseases are spreading faster and further than ever [7]. Mathematical models are important tools in understanding emerging infectious disease dynamics and predicting the spread and persistence of these viruses [8]. Where expansion of a disease has occurred, spatial models can be employed to explore the potential drivers of this emergence and identify areas at risk of future expansion whilst accounting for spatial autocorrelation or connectivity across a geographical area.

In this thesis, a spatial modelling framework is developed to quantify the role of climate change, socioeconomic factors, and human movement on the expansion of emerging infectious diseases. This framework is developed in the context of dengue re-emergence and expansion in Brazil. In this chapter I introduce dengue and discuss its epidemiology and global burden. Following this, I discuss the re-emergence of dengue in Brazil and introduce the hypothesised drivers for the expansion of the dengue transmission zone. Finally, I introduce statistical modelling methods that can be used to quantify the impact of potential drivers on emerging infectious diseases. This chapter is concluded with the thesis aims, objectives, and the structure of the thesis.

1.2 Dengue

Dengue is one of the world's fastest growing communicable diseases, with the number of cases doubling every decade over the past 30 years [9,10]. Dengue is transmitted to humans by the bite of an infected female mosquito of the *Aedes* genus, primarily *Aedes aegypti* although *Aedes albopictus* can transmit the virus less efficiently [9]. *Aedes aegypti* are now ubiquitous in tropical and sub-tropical regions worldwide, while *Aedes albopictus* continue to expand into

temperate regions [11]. The rapid expansion of the global distribution of *Aedes* mosquitoes and, subsequently, the arboviruses transmitted by them, has been attributed to rising temperatures, increasing urbanisation, and increased global connectivity arising from human movement [9,12–17]. Around half the world's population are believed to live in areas at risk of dengue infection [18].

Dengue is caused by a virus of the Flaviviridae family of which there are four genetically distinct dengue serotypes (DENV-1 – 4). Infection from one serotype provides long-term immunity against that particular serotype and short-term protection against others. However, the risk of severe dengue increases upon secondary infection from another serotype [9]. Most people with dengue will remain asymptomatic or develop mild symptoms, only around 25% of cases are clinically apparent [18]. Dengue symptoms include fever, nausea, rashes, aches, and lethargy. A small proportion of clinically apparent cases will progress to severe dengue or dengue haemorrhagic fever (DHF), a potentially lethal form of the disease, characterised by severe bleeding, plasma leakage and organ involvement [19]. Although there is no specific treatment for dengue, supportive care can reduce mortality rates of severe dengue to below 1% [20]. Despite low mortality rates compared to other mosquito-borne diseases like malaria, dengue still has a huge burden, contributing to an estimated 2.9 million disability-adjusted life-years (DALYs) across the globe in 2017 [21].

Despite the huge global dengue burden, there is still no consensus on the optimal disease control strategy. The World Health Organisation's (WHO) current recommendations focus on vector control, promoting an integrated vector management programme that combines targeting mosquitoes directly (i.e., kill mosquitoes or their larvae using insecticides or biological control agents), and indirectly (e.g., improvements in sanitation or environmental modifications that aim to reduce potential mosquito habitat) [19]. However, there is a lack of evidence that these measures are effective at reducing disease incidence or would be sustainable in a wide-scale application [22,23]. Although several dengue vaccine candidates are in development, only one has completed phase 3 trials, DENGVAXIA® [24,25]. The vaccine has been licensed for use in 20 dengue-endemic countries but has been found to significantly increase the risk of serious dengue and hospitalisations in patients that were seronegative at administration [25,26]. The WHO currently recommends pre-vaccination screening of serostatus before administering DENVAXIA® [27]. However, no rapid diagnostic test validated for this purpose currently exists, and even if there were, this would present major

practical issues in a large-scale vaccination programme. A promising innovation in vector control is releasing mosquitoes infected with Wolbachia, a bacteria that has been shown to inhibit arbovirus infection in *Aedes* mosquitoes [28,29]. Modelling studies predict that releasing Wolbachia-infected mosquitoes into dengue-endemic settings would significantly reduce dengue infections, and in certain settings potentially lead to elimination [30,31]. These were supported by experimental and quasi-experimental studies showing significant reductions in dengue infections and hospitalisations in areas following the release of Wolbachia-infected mosquitoes [32–36]. For these interventions to be as successful as possible, it is important to target the most affected locations.

1.3 Dengue in Brazil

Dengue is hyper-endemic in Brazil, which means it experiences a persistent high level of transmission with all four genetically distinct serotypes co-circulating [37]. Dengue was re-introduced to Brazil in the 1980s, following the end of a successful mosquito eradication campaign which led to the reinfestation of Brazil with *Aedes aegypti* [38]. Although the first outbreak occurred in Boa Vista in 1981, dengue only became a public health problem 5 years later when an outbreak in Rio de Janeiro spread rapidly along the coast to different states due to the large number of people moving to and from the city [37–39]. The frequency and intensity of dengue epidemics in Brazil have increased over the last 20 years [40]. The number of areas experiencing outbreaks has also increased over this period [41]. Previous studies had identified geographical barriers to dengue transmission, beyond which areas were relatively protected from outbreaks [42]. A barrier to transmission was identified in South Brazil, a temperate region which experiences cold winters, too cold for *Aedes* mosquitoes to effectively transmit dengue virus [43,44]. However, temperatures in this region are increasing [45], and recent data shows that outbreaks are now occurring as far south as Rio Grande do Sul [46]. Another barrier was identified in the western Amazon, a relatively remote region which is mostly disconnected from the Brazilian urban network [47], providing protection from the introduction of infectious hosts and vectors. However, recent rapid and unplanned urbanisation without adequate infrastructure, as well as improvements in transportation networks in the region have led to explosive dengue outbreaks [48,49].

Dengue is a notifiable disease in Brazil and all suspected cases must be recorded in the Sistema de Informação de Agravos de Notificação (SINAN) portal within 60 days [50]. During non-epidemic periods, it is recommended all suspected cases are laboratory confirmed. However, during epidemic periods cases are more often notified based on the Ministry of Health's syndromic definition. That is, someone that resides in an area or has travelled in the past 14 days to an area with occurrence of *Aedes aegypti* mosquitoes, experiences fever for between 2 and 7 days, and has at least 2 of the following symptoms: nausea/vomiting, rash, myalgia, headache, petechiae and leukopenia [50]. Although the system is unified across the country and covers all 5,570 municipalities, reporting rates can differ drastically across the country and between epidemic and non-epidemic periods [51,52]. Surveillance is passive in Brazil which means cases are likely underreported as many cases are mild or asymptomatic [18] and patients may not seek medical assistance due to perceived low quality of healthcare or long waits during epidemic periods [53]. One study estimated that there were 12 actual cases for every notified case in Brazil, rising to over 17 in periods of low transmission [52]. Despite these issues, the surveillance system is considered sensitive and robust [53], and although the accuracy of case counts may differ, the Brazilian Observatory of Climate and Health noted that "there is no way to hide an epidemic" [42].

1.3.1 Dengue and climate in Brazil

Due to its size, Brazil experiences a wide range of climatic, socioeconomic, and geographical settings that may contribute to an areas' suitability for dengue transmission. Brazil consists of 6 distinct biomes, defined by their vegetation and climate system, according to the Brazilian Institute of Geography and Statistics (IBGE) [54]. These are Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa and Pantanal (Figure 1a). The largest biome is the Amazon which covers 49% of Brazil and mostly consists of tropical rainforest. The Amazon biome experiences a humid, equatorial climate which means year-round high temperatures and high levels of precipitation. This contrasts greatly with the Caatinga biome which covers around 10% of Brazil and is characterised by a semi-arid climate with very little rainfall, and the Pampa biome which consists temperate grasslands with cold winter temperatures. An alternative definition of Brazil's climate system is the Köppen climate classification, which uses temperature and precipitation to categorise climates [55]. Brazil consists of 3 zones: A (tropical climate), B (dry climate), and C (temperate climate), and 9 distinct climate types (Figure 1b) [56]. However,

there is evidence to suggest that these zones are no longer stable, and that climate change is leading to a reduction in the proportion of wet tropical and temperate climate zones [45].

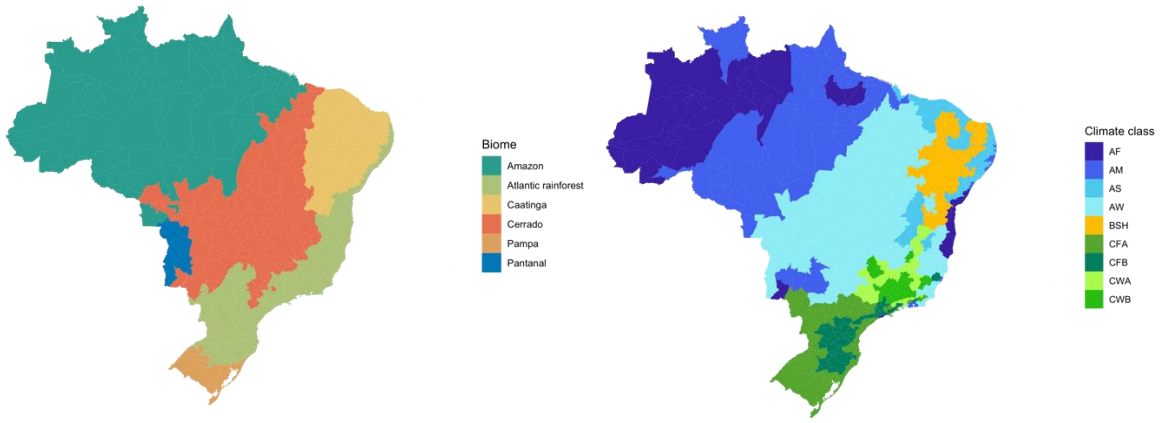


Figure 1: Climate zones of Brazil described using a) biomes and b) Köppen climate classification. Köppen classification groups: Af: Tropical rainforest climate, Am: Tropical monsoon climate, As: Tropical dry savanna climate, Aw: Tropical wet savanna, BSh: Hot semi-arid climate, Cfa: Humid sub-tropical climate, Cfb: temperate oceanic climate, Cwa: Monsoon subtropical climate, Cwb: Subtropical highland climate.

Ambient temperature is an important consideration in mosquito-borne disease dynamics as it affects mosquito biology. The ability of mosquitoes to survive, reproduce and transmit viruses are altered by temperature [17,43,44]. Temperatures can also modify human behaviour, which impacts interaction between hosts and vectors. In Brazil, dengue is transmitted by *Aedes aegypti* and, to a lesser extent, *Aedes albopictus* mosquitoes. A modelling study showed that *Aedes* mosquitoes were only able to transmit dengue virus between 16.2°C and 34.5°C [44]. Seasonal temperature variation has also been shown to impact the size and duration of epidemics [57]. Most of Brazil experiences year-round temperature suitability for dengue transmission, except South Brazil that experiences cold winters.

Hydrometeorological factors, such as precipitation and drought, are also important drivers of dengue transmission, although their impacts are complex and often delayed. As *Aedes* mosquitoes breed in pools of standing water, increased precipitation can lead to additional habitat and therefore increased cases. However, too much precipitation can flush the larvae out and reduce the number of mosquitoes in an area [58,59]. Alternatively, periods of extreme

drought, particularly in areas with inadequate access to piped water, can change water storage behaviours and create additional breeding sites [60,61]. One study found that the risk of dengue in Brazil increased shortly after (0 – 3 months) extremely wet conditions but at a longer lag (3 – 5 months) following drought conditions, particularly in areas that experienced a higher frequency of water supply shortages [62].

1.2.2 Dengue, cities and connectivity in Brazil

Brazil comprises 5 geo-political regions (North, Northeast, South, Southeast and Centre-West, Figure 2a) which consist of 26 states and one federal district containing the capital, Brasilia (Figure 2b). These regions vary greatly in terms of wealth, level of urbanisation and access to basic services. Dengue is typically an urban disease due to the evolution of *Aedes* mosquitoes to thrive alongside humans, breeding in manmade water storage containers or pools of water created by refuse [14]. Both *Aedes aegypti* and *Aedes albopictus* are widespread across Brazil [11], with *Aedes aegypti* predominantly found in urban settings and *Aedes albopictus* typically found in peri-urban and rural areas [12]. However, there is evidence that the primary vector, *Aedes aegypti*, is expanding into peri-urban and rural areas across Latin America [63,64].

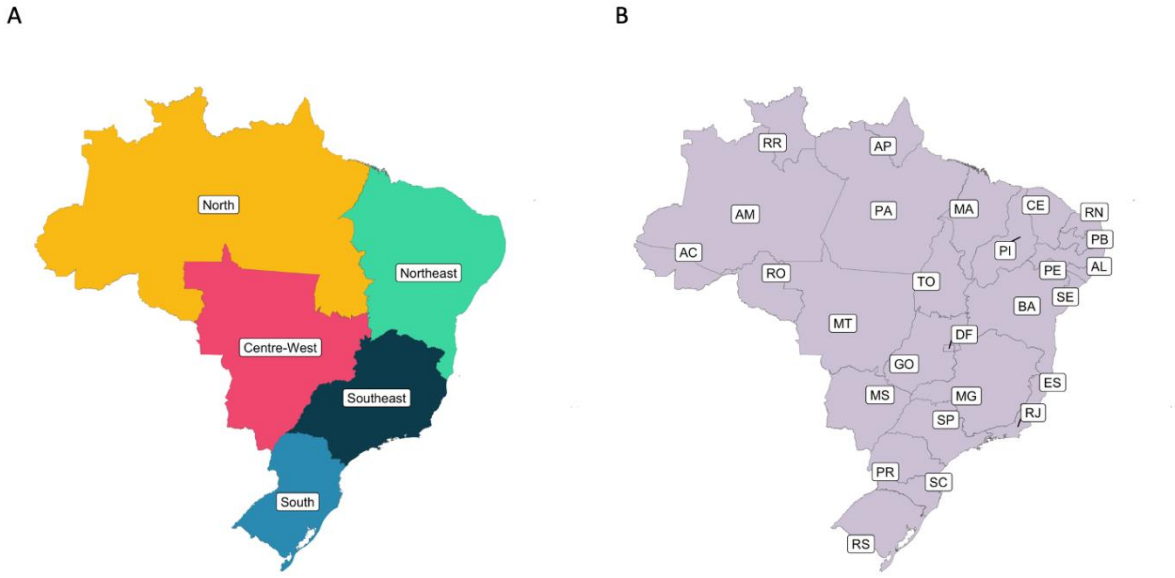


Figure 2: The organisation of Brazil into a) 5 geo-political regions, and b) 27 federal units. Abbreviations: AC = Acre, AL = Alagoas, AP = Amapá, AM = Amazonas, BA = Bahia, CE = Ceará, DF = Distrito Federal, ES = Espírito Santo, GO = Goiás, MA = Maranhão, MT = Mato Grosso, MS = Mato Grosso do Sul, MG = Minas Gerais, PA = Pará, PB = Paraíba, PR = Paraná, PR = Pernambuco, PI = Piauí, RJ = Rio de Janeiro, RN = Rio Grande do Norte, RS = Rio

Grande do Sul, RO = Rondônia, RR = Roraima, SC = Santa Catarina, SP = São Paulo, SE = Sergipe, TO = Tocantins.

According to the 2010 census, the percentage of households with access to piped water, the sewage network, and refuse collection ranged from 0 to 100% [65]. Access to basic services is highly correlated to the level of urbanisation in Brazil (Appendix B.1, Figure D). In general, rural communities in the historically poorer North region had the lowest levels of access, compared to the traditionally wealthier, highly urbanised South and Southeast regions (Figure 3a). However, levels of access to the piped water network does not guarantee this is reliable, as is the case in São Paulo which experiences intermittent water supply due to increasingly severe droughts despite being one of the wealthiest cities in Brazil with the highest access to basic services [66]. Residents without access to reliable water networks must rely on alternative approaches, such as rainwater collection, which may create *Aedes* breeding habitats if not properly maintained [67].

A B

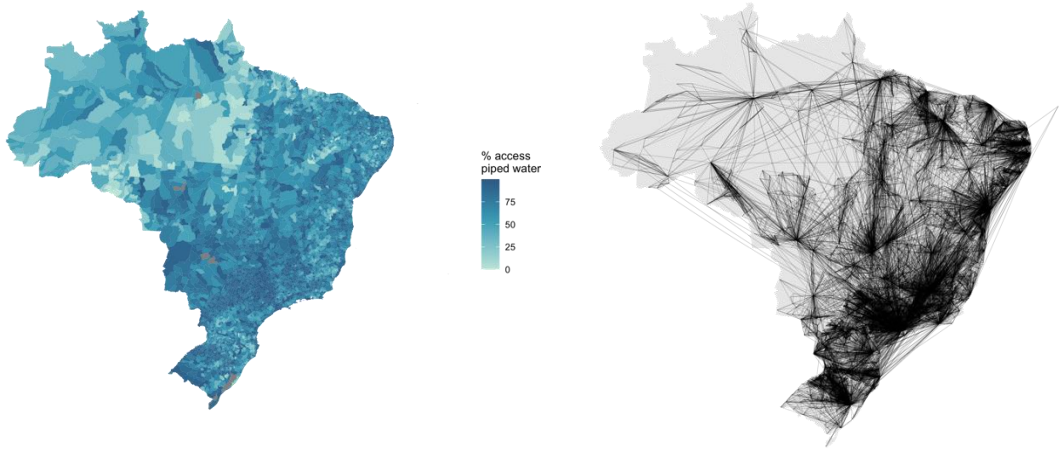


Figure 3: a) Percentage of residents with access to the piped water network in 2010 (source: IBGE demographic census [65]) and b) links between cities arising from the movement of people and goods around the Brazilian urban network (source: REGIC 2018 [47]).

In addition to socioeconomic differences between regions, patterns of human movement also differ greatly across Brazil. Approximately every 10 years, IBGE publishes the Regions of Influence of Cities ("Regiões de Influência das Cidades", REGIC) study which aims to recreate

the Brazilian urban network and describe the hierarchies and links between cities across Brazil. The latest REGIC study uses data from 2018 [47]. The urban network aims to explain the flow of people, goods, and services between cities within Brazil. This is based on a survey carried out in every municipality in Brazil which collects information including public transport connections, air travel connections, and residents' travel habits: where residents travel for work, education, healthcare, shopping, and other recreational activities. Cities are categorised into five levels of influence based on their 'attractiveness' within the urban network. These are:

1. Metropolis: the largest and most connected cities in Brazil, characterised by a strong connection across the entire country and internationally. There were 15 metropoles in the 2018 REGIC network (increased from 12 in the 2007 study [68]), including the capital Brasilia, other large state capitals such as São Paulo and Rio de Janeiro, and Campinas in São Paulo state, the only metropolis which is not a state capital. All connections within the urban network converge to one or more metropolis.
2. Regional capital: large cities that are characterised by strong attraction within the region they are located and to metropoles. There were 97 municipalities classified as regional capitals in the 2018 study, including state capitals not classified as metropoles such as Rio Branco, Campo Grande and Porto Velho, and other large cities.
3. Sub-regional capital: cities with a lower level of connectivity than regional capitals but still have visitors from smaller cities for 'less complex' activities (for example non-specialist healthcare). These are mostly connected locally and to the three largest metropoles.
4. Zone centre: smaller cities, mostly visited by residents from neighbouring cities for commerce.
5. Local centre: smallest cities in the network that are generally used by residents for daily activities such as shopping. Local centres do not typically have any influence outside their limits and are not the main destination for any other city.

The REGIC study found that patterns of movement and connectivity differed vastly between regions of Brazil. For example, the Southeast has the most 'high-level' centres (e.g., metropoles and regional capitals) as it is home to some of the largest cities in Brazil, contributing a large proportion of the national income. In contrast, the Northeast comprises mostly 'lower-level' cities (e.g., zone and local centres) as residents tend to travel to neighbouring cities rather than travel large distances (Figure 3b). There was only one metropolis located in the North region,

Manaus. Although Manaus is far from being classified as the most influential city in the urban network, it has the largest average distance travelled to reach the city (316km, over double the distance travelled to the second highest which is Brasilia at 145km). Despite this huge reach, Manaus' region of influence has the lowest GDP of any metropolis (less than R$100 billion annually) [47].

Human movement is an important consideration when modelling any infectious disease as movement drives contact between susceptible populations and infectious hosts. The contribution of human movement to disease transmission depends on the timescale and mode of transmission, and the geographical scale of interest [69,70]. Studies have shown that simplifying human movement behaviour, either by assuming no movement or random mixing, can lead to inaccurate and exaggerated inferences about mosquito-borne disease transmission patterns [71,72]. Within Brazil, there is evidence to suggest that long-distance travel between cities has contributed to the expansion of dengue outbreaks across the country [42,48,73] and that intra-city movement is important in predicting outbreaks at a local level [74].

1.4 Statistical models of emerging infectious diseases

Statistical models of emerging infectious diseases allow us to identify key drivers of disease expansion and predict regions at risk of future emergence based on observed data. Models can vary in complexity depending on the nature of relationships between the response and explanatory variables, including nonlinear and lagged variables where necessary. This is particularly important when investigating the impact of climatic change on arbovirus expansion, which is complex, nonlinear and often delayed by periods of up to several months [62,75]. When considering disease emergence across a geographical area, it is also important to consider the underlying spatial structure of the data, in particular spatial connectivity that can play an important role in the importation of diseases from endemic areas with ongoing transmission to disease-free areas [69,76,77].

Statistical models assume some underlying probability model of the outcome described by unobserved parameters. These parameters are estimated via an inferential approach using observed data. There are many different approaches to making these inferences, this thesis uses a Bayesian approach. Bayesian inference is a branch of statistics based on Bayes theorem

(Equation 1), which allows the inclusion of some prior belief in the model fitting process. In Bayesian analysis, parameter estimates and model predictions are based on the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, the probability of parameters $\boldsymbol{\theta} = \theta_1, \theta_2, ...$ conditional on data $\boldsymbol{y}$. This posterior distribution represents our current knowledge of the problem by updating the prior belief (represented by the prior distribution $p(\boldsymbol{\theta})$) using observed data (represented by the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$) [78,79]:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})} \qquad (1)$$

Where $p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ are continuous.

Bayesian hierarchical models are a common approach used to model spatial and spatiotemporal data as expected spatial and/or temporal structures can be incorporated into their prior distribution [78,80,81]. Random terms are included in the model to account for residual autocorrelation and/or account for pooling within the data. In spatial models for emerging diseases, this autocorrelation can arise due to unobserved shared characteristics between close areas (for example, due to similar climates, vector-control programmes, and levels of immunity within communities) or from human and vector movement creating connections between areas. The prior distribution of these random terms is defined using some unknown hyperparameter(s), $\boldsymbol{\varphi}$, which has its own prior distribution and can be defined using some spatial and/or temporal structure (for examples, see [78–83]). The prior distribution therefore becomes a joint prior, defining the prior belief of model parameters $\boldsymbol{\theta}$ and their hyperparameters $\boldsymbol{\varphi}$, $p(\boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\boldsymbol{\theta}|\boldsymbol{\varphi})p(\boldsymbol{\varphi})$. The posterior distribution defined in Equation 1 can be rewritten as:

$$p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\varphi})p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\varphi})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{\varphi})p(\boldsymbol{\varphi})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})} \qquad (2)$$

Note that $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ as the data distribution is only dependent on the model parameters $\boldsymbol{\theta}$, as the hyperparameters only affect the data through $\boldsymbol{\theta}$.

1.4.1 Bayesian hierarchical model estimation

Although the formulation of the posterior distribution is generally well defined (Equation 2), the multi-dimensional integral in the denominator is often hard, if not impossible, to calculate. There are several approaches that have been developed to aid the derivation of the posterior distribution which can broadly be classified into three groups: simulation-based, approximation, and empirical approaches.

Simulation-based Bayesian inference

Markov chain Monte Carlo (MCMC) methods are a collection of algorithms used to simulate samples from probability distributions. Within Bayesian statistics, they remove the need to calculate the complex integral in the denominator of the posterior distribution (Equation 2). They work by drawing samples of parameters $\boldsymbol{\theta}$ from approximate distributions and then correct these to obtain better approximations from the target posterior distribution [78]. MCMC is an iterative process, with each sampled value depending on the last value drawn, forming a Markov chain. As the approximate distribution is improved at each step, the basis of MCMC is that it will eventually converge to the target population, in this case the posterior distribution.

One of the main advantages of MCMC is that we do not need to know the full form of the distribution we are sampling from (only up to the normalising constant), allowing estimation of very complex distributions with many parameters (as is often the case when including random terms into a model). The development of the user-friendly, flexible BUGS language [87] and the programmes inspired by this [87–91] have made implementation more accessible and efficient than ever. In particular, the NIMBLE package allows users to define novel modelling frameworks and sampling algorithms which allows inference beyond general models included in other software packages [91,92]. Despite these advances, MCMC methods are still extremely computationally intensive and issues can arise with convergence.

Approximation methods for Bayesian inference

Integrated nested Laplace approximation (INLA) presents a less computationally intensive alternative to Bayesian simulation approaches [93,94]. INLA performs approximate Bayesian inference on a class of models known as latent Gaussian models (LGMs) and returns an approximation of the posterior distribution. LGMs are a class of additive models that aim to explain the relationship between an outcome of interest and covariates whilst accounting for some unobservable (latent) structure, including hierarchical models (where the latent structure can define spatial and/or temporal autocorrelation [81,82]). The latent structure, consisting of

unobserved parameter and hyperparameters, is assumed to follow a Gaussian Markov random field (GMRF). That is, a random field that follows a multivariate Gaussian distribution and satisfies the Markov property of conditional independence [95]. A common example of a latent structure that satisfies the GMRF assumption is the conditional autoregressive (CAR) spatial structure where regions are considered connected if and only if they share a border [84]. This GMRF structure means that the parameters and hyperparameters will have a sparce precision matrix, speeding up computation. Under these conditions, INLA can calculate the joint posterior distribution of the latent field (model parameters and hyperparameter) and obtain the posterior marginal distributions using Laplace approximation, a mathematical approach used to estimate integrals of the form $\int_a^b e^{Mf(x)}dx$ [93].

Bayesian inference using INLA can be implemented using the R-INLA package which provides a range of pre-built options to specify model priors, including spatial and temporal structures [83,85,94,96]. R-INLA includes the most commonly used prior structures, including CAR models. However, this approximation method lacks the flexibility of MCMC approaches, which allow users to fully specify the model.

Empirical Bayesian inference

Both MCMC and INLA are examples of fully Bayesian approaches which require users to specify their prior beliefs of structures within the data before model fitting. Empirical approaches allow these underlying structures to be estimated as part of the model fitting process using the data. This is particularly useful in scenarios where the spatial structure of the data is not fully understood. Although generalised additive models (GAMs) are typically considered a frequentist modelling approach, they can be extended to include smooth spatial and/or temporal random terms which can be interpreted from a Bayesian perspective [97]. This class of models are also known as structured additive regression (STAR) models [98].

Smooth terms are generated by applying smoothing splines to some representation of the structure within the data (e.g., coordinates if assuming spatial connectivity arising between close observations). Smoothing splines are estimated using restricted maximum likelihood (REML) which imposes a penalty matrix, ensuring that they are not ‘overly wiggly’ [99]. This penalty matrix can be viewed as a prior belief from a Bayesian perspective and can be formalised into a prior distribution [97]. The mgcv package in R can fit these models and

generate simulations from the posterior distribution of model parameters [100]. The empirical smoothing spline approach presents an efficient, flexible alternative to fully Bayesian approaches when the structure of the data is not fully understood. A more detailed description of this approach can be found in Appendix A.

1.5 Aim

The overall aim of this research is to understand the complex, interacting drivers of dengue expansion in Brazil. In particular, the goal is to understand the contribution of increasing temperatures in South Brazil and connectivity between cities arising from human movement to the expansion of the dengue transmission zone in Brazil.

1.5.1 Objectives

To achieve this aim, I will address the following objectives:

1. Explore the impact of temperature suitability, urbanisation, and connectivity of cities to the Brazilian urban network on the expansion of the dengue transmission zone in Brazil.
2. Identify spatial modelling techniques currently used to study mosquito-borne disease transmission and the assumptions made in modelling studies about how spatial connectivity arises, and describe the data used to inform spatial models.
3. Develop a statistical modelling framework capable of accounting for multiple sources of spatial connectivity and quantifying the relative contribution of each source to the overall spatial structure of the data.
4. Quantify the relative contribution of human movement to the expansion of dengue outbreaks in Brazil.

1.6 Thesis structure

This thesis is written in a research paper style, where analysis chapters are written in the style of a scientific paper for publication and are preceded by a bridging section. This thesis includes four analysis chapters, preceded by the current introductory chapter, followed by a discussion section. The remaining chapters are as follows:

Chapter 2: The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. This chapter was published *in PLOS Neglected Tropical Diseases* in December 2021 [101] and addresses Objective 1. It uses a Bayesian spatiotemporal model to explore the association between dengue outbreaks in Brazil between 2001 – 2020, and temperature suitability, urbanisation and level of influence of cities in the Brazilian urban network. The paper also quantifies the relative contribution of each factor in turn and provides refined geographical barriers and updated limits to the dengue transmission zone in Brazil based on new data and modelling results.

Chapter 3: Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. This chapter was published in the *Journal of the Royal Society Interface* in May 2021 [102]. This systematic review addresses Objective 2 by synthesising the spatial methods described in the literature to model mosquito-borne diseases, their spatial connectivity assumptions and the data used to inform spatial model components.

Chapter 4: A Bayesian model framework to quantify multiple sources of spatial variation for disease mapping. This chapter was published in the *Journal of the Royal Society Interface* in September 2022 [103]. The paper addresses Objective 3 and presents a novel modelling framework which allows multiple sources of spatial connectivity to be included within a statistical model (in this case, autocorrelation between close regions and connections arising due to human movement). Through Bayesian inference and simulations, the relative contribution of each spatial connectivity element to the overall model structure can be quantified. This chapter contains model derivation, results of simulation studies, and a case study in which this method is applied to dengue incidence data from South Brazil.

Chapter 5: Quantifying the relative contribution of human movement to the expansion of dengue outbreaks in Brazil. This paper is currently being prepared for submission and addresses Objective 4. The modelling framework presented in Chapter 4 is applied to dengue outbreak data between 2001 – 2020 from Brazil. The model includes spatially structured terms to account for spatial autocorrelation between close regions and connections arising due to human movement based on commuting. The relative contribution of the human movement-based term is produced for Brazil as a whole, and for each region separately.

Chapter 6: Discussion. This chapter provides an overall conclusion of the thesis and presents strengths and limitations of the work carried out. Suggestions for future work are also presented.

References

1. Gulland A. 2016 Zika virus is a global public health emergency, declares WHO.

2. Lowe R, Barcellos C, Brasil P, Cruz OG, Honório NA, Kuper H, Carvalho MS. 2018 The Zika Virus Epidemic in Brazil: From Discovery to Future Implications. *Int. J. Environ. Res. Public. Health* **15**.

3. World Health Organization. 2016 WHO: Ebola situation report.

4. Coltart CE, Lindsey B, Ghinai I, Johnson AM, Heymann DL. 2017 The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160297.

5. Mahase E. 2020 Covid-19: WHO declares pandemic because of "alarming levels" of spread, severity, and inaction. *Bmj* **368**, 1036.

6. Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM, Eskew EA, Olival KJ, Ross N, Bansal S. 2022 Climate change increases cross-species viral transmission risk. *Nature* , 1–8.

7. Baker RE *et al.* 2022 Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205.

8. Metcalf CJE, Lessler J. 2017 Opportunities and challenges in modeling emerging infectious diseases. *Science* **357**, 149–152.

9. Wilder-Smith A, Ooi E-E, Horstick O, Wills B. 2019 Dengue. *The Lancet* **393**, 350–363.

10. Stanaway JD *et al.* 2016 The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *Lancet Infect. Dis.* **16**, 712–723.

11. Kraemer MU *et al.* 2015 The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife* **4**.

12. Kraemer MUG *et al.* 2019 Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nat. Microbiol.* **4**, 854–863.

13. Yang X, Quam MB, Zhang T, Sang S. 2021 Global burden for dengue and the evolving pattern in the past 30 years. *J. Travel Med.* **28**.

14. Gubler DJ. 2011 Dengue, Urbanization and Globalization: The Unholy Trinity of the 21st Century. *Trop. Med. Health* **39**, 3–11.

15. Kolimenakis A, Heinz S, Wilson ML, Winkler V, Yakob L, Michaelakis A, Papachristos D, Richardson C, Horstick O. 2021 The role of urbanisation in the spread of Aedes mosquitoes and the diseases they transmit—A systematic review. *PLoS Negl. Trop. Dis.* **15**, e0009631.

16. Lwande OW, Obanda V, Lindström A, Ahlm C, Evander M, Näslund J, Bucht G. 2020 Globe-trotting Aedes aegypti and Aedes albopictus: risk factors for arbovirus pandemics. *Vector-Borne Zoonotic Dis.* **20**, 71–81.

17. Reinhold JM, Lazzari CR, Lahondère C. 2018 Effects of the Environmental Temperature on Aedes aegypti and Aedes albopictus Mosquitoes: A Review. *Insects* **9**.

18. Bhatt S *et al.* 2013 The global distribution and burden of dengue. *Nature* **496**, 504–507.

19. World Health Organization. 2009 Dengue guidelines for diagnosis, treatment, prevention and control: new edition.

20. World Health Organization. In press. Dengue and severe dengue. See https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed on 9 June 2022).

21. Zeng Z, Zhan J, Chen L, Chen H, Cheng S. 2021 Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017. *EClinicalMedicine* **32**, 100712.

22. Bowman LR, Donegan S, McCall PJ. 2016 Is dengue vector control deficient in effectiveness or evidence?: Systematic review and meta-analysis. *PLoS Negl. Trop. Dis.* **10**, e0004551.

23. Reiner Jr RC *et al.* 2016 Quantifying the epidemiological impact of vector control on dengue. *PLoS Negl. Trop. Dis.* **10**, e0004588.

24. Villar L *et al.* 2015 Efficacy of a tetravalent dengue vaccine in children in Latin America. *N. Engl. J. Med.* **372**, 113–123.

25. Wilder-Smith A. 2020 Dengue vaccine development: status and future. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* **63**, 40–44.

26. Sridhar S *et al.* 2018 Effect of dengue serostatus on dengue vaccine safety and efficacy. *N. Engl. J. Med.* **379**, 327–340.

27. World Health Organization. 2019 Dengue vaccine: WHO position paper, September 2018-Recommendations. *Vaccine* **37**, 4848–4849.

28. Moreira LA *et al.* 2009 A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya, and Plasmodium. *Cell* **139**, 1268–1278.

29. Flores HA, O'Neill SL. 2018 Controlling vector-borne diseases by releasing modified mosquitoes. *Nat. Rev. Microbiol.* **16**, 508–518.

30. Ferguson NM *et al.* 2015 Modeling the impact on virus transmission of Wolbachia-mediated blocking of dengue virus infection of Aedes aegypti. *Sci. Transl. Med.* **7**.

31. O'Reilly KM *et al.* 2019 Estimating the burden of dengue and the impact of release of wMel Wolbachia-infected mosquitoes in Indonesia: a modelling study. *BMC Med.* **17**, 1–14.

32. Utarini A *et al.* 2021 Efficacy of Wolbachia-infected mosquito deployments for the control of dengue. *N. Engl. J. Med.* **384**, 2177–2186.

33. Indriani C *et al.* 2020 Reduced dengue incidence following deployments of Wolbachia-infected Aedes aegypti in Yogyakarta, Indonesia: a quasi-experimental trial using controlled interrupted time series analysis. *Gates Open Res.* **4**.

34. Ryan PA *et al.* 2019 Establishment of wMel Wolbachia in Aedes aegypti mosquitoes and reduction of local dengue transmission in Cairns and surrounding locations in northern Queensland, Australia. *Gates Open Res.* **3**.

35. O'Neill SL *et al.* 2018 Scaled deployment of Wolbachia to protect the community from dengue and other Aedes transmitted arboviruses. *Gates Open Res.* **2**.

36. Pinto SB *et al.* 2021 Effectiveness of Wolbachia-infected mosquito deployments in reducing the incidence of dengue and other Aedes-borne diseases in Niterói, Brazil: A quasi-experimental study. *PLoS Negl. Trop. Dis.* **15**, e0009556.

37. Martins ABS, Alencar CH. 2022 Ecoepidemiology of dengue in Brazil: from the virus to the environment. *One Health Implement. Res.* **2**, 1–14.

38. Fares RCG, Souza KPR, Añez G, Rios M. 2015 Epidemiological Scenario of Dengue in Brazil. *BioMed Res. Int.* **2015**, e321873.

39. Nunes PCG, Daumas RP, Sánchez-Arcila JC, Nogueira RMR, Horta MAP, dos Santos FB. 2019 30 years of fatal dengue cases in Brazil: a review. *BMC Public Health* **19**, 329.

40. Andrioli DC, Busato MA, Lutinski JA. 2020 Spatial and temporal distribution of dengue in Brazil, 1990-2017. *PLoS One* **15**, e0228346.

41. de Azevedo TS, Lorenz C, Chiaravalloti-Neto F. 2020 Spatiotemporal evolution of dengue outbreaks in Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **114**, 593–602.

42. Barcellos C, Lowe R. 2014 Expansion of the dengue transmission area in Brazil: the role of climate and cities. *Trop. Med. Int. Health* **19**, 159–168.

43. Mordecai EA *et al.* 2017 Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLoS Negl. Trop. Dis.* **11**, e0005568.

44. Mordecai EA *et al.* 2019 Thermal biology of mosquito-borne disease. *Ecol. Lett.* **22**, 1690–1708.

45. Dubreuil V, Fante K, Planchon O, Sant'anna Neto J. 2019 Climate change evidence in Brazil from Köppen's climate annual types frequency. *Int. J. Climatol.* **39**, 1446–1456.

46. Secretaria de Vigilância em Saúde. 2022 Monitoramento dos casos de arboviroses até a semana epidemiológica 18 de 2022.

47. IBGE (Instituto Brasileiro de Geografia e Estatística). 2020 *Regiões de influência das cidades 2018*.

48. Lana RM, Gomes MF da C, de Lima TFM, Honório NA, Codeço CT. 2017 The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: A network-based analysis. *PLoS Negl. Trop. Dis.* **11**.

49. Lowe R, Lee S, Lana RM, Codeço CT, Castro MC, Pascual M. 2020 Emerging arboviruses in the urbanized Amazon rainforest. *BMJ* **371**, m4385.

50. BRASIL. 2020 Ministério da Saúde. Guia de Vigilância em Saúde.

51. Coelho GE, Leal PL, Cerroni M de P, Simplicio ACR, Siqueira Jr JB. 2016 Sensitivity of the dengue surveillance system in Brazil for detecting hospitalized cases. *PLoS Negl. Trop. Dis.* **10**, e0004705.

52. Silva MMO *et al.* 2016 Accuracy of Dengue Reporting by National Surveillance System, Brazil. *Emerg. Infect. Dis.* **22**, 336–339.

53. Angelo M, Ramalho WM, Gurgel H, Belle N, Pilot E. 2020 Dengue surveillance system in Brazil: A qualitative study in the federal district. *Int. J. Environ. Res. Public. Health* **17**, 2062.

54. IBGE (Instituto Brasileiro de Geografia e Estatística). 2019 Biomas e Sistema Costeiro-Marinho do Brasil.

55. Köppen W. 1884 Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet. *Meteorol. Z.* **1**, 5–226.

56. Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G. 2013 Köppen's climate classification map for Brazil. *Meteorol. Z.* , 711–728.

57. Huber JH, Childs ML, Caldwell JM, Mordecai EA. 2018 Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. *PLoS Negl. Trop. Dis.* **12**, e0006451.

58. Benedum CM, Seidahmed OME, Eltahir EAB, Markuzon N. 2018 Statistical modeling of the effect of rainfall flushing on dengue transmission in Singapore. *PLoS Negl. Trop. Dis.* **12**, e0006935.

59. Koenraadt C, Harrington L. 2008 Flushing effect of rain on container-inhabiting mosquitoes Aedes aegypti and Culex pipiens (Diptera: Culicidae). *J. Med. Entomol.* **45**, 28–35.

60. Martin JL, Lippi CA, Stewart-Ibarra AM, Ayala EB, Mordecai EA, Sippy R, Heras FH, Blackburn JK, Ryan SJ. 2021 Household and climate factors influence Aedes aegypti presence in the arid city of Huaquillas, Ecuador. *PLoS Negl. Trop. Dis.* **15**, e0009931.

61. Akanda AS, Johnson K, Ginsberg HS, Couret J. 2020 Prioritizing Water Security in the Management of Vector-Borne Diseases: Lessons From Oaxaca, Mexico. *GeoHealth* **4**, e2019GH000201.

62. Lowe, R., Lee, S.A., O'Reilly, K.M., Brady, O.J., Bastos, L., Carrasco-Escobar, G., de Castro Catão, R., Colón-González, F.J., Barcellos, C., Carvalho, M.S., Blangiardo, M. 2021 Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *Lancet Planet. Health* **5**, e209–e219.

63. Pérez-Castro R, Castellanos JE, Olano VA, Matiz MI, Jaramillo JF, Vargas SL, Sarmiento DM, Stenström TA, Overgaard HJ. 2016 Detection of all four dengue serotypes in Aedes aegypti female mosquitoes collected in a rural area in Colombia. *Mem. Inst. Oswaldo Cruz* **111**, 233–240.

64. Guagliardo SA, Morrison AC, Barboza JL, Requena E, Astete H, Vazquez-Prokopec G, Kitron U. 2015 River Boats Contribute to the Regional Spread of the Dengue Vector Aedes aegypti in the Peruvian Amazon. *PLoS Negl. Trop. Dis.* **9**, e0003648.

65. IBGE (Instituto Brasileiro de Geografia e Estatística). 2012 Censo demográfico 2010: resultados gerais da Amostra.

66. Soriano É, Londe L de R, Di Gregorio LT, Coutinho MP, Santos LBL. 2016 Water crisis in São Paulo evaluated under the disaster's point of view. *Ambiente Soc.* **19**, 21–42.

67. Cavalcanti LP de G, Oliveira R de MAB, Alencar CH. 2016 Changes in infestation sites of female *Aedes aegypti* in Northeast Brazil. *Rev. Soc. Bras. Med. Trop.* **49**, 498–501.

68. IBGE (Instituto Brasileiro de Geografia e Estatística). 2008 *Regiões de influência das cidades 2007*.

69. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis* **3**, e481.

70. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V. 2014 On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716.

71. Massaro E, Kondor D, Ratti C. 2019 Assessing the interplay between human mobility and mosquito borne diseases in urban environments. *Sci. Rep.* **9**, 1–13.

72. Kraemer MUG *et al.* 2018 Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan. *Epj Data Sci.* **7**, 17.

73. Churakov M, Villabona-Arenas CJ, Kraemer MUG, Salje H, Cauchemez S. 2019 Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony. *PLoS Negl. Trop. Dis.* **13**.

74. Bomfim R, Pei S, Shaman J, Yamana T, Makse HA, Andrade JS, Lima Neto AS, Furtado V. 2020 Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *J. R. Soc. Interface* **17**, 20200691.

75. Lowe, R., Gasparrini, A., Van Meerbeeck, C.J., Lippi, C.A., Mahon, R., Trotman, A.R., Rollock, L., Hinds, A.Q., Ryan, S.J., Stewart-Ibarra, A.M. 2018 Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. *PLOS Med.* **15**, e1002613.

76. Kraemer, M.U., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D.M., Ray, S.E., Brady, O.J., Brownstein, J.S., Faria, N.R., Cummings, D.A.T., Pybus, O.G. 2019 Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 1–11.

77. Findlater A, Bogoch II. 2018 Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends Parasitol.* **34**, 772–783.

78. Gelman A, Carlin JB, Stern HS, Rubin DB. 1995 *Bayesian data analysis*. Chapman and Hall/CRC.

79. van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., Yau, C. 2021 Bayesian statistics and modelling. *Nat. Rev. Methods Primer* **1**, 1–26.

80. Banerjee S, Carlin BP, Gelfand AE. 2003 *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

81. Moraga P. 2019 *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press.

82. Gómez-Rubio V. 2020 *Bayesian inference with INLA*. CRC Press.

83. Riebler A, Sørbye SH, Simpson D, Rue H. 2016 An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25**, 1145–1165.

84. Besag J. 1974 Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Methodol.* **36**, 192–225.

85. Besag J, York J, Mollié A. 1991 Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20.

86. Lee D. 2011 A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat. Spatio-Temporal Epidemiol.* **2**, 79–89.

87. Gilks WR, Thomas A, Spiegelhalter DJ. 1994 A language and program for complex Bayesian modelling. *J. R. Stat. Soc. Ser. Stat.* **43**, 169–177.

88. Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000 WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337.

89. Thomas A, O'Hara R, Ligges U, Sturtz S. 2006 Making BUGS open. *R News* **6**, 12–17.

90. Plummer M. 2003 JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. pp. 1–10. Vienna, Austria.

91. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. 2017 Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.* **26**, 403–413.

92. De Valpine P *et al.* 2021 Nimble: MCMC, particle filtering, and programmable hierarchical modeling. *R Package Version 011* **1**.

93. Rue H, Martino S, Chopin N. 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 319–392.

94. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. 2017 Bayesian computing with INLA: a review. *Annu. Rev. Stat. Its Appl.* **4**, 395–421.

95. Rue H, Held L. 2005 *Gaussian Markov random fields: theory and applications*.

96. Lee D. 2011 A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat. Spatio-Temporal Epidemiol.* **2**, 79–89.

97. Wood SN. 2017 *Generalized additive models: an introduction with R*.

98. Fahrmeir L, Kneib T, Lang S. 2004 Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat. Sin.* , 731–761.

99. Wood SN. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36.

100. Wood S, Wood MS. 2015 Package 'mgcv'. *R Package Version* **1**, 729.

101. Lee SA, Economou T, Catão R de C, Barcellos C, Lowe R. 2021 The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. *PLoS Negl. Trop. Dis.* **15**, e0009773.

102. Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021. Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**, 20210096.

103. Lee SA, Economou T, Lowe R. 2022 A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping. *J. R. Soc. Interface* **19,** 20220440.

2. The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil

Bridging section

In this chapter, I present a published research study, which explores the impact of temperature suitability, urbanisation, and connectivity to the Brazilian urban network on the expansion of the dengue transmission zone in Brazil between 2001 and 2020 (Objective 1).

Previous studies identified geographical barriers to dengue transmission, beyond which areas of the country were relatively protected from outbreaks[1]. These barriers included South Brazil, a temperate part of the country which experiences cold winter temperatures, too cold for *Aedes* mosquitoes to effectively transmit dengue virus to humans. The other dengue transmission barrier existed in the western Amazon region of Brazil, a relatively remote area of the country with many municipalities only accessible by long boat journeys. More recent epidemiological reports showed that these barriers have been eroded and dengue outbreaks now occur beyond these previous barriers[2].

This chapter uses dengue incidence data from the Brazilian surveillance system between 2001 – 2020 to redefine the geographical limits of the dengue transmission zone in Brazil[1] and investigate potential drivers of the erosion of transmission barriers. I use a Bayesian spatiotemporal model to quantify the role of increased temperatures in South Brazil and increased connectivity to and within the western Amazon region in the expansion of the dengue transmission zone.

This chapter was published in *PLOS Neglected Tropical Diseases* in December, 2021[3]. I have included the published version of this paper. Supplementary materials referred to in the paper can be found in Appendix B.

[1.] Barcellos C, Lowe R. 2014 Expansion of the dengue transmission area in Brazil: the role of climate and cities. *Trop. Med. Int. Health* **19**, 159–168. (doi:10.1111/tmi.12227)

[2.] Secretaria de Vigilância em Saúde. 2022 Monitoramento dos casos de arboviroses até a semana epidemiológica 18 de 2022.

[3.] Lee SA, Economou T, Catão R de C, Barcellos C, Lowe R. 2021 The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. *PLoS Negl. Trop. Dis.* **15**, e0009773. (doi:10.1371/journal.pntd.0009773)

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1806358	**Title**	Ms
First Name(s)	Sophie Alice		
Surname/Family Name	Lee		
Thesis Title	Spatial Modelling of Emerging Infectious Diseases: Quantifying the Role of Climate, Cities and Connectivity on Dengue Expansion in Brazil		
Primary Supervisor	Dr Rachel Lowe		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	PLOS Neglected Tropical Diseases		
When was the work published?	9 December 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My contribution: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing – original draft, Writing – review & editing

SECTION E

Student Signature	Sophie Lee
Date	08/09/2022

Supervisor Signature	Rachel Lowe
Date	08/09/2022

PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil

Sophie A. Lee[1,2]*, **Theodoros Economou**[3], **Rafael de Castro Catão**[4], **Christovam Barcellos**[5], **Rachel Lowe**[1,2,6]

1 Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, **2** Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, United Kingdom, **3** Climate and Atmosphere Research Centre, The Cyprus Institute, Nicosia, Cyprus, **4** Departamento de Geografia, Universidade Federal do Espírito Santo, Vitoria, Brazil, **5** Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, **6** Barcelona Supercomputing Center, Barcelona, Spain

* sophie.lee@lshtm.ac.uk

Citation: Lee SA, Economou T, de Castro Catão R, Barcellos C, Lowe R (2021) The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. PLoS Negl Trop Dis 15(12): e0009773. https://doi.org/10.1371/journal.pntd.0009773

Editor: Hannah E. Clapham, National University Singapore Saw Swee Hock School of Public Health, SINGAPORE

Received: August 27, 2021

Accepted: November 24, 2021

Published: December 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pntd.0009773

Data Availability Statement: All data used in this study is open access and available freely on the internet, see the methods section for more details.

Abstract

Dengue is hyperendemic in Brazil, with outbreaks affecting all regions. Previous studies identified geographical barriers to dengue transmission in Brazil, beyond which certain areas, such as South Brazil and the Amazon rainforest, were relatively protected from outbreaks. Recent data shows these barriers are being eroded. In this study, we explore the drivers of this expansion and identify the current limits to the dengue transmission zone. We used a spatio-temporal additive model to explore the associations between dengue outbreaks and temperature suitability, urbanisation, and connectivity to the Brazilian urban network. The model was applied to a binary outbreak indicator, assuming the official threshold value of 300 cases per 100,000 residents, for Brazil's municipalities between 2001 and 2020. We found a nonlinear relationship between higher levels of connectivity to the Brazilian urban network and the odds of an outbreak, with lower odds in metropoles compared to regional capitals. The number of months per year with suitable temperature conditions for *Aedes* mosquitoes was positively associated with the dengue outbreak occurrence. Temperature suitability explained most interannual and spatial variation in South Brazil, confirming this geographical barrier is influenced by lower seasonal temperatures. Municipalities that had experienced an outbreak previously had double the odds of subsequent outbreaks. We identified geographical barriers to dengue transmission in South Brazil, western Amazon, and along the northern coast of Brazil. Although a southern barrier still exists, it has shifted south, and the Amazon no longer has a clear boundary. Few areas of Brazil remain protected from dengue outbreaks. Communities living on the edge of previous barriers are particularly susceptible to future outbreaks as they lack immunity. Control strategies should target regions at risk of future outbreaks as well as those currently within the dengue transmission zone.

Data and code used to produce this analysis is available from a Github repository cited in the manuscript (https://github.com/sophie-a-lee/Dengue_expansion).

Funding: S.A.L. was supported by a Royal Society Research Grant for Research Fellows. https://royalsociety.org/. R.L. was supported by a Royal Society Dorothy Hodgkin Fellowship https://royalsociety.org/. T.E. was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 856612 https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en and the Cyprus Government. C.B. was funded by the National Council for Scientific and Technological Development under grant No. 303985/2019-4 (http://www.cnpq.br/) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Dengue is a mosquito-borne disease that has expanded rapidly around the world due to increased urbanisation, global mobility and climate change. In Brazil, geographical barriers to dengue transmission exist, beyond which certain areas including South Brazil and the Amazon rainforest are relatively protected from outbreaks. However, we found that the previous barrier in South Brazil has shifted further south as a result of increased temperature suitability. The previously identified barrier protecting the western Amazon no longer exists. This is particularly concerning as we found dengue outbreaks tend to become established in areas after introduction. Highly influential cities with many transport links had increased odds of an outbreak. However, the most influential cities had lower odds of an outbreak than cities connected regionally. This study highlights the importance of monitoring the expansion of dengue outbreaks and designing disease prevention strategies for areas at risk of future outbreaks as well as areas in the established dengue transmission zone.

Introduction

Dengue is considered one of the top 10 threats to global health [1], with around half the world's population living in areas at risk of infection [2]. Incidence rates have doubled each decade in the past 30 years as a result of increased urbanisation, global mobility and climate change [2–4]. All 4 dengue serotypes are endemic to Brazil, which experiences frequent outbreaks across the country [5]. Previous studies identified geographical barriers to dengue transmission beyond which regions were relatively protected. This included South Brazil, where seasonal temperatures are too cold for vectors to efficiently transmit the virus, areas of high altitude in Southeast Brazil and remote regions of the western Amazon [6]. However, these barriers are being eroded and the dengue transmission area in Brazil has expanded over the past decade. This expansion is thought to be linked to increased human mobility and changes in climate [7,8].

For dengue to become established in a new region, the environment must be suitable to support the propagation of the dengue vector, *Aedes* mosquitoes. There are two vectors present in Brazil capable of transmitting the dengue virus: *Aedes aegypti* and *Aedes albopictus*. Currently only *Aedes aegypti* are considered responsible for dengue transmission in Brazil [9,10], however a recent study identified *Aedes albopictus* infected by dengue virus in a rural area of Brazil during an outbreak, which could indicate their involvement in the introduction of dengue to rural areas [11]. *Aedes aegypti* have evolved to live in urban environments close to humans [12] but there is evidence to suggest they are becoming established in peri-urban and rural regions of South America [13,14]. Conversely, *Aedes albopictus* are typically found in peri-urban areas but have been identified in densely urbanised areas such as urban slums in Brazil [9,15]. *Aedes* mosquitoes breed in pools of standing, clean water created by water storage containers or uncollected refuse. These conditions arise when rapid urbanisation occurs without adequate improvements to infrastructure, such as access to piped water and refuse collection [16,17]. There is evidence that areas lacking reliable access to piped water are more susceptible to dengue outbreaks, particularly in highly urbanised areas following drought [18]. Prior studies have found that extremely wet conditions also increased the risk of dengue outbreaks, thought to be linked to the creation of larval habitat in the short term [18,19]. Suitable temperature conditions are required for the mosquitoes to breed and transmit the virus. *Aedes aegypti* are unable to survive in temperatures below 10˚C or above 40˚C [20] and can only

transmit the virus between 17.8˚C and 34.5˚C [21,22]. *Aedes albopictus* are more suited to cooler temperatures and can transmit the virus between 16.2˚C and 31.4˚C [21,22]. Recent outbreaks in temperate cities of South America have shown that epidemics are still possible in regions that experience seasonal temperatures outside of this range due to human movement [23–25].

The expansion of *Aedes aegypti* and the arboviruses they transmit into rural parts of the Amazon has been linked to connections to and within the area by air, road or boat [13,26]. Despite this, the investigation of spatial connections created by human movement is little explored in the literature and the vast majority of spatial modelling studies of mosquito-borne diseases assume connectivity is based on distance alone [27]. Brazilian cities are connected to one another within a complex urban network, described within the Regions of influence of cities ("Regiões de Influência das Cidades", REGIC) studies carried out by the Brazilian Institute of Geography and Statistics [28,29]. People often travel great distances to reach large urban centres as they contain important educational, business or cultural institutions. Failure to account for long-distance movements may miss important drivers of dengue expansion, particularly in areas such as the Amazon where the average distance travelled to Manaus, the capital of Amazonas state, was 316km. Important cities can have influence over vast areas of Brazil, for example the region of influence connected to the capital city of Brasilia corresponds to over 20% of the country and spans 1.8 million km^2 [29].

Although previous studies have shown the expansion of dengue outbreaks in Brazil [7] and the association between the dengue transmission zone and climate [6], neither formally investigated the link between this expansion and human movement. In this study, we use the level of influence of cities from the REGIC studies [28,29] as a proxy for human movement, and aim to better understand how climate suitability, connectivity between cities and socioeconomic factors have contributed to the recent expansion of dengue. It is hoped that by understanding the drivers of dengue expansion in Brazil, we can identify its spatial trends and regions at risk from future outbreaks.

Methods

Epidemiological data

Brazil is the 6th most populous country in the world with an estimated population of over 212 million in 2020 [30]. The country can be separated into 5 distinct geo-political regions (S1 Fig), 27 federal units (26 states and a federal district containing the capital city Brasilia, S1 Fig), and 5,570 municipalities. We obtained monthly notified dengue cases for each of Brazil's 5,570 municipalities between January 2001 and December 2020 from Brazil's Notifiable Diseases Information System (SINAN), freely available via the Health Information Department, DATASUS (https://datasus.saude.gov.br/informacoes-de-saude-tabnet/). Cases were aggregated by month of first symptom and municipality of residence. Dengue cases are considered confirmed if they test positive in a laboratory or, more commonly, based on the Ministry of Health's syndromic definition Due to its passive nature, the accuracy of the dengue surveillance system differs between municipalities and between periods of high and low incidence [31]. To reduce the bias introduced by differences in case reporting and health seeking behaviour, we chose to model binary outbreak indicators rather than incidence rates because, as stated by the Brazilian Observatory of Climate and Health, "there is no way to hide an epidemic" [6]. Between 2001 and 2020, municipality boundaries in Brazil have changed and several new municipalities were created. To ensure data were consistent over the study period, we aggregated data to the 5,560 municipalities that were present in 2001 by combining the new municipalities with their parent municipalities. The data and code used to aggregate the dengue case data are available from https://github.com/sophie-a-lee/Dengue_expansion [32].

Meteorological data

Monthly mean temperatures (K) were obtained from the European Centre for Medium-Range Weather Forecasts' (ECMRWF) ERA5-Land dataset [33] for the period January 2001—December 2020, at a spatial resolution of 0.1˚ x 0.1˚ (~9km). The ERA5-Land database was chosen because of its fine spatial scale, necessary when analysing small administrative units such as municipalities. Temperatures were converted from Kelvin to degrees Celsius (˚C) by subtracting 273.15. Mean temperature was aggregated to each municipality using the exactextractr package [34] in R (version 4.0.3) by calculating the mean of the grid boxes lying within each municipality. Grid boxes partially covered by a municipality were weighted by the percentage of area that lay within the municipality.

Due to its size, Brazil experiences a wide range of climate systems and ecosystems. The northern part of the country lies on or close to the equator, meaning regions experience year-round high temperatures. In contrast, the South and Southeast regions have clear seasonality in temperatures with cooler winters (S2 Fig), often falling below the optimal temperature range for dengue transmission (between 17.8˚C and 34.5˚C for *Aedes aegypti* and 16.2˚C and 31.4˚C for *Aedes albopictus* [21,22]). To understand how temperature suitability has contributed to the expansion of the dengue transmission zone in Brazil, we calculated the number of months per year each municipality lay within the suitable temperature ranges (between 16.2˚C and 34.5˚C). Most of Brazil experiences year-round temperature suitability except for the temperate South and mountainous regions in the Southeast (S3 Fig), although the number of months suitable has increased in these regions over the past decade (Fig 1). As *Aedes aegypti* is the only vector proven to transmit dengue in Brazil, we also tested the number of months considered suitable for *Aedes aegypti* transmission (between 17.8˚C and 34.5˚C) within the model.

Urbanisation

We obtained the percentage of residents in each municipality living in urban areas from the 2000 and 2010 censuses via DATASUS. In 2010, just under 85% of Brazil's population lived in urban areas, mostly concentrated in the large cities of South and Southeast Brazil. The North region, except for some state capitals, has a larger rural population (S4 Fig). The percentage of residents living in urban areas was converted to the proportion to make interpretation and comparison of model coefficients easier. Data from the 2000 census was used for the years 2001–2009 and data from 2010 was used for the years 2010–2020 to account for changes in urbanisation over the period. Further details on the socioeconomic variables considered in this analysis are given in S1 Text.

Hierarchical levels of influence of cities

As a proxy for human movement, we obtained the hierarchical level of influence of cities from IBGE's REGIC studies, carried out in 2007 and 2018 [28,29]. REGIC aims to recreate the complex urban network of Brazil using information from surveys about the frequency and reasons for the movement of people and goods around the country. Part of this study involved classifying cities based on their hierarchical level of influence within this network (see S1 Text for more details). Cities were classified into five levels:

1. Metropolis: the largest cities in Brazil, with strong connections throughout the entire country. This includes São Paulo, the capital Brasilia, and Rio de Janeiro.
2. Regional capital: large cities which are connected throughout the region in which they are located and to metropoles. This includes state capitals that were not classified as metropoles, such as Rio Branco, Campo Grande and Porto Velho.

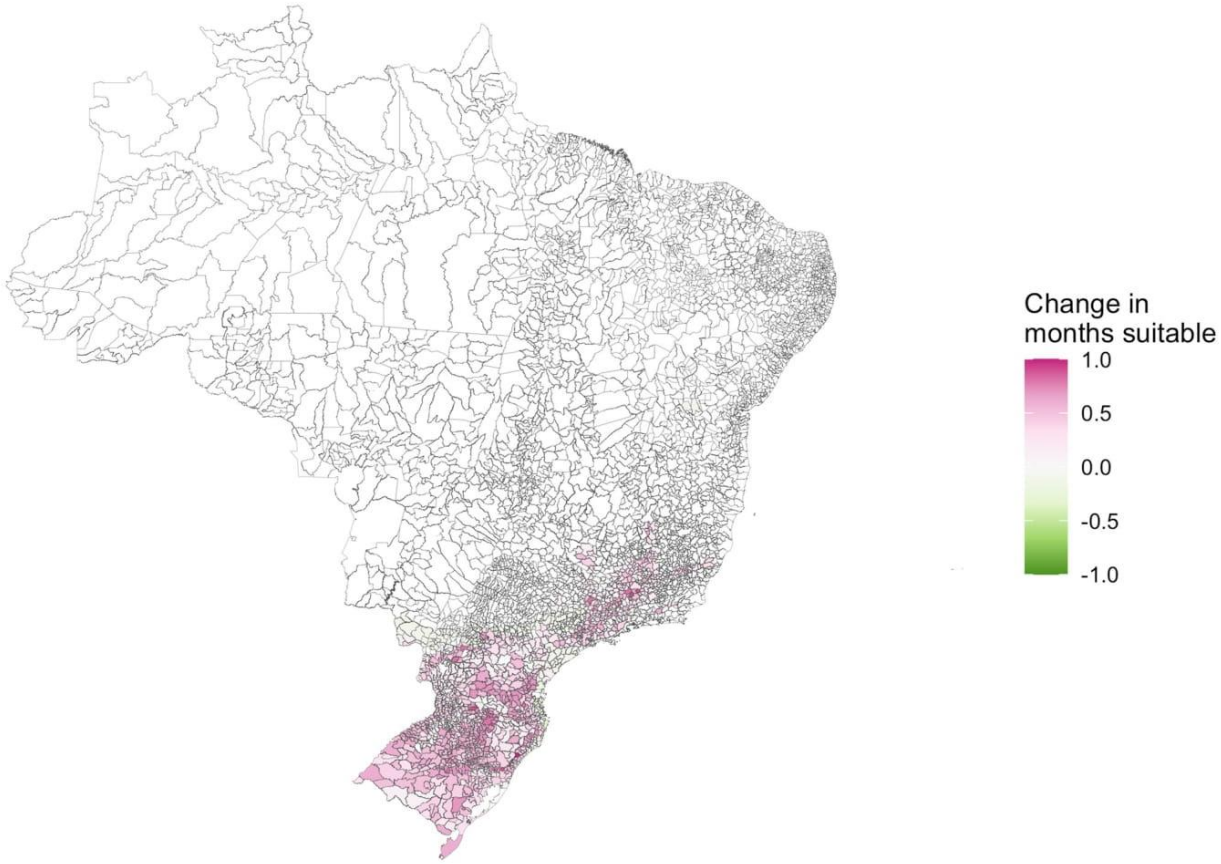


Fig 1. The difference between the average number of months with suitable temperatures for dengue transmission in 2001–2010 and 2011–2020. The number of months with temperatures between 16.2˚C and 34.5˚C has increased on average (shown in pink) in parts of South and Southeast Brazil which were previously considered 'protected' from dengue transmission. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g001

3. Sub-regional capital: cities with a lower level of connectivity, mostly connected locally and to the three largest metropoles.
4. Zone centre: smaller cities with influences restricted to their immediate area, often neighbours.
5. Local centre: the smallest cities in the network which typically only serve residents of the municipality and are not connected elsewhere.

The REGIC study aggregated data to population concentration areas ("Áreas de Concentração de População", ACPs), defined in [36]. Smaller or isolated ACPs consisted of a single municipality, while large urban centres consisted of multiple municipalities. Levels of influence were extracted for each municipality based on the ACP they belonged to, meaning small municipalities neighbouring large cities may have a high level of influence. The distribution of highly connected urban centres is uneven across the country; the South and Southeast regions are particularly well connected, while the North and Northeast contain fewer high-level centres (Fig 2 and S1 Table). To account for any changes in connectivity over the study period, we

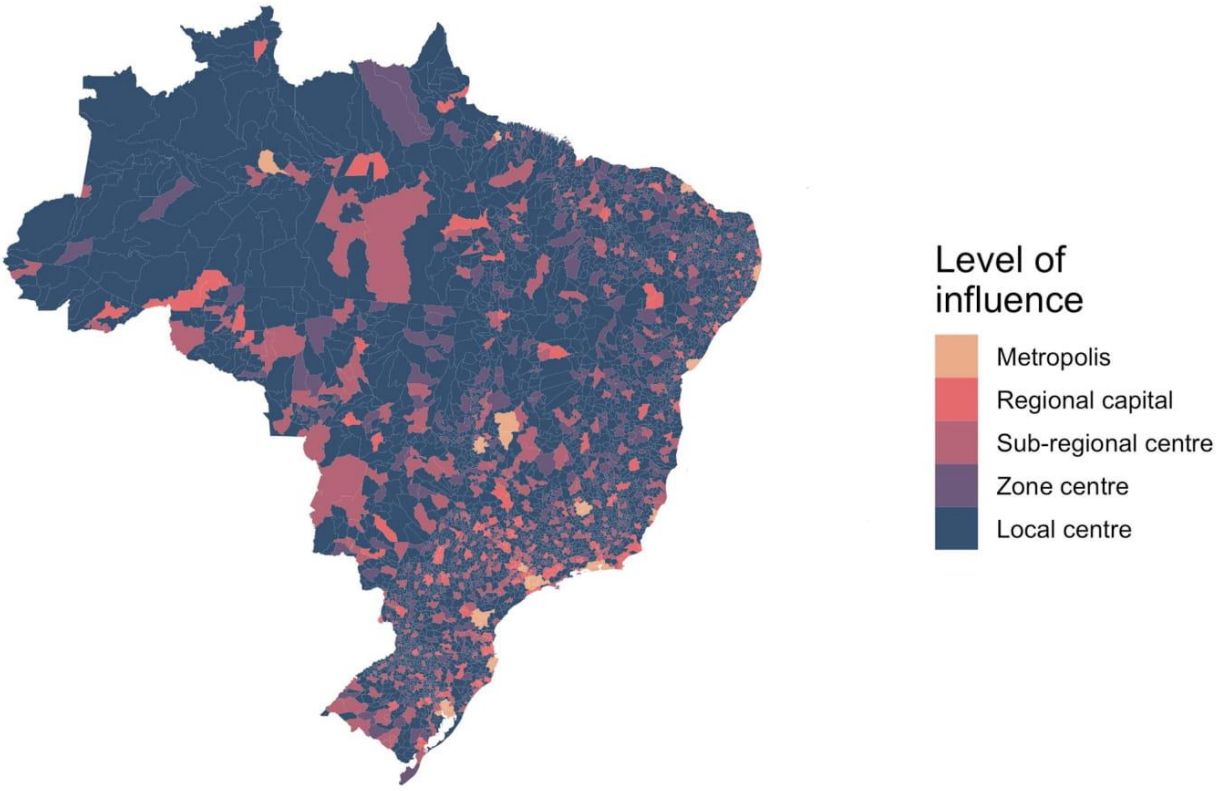


Fig 2. The level of influence of cities within the Brazilian urban network from REGIC 2018. The Amazon region is far less connected to the urban network than the rest of the country. As there is only one metropolis in North Brazil (Manaus), people often travel great distances, far greater than in other regions, to reach cities. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g002

used the levels extracted from the 2007 study for the years 2001–2010, and levels from the 2018 study for the years 2011–2020.

Modelling approach

To understand how the dengue transmission zone has expanded between 2001 and 2020, we aggregated dengue cases by year and created a binary outbreak indicator. We used an outbreak threshold of more than 300 cases per 100,000 residents, defined as 'high risk' by the Brazilian Ministry of Health [37]. We also tested a 'medium risk' indicator, defined as more than 100 cases per 100,000 residents, and a threshold defined as the 75th percentile of the dengue incidence rate between 2001–2020 for each municipality to ensure our analyses were robust to this outbreak definition. The annual dengue incidence rate was calculated using estimates of the annual population for each municipality obtained from the Brazilian Institute of Statistics and Geography (IBGE) via DATASUS (http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/poptbr.def). Further details about the dengue surveillance system in Brazil and outbreak definitions are given in S1 Text. We formulated a binomial spatio-temporal generalised additive model (GAM) using the binary outbreak indicator as the response variable. We included the number of months per year with temperature suitable for *Aedes* mosquitoes to transmit dengue, the level of influence from REGIC, the proportion of residents living in urban areas, and a 'prior outbreak' indicator which took the value 0 until the year of the first outbreak in a municipality and 1 in every year after as covariates. We also considered the number of extremely wet

months as a covariate, but we found this did not improve the model (further details can be found in S1 Text). To account for spatial and temporal patterns in the data, smooth functions of the year and the coordinates of the centroids of municipalities were included in the model (see S1 Text for further details). Inference was performed using an empirical Bayesian approach with estimates calculated using restricted maximum likelihood (REML) as part of the mgcv package in R [38].

Model fit was assessed using a receiver operating characteristic (ROC) curve which plots the true positive rate against the true negative rate at different thresholds to test the predictive ability of the model. The area under the ROC curve was calculated as this gives a measure of predictive ability compared to chance, which would return a value of 0.5. The closer the area under the ROC curve is to 1, the better the model fits the data. The predictive ability of models were also compared using the Brier score [39]. The Brier score is the mean squared difference between the observed and expected outcomes; a lower Brier score represents a better fitting model.

To assess the relative contribution of the covariates, we compared the spatio-temporal structured residual terms between the final model and a baseline model, containing only the spatio-temporal smooth terms. If the covariates explained variation in the data, the smooth functions would shrink towards zero in the final model and the difference between the absolute estimates of these functions would be negative. To assess the contribution of the covariates over the entire period, we took the median difference for each municipality. The contribution of each individual covariate was also assessed by taking the difference between the structured residuals from the baseline model and models with each covariate added in turn.

To understand how the risk of outbreaks have changed between 2001 and 2020, we drew 1000 simulations from the posterior distribution of the response and estimated the probability of an outbreak for each municipality per year. These estimates were aggregated to the first (2001–2010) and second (2011–2020) decades by taking the mean probability for each municipality per decade to observe how the dengue transmission zone had changed after the large-scale outbreak of the 21st century in 2010. The estimated probabilities were then used to determine the current dengue transmission barriers by identifying regions where the average probability of an outbreak lay below 10%, other barrier thresholds were also considered.

Results

There were 13,860,348 cases of dengue notified between January 2001 and December 2020 in Brazil. The dengue incidence rate has increased across all regions of the country (Fig 3) particularly in the Centre-West and Southeast. Outbreaks were more widespread since 2010 with around 80% of all municipalities in the Centre-West now regularly experiencing outbreaks (S5 Fig). Although the South had the highest incidence in 2020, this was still concentrated in a small number of municipalities in Paraná, around the fringe area of the previously identified geographical barrier. The previous barriers to dengue transmission have been eroded over the past decade. This is particularly noticeable in the western Amazon where there are now very few municipalities yet to experience an outbreak. The erosion of the barrier in the South was particularly noticeable in 2020 when it had the highest incidence rate of any region (Fig 3), and many municipalities close to the previous barrier experienced outbreaks for the first time (Fig 4). We observed that once dengue was introduced to municipalities, the virus became established and future outbreaks were likely to occur (Fig 5).

Model results

We found municipalities that were highly urbanised, highly connected, and had temperatures suitable for dengue transmission year-round had a significantly increased odds of an outbreak

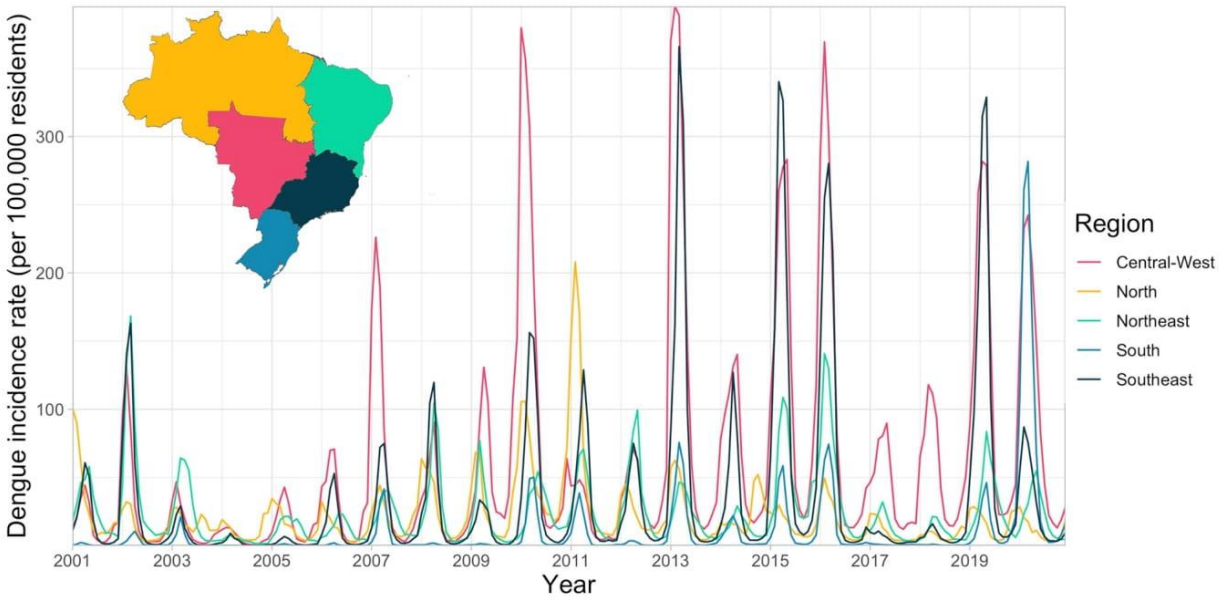


Fig 3. Monthly incidence rate per 100,000 residents in regions of Brazil 2001–2020. Incidence rates have increased in every region of the country between 2001–2020. The first regional outbreak occurred in 2010, outbreaks have occurred more frequently and in more regions since then. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g003

(Table 1 and Fig 6). Municipalities that had previously experienced outbreaks had around double the odds of experiencing another compared to municipalities that were still protected (adjusted odds ratio (aOR): 2.03, 95% credible interval (CI): 1.93, 2.15). This could indicate that the virus becomes established following its introduction, however the increased incidence may be a result of increased surveillance following an outbreak or due to the increased probability of severe cases following the introduction of new serotypes [40]. Municipalities with

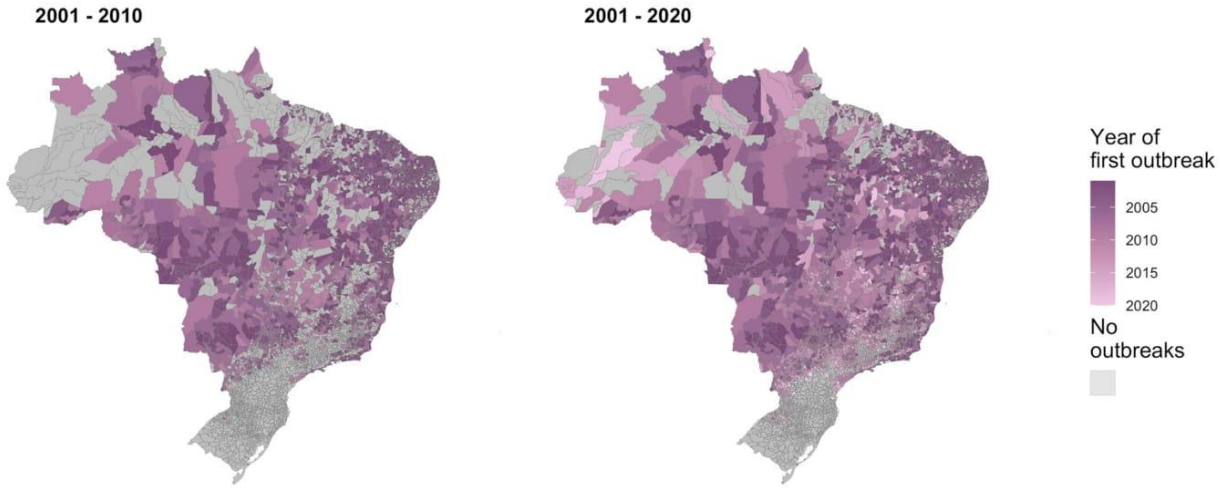


Fig 4. The first year each municipality experienced an outbreak for the first time in the period 2001–2010 and 2001–2020. The year each municipality first recorded over 300 cases per 100,000 residents. Recent data shows the previous barriers to dengue outbreaks in the Amazon and South are being eroded. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g004

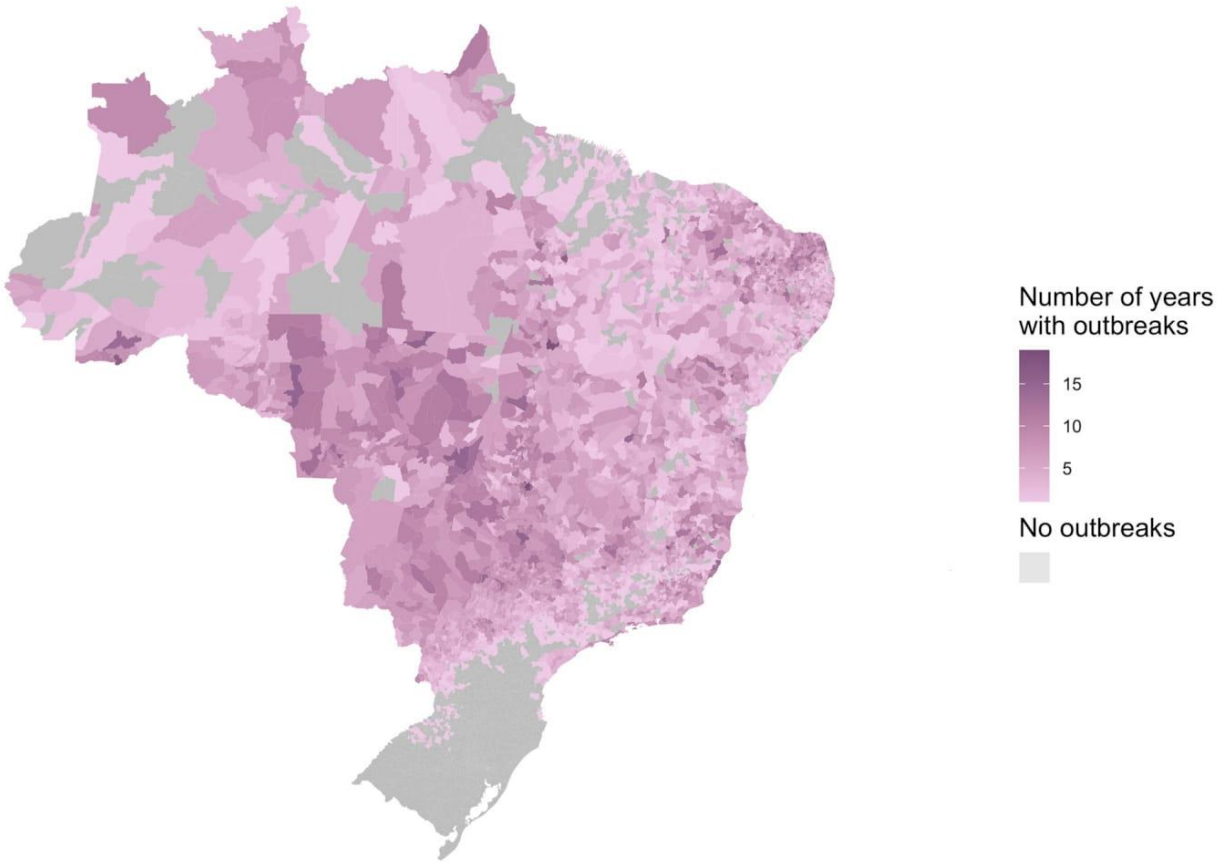


Fig 5. The number of years each municipality experienced an outbreak between 2001 and 2020. Municipalities that experienced outbreaks earlier in the 21st century continued to experience outbreaks throughout the period. This suggests that once dengue is introduced to a region, it becomes established. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g005

Table 1. Posterior mean and 95% credible interval (CI) estimates for linear effect parameters, shown on the adjusted odds ratio (aOR) scale.

Coefficient	aOR (95% CI)
Urbanisation	3.26 (2.85, 3.72)
REGIC level: metropolis	1.39 (1.22, 1.59)
REGIC level: regional capital	1.52 (1.38, 1.66)
REGIC level: sub-regional centre	1.23 (1.14, 1.33)
REGIC level: zone centre	1.23 (1.15, 1.31)
Prior outbreak: yes	2.03 (1.93, 2.15)
Months with suitable temperature	1.42 (1.30, 1.55)

Posterior mean and credible interval estimated taking the 50th, 2.5th and 97.5th quantiles from the simulated posterior distribution. The response variable is a dengue outbreak, defined as over 300 cases per 100,000 residents. Urbanisation is the proportion of residents living in urban areas. REGIC covariates are in comparison to the reference group, local centre. A suitable temperature is defined as between 16.2˚C and 34.5˚C (suitable for both *Aedes aegypti* and *Aedes albopictus*).

https://doi.org/10.1371/journal.pntd.0009773.t001

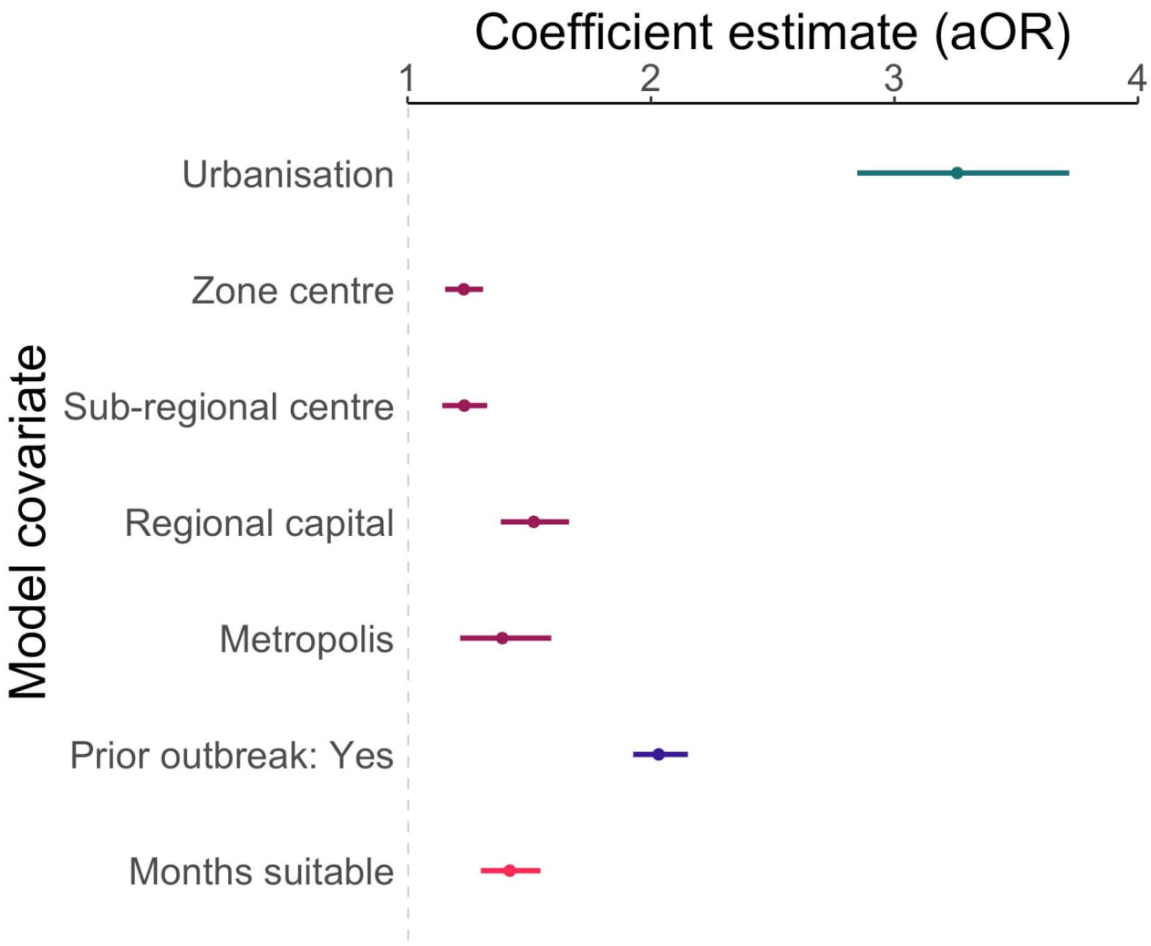


Fig 6. The mean and 95% credible interval of the posterior distribution for each model covariate. Results show that municipalities with a higher proportion of residents living in urban areas, in cities with a higher connectivity than local centres, with a higher number of month per year suitable for dengue transmission, which had previously experienced an outbreak have significantly higher odds of an outbreak.

https://doi.org/10.1371/journal.pntd.0009773.g006

year-round temperature suitability had increased risk of outbreaks, whether we consider suitability for both species of *Aedes* mosquitoes (Table 1) or just *Aedes aegypti* (S2 Table). On average, the odds of an outbreak increased by 42% (aOR: 1.42, 95% CI: 1.30, 1.55) for every additional month of suitable temperature per year.

Although higher levels of connectivity had significantly higher odds of an outbreak than local centres, this difference was highest on average for regional centres (aOR: 1.52, 95% CI: 1.38, 1.66) despite being considered less connected to the urban network than metropoles (aOR: 1.39, 95% CI: 1.22, 1.59). This is potentially due to the structure of the urban network which connects smaller cities to larger centres until they converge to metropoles, meaning that regional capitals are important intermediate urban centres, that influences wide hinterland areas [29]. Alternatively, despite the regional capitals having similar levels of access to basic services as metropoles when aggregated to the municipality level (S6 Fig), metropoles have larger economies and greater access to healthcare than regional capitals [29] which may mean improved infrastructure which is not reflected by census variables on this scale.

The area under the ROC curve for the final model was 0.86 (95% confidence interval: 0.856, 0.861, S7 Fig), indicating that the model fit the data well. We found the conclusions drawn from the models using alternative outbreak definitions remained consistent, however the coefficient values differed (S8 Fig and S2 Table). In particular, the model based on the 75th percentile produced lower coefficient estimates than the fixed threshold models, and the model using a threshold of over 100 cases per 100,000 residents estimated an increased odds following a previous outbreak compared to the primary analysis (S8 Fig and S2 Table). We found that the fixed threshold models outperformed the 75th percentile according to the ROC curve (S7 Fig) and Brier score (S3 Table). The temporal smooth function showed increasing odds of an outbreak over the period not explained by the model covariates (Fig 7). The spatial smooth field showed that the risk around Rio Branco in Acre, the Centre-West region, and in Rio Grande do Norte in Northeast Brazil were higher on average than explained by the model covariates (Fig 7). In contrast, areas in South Brazil, along the northern Brazilian coast, and in parts of the Amazon had lower risk of dengue outbreak occurrence than expected given the covariates.

The structured residuals for the full model were closer to zero on average for the vast majority of the country than the baseline model (92.33% of municipalities, Fig 8), indicating that the covariates are indeed explaining spatio-temporal variation in the data. The inclusion of temperature suitability into the baseline model shrank the structured residuals towards zero for 91.16% of municipalities. This was particularly noticeable in South Brazil (Fig 9), supporting the hypothesis that the dengue transmission barrier here was a result of lower temperatures. The inclusion of the prior outbreak indicator also shrank the structured residuals towards zero across Brazil (in 94.28% of municipalities, Fig 9) showing its relative importance in this model. The relative importance of urbanisation and REGIC levels of influence were less clear; despite the model finding both these variables significantly associated with increased odds of an outbreak, there were fewer municipalities in which the structured residuals had shrank towards (57.5% for urbanisation, Fig 9, and 45.08% for REGIC levels of influence, Fig 9). One potential reason for this is that both variables are only measured once per decade and therefore do not

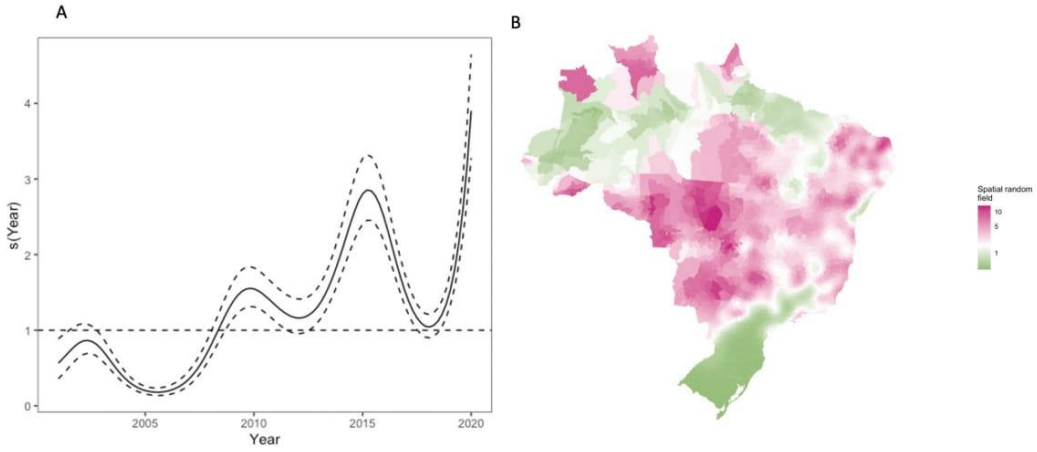


Fig 7. Temporal (a) and spatial (b) smooth functions from the final model transformed to show the change in odds. The odds of an outbreak has increased over the period due to unexplained factors not included in the model. The spatial random field highlights that more information is needed in the model to understand the explosive outbreaks that have taken place in Rio Branco, Acre and the Centre-West region as these hotspots are not fully explained by the model covariates. Pink (green) regions of the map represent areas where the odds of an outbreak was higher (lower) on average than estimated by the covariates. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g007

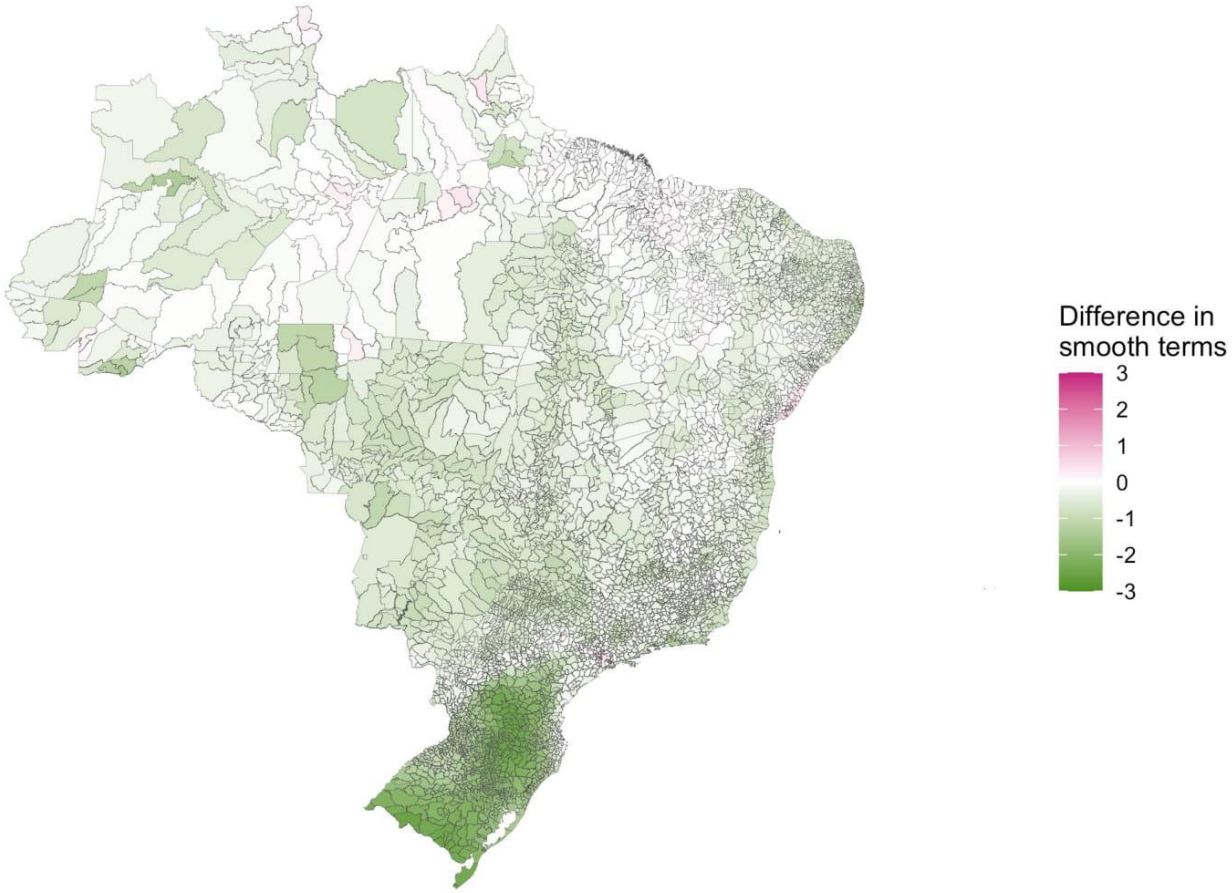


Fig 8. The median difference between absolute values of the smooth function estimates calculated from the full model and from a baseline model. A reduction in the absolute smooth functions (shown in green) indicates that the estimates have shrunk towards zero when the covariates were added to the model and these covariates are explaining some of the variability in the data. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g008

differ annually; there may be changes in municipalities that contribute to dengue transmission but are not captured by these stationary variables. Another potential reason is that these variables are not able to account for within-city variation at this spatial resolution that may contribute to outbreaks of dengue.

The probability of an outbreak increased across most of Brazil since the first decade of the 21st century except for the 2 most southern states and some areas of the Northeast (Fig 10). The largest increases in risk were seen in the Centre-West, which has been the epicentre of the explosive outbreaks taking place since 2010. In the regions previously protected from outbreaks (the western Amazon and the South (Fig 10)), the erosion of the geographic barriers can clearly be seen. Although a southern border still exists, it has shifted south, and the Amazon no longer has a clear boundary.

Current barriers to dengue transmission

To determine the current dengue transmission barriers, we identified regions where the average probability of an outbreak lay below 10% (Fig 11). We chose the threshold 10% as this gave

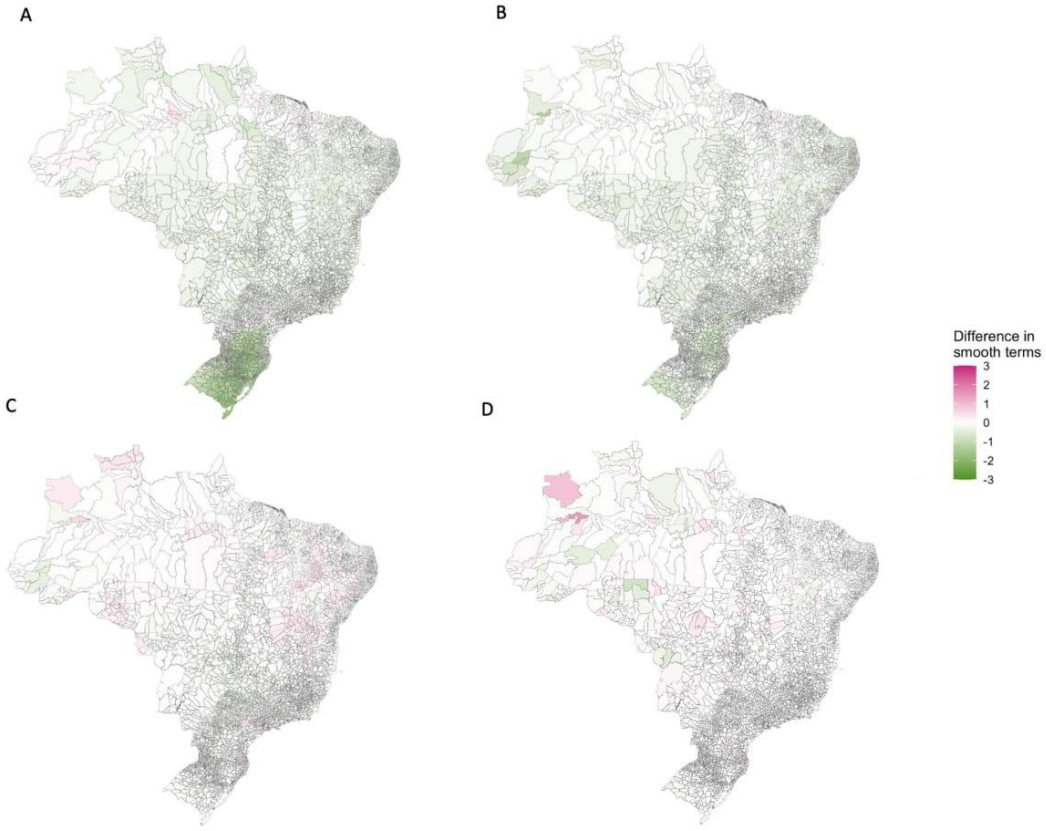


Fig 9. The median difference between absolute values of the smooth function estimates calculated from the baseline model and models with a) the climate suitability covariate added, b) the prior outbreak indicator added, c) the proportion of urbanisation added, and d) the level of connectivity covariate added. A reduction in the absolute estimates of the smooth functions (shown in green here) indicates that the functions have shrunk towards zero and the covariate has explained variation in the data. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g009

barriers comparable to those identified in a previous study [6] (S10 Fig). The number of municipalities considered protected declined from 2689 in 2001–2010 to 1599 in 2011–2020. Between 2011 and 2020 there were no municipalities in the Centre-West region that were considered protected, compared to 92 in 2001–2010. Northeast Brazil was the only region that had more protected municipalities in 2011–2020 than 2001–2020 (366 compared to 315). The southern barrier to dengue transmission now begins in the southern part of Paraná and extends through the west of Rio Grande do Sul and Santa Catarina. Areas of high altitude in Southeast Brazil, mostly found in Minas Gerais, are still considered protected. There are still areas of the Amazon protected from dengue outbreaks, but this barrier is no longer clearly defined. In addition to the previously identified barriers in the South region and Amazon rainforest, we found that there was a protected region along the north coast of Brazil in northern Pará and Maranhão. This barrier was not explained by the covariates in our model indicated by the low values of the spatial smooth function (Fig 7). This area is predominantly warm and humid climate, with higher precipitation during winter ('Am' type in Köppen climate classification) [41]. Although temperature and humidity are relatively stable along seasons in this area, the interaction between these variables and increased precipitation may inhibit the mosquito populations [42].

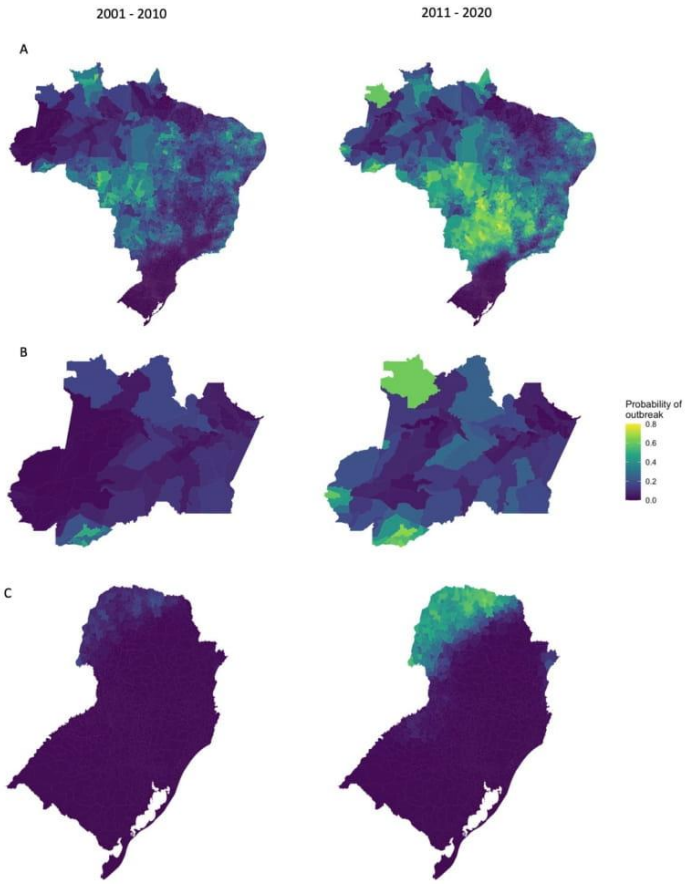


Fig 10. The average probability of an outbreak 2001–2010 and 2011–2020 in a) Brazil, b) Acre and Amazonas, and c) South Brazil. The probability of an outbreak estimated using simulations from the posterior distribution of the response from the final model, averaged over the first and second decade of the time period. The probability of an outbreak has increased across most of Brazil. The Amazonian barrier has almost completely been eroded and the South Brazil border has moved further south. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g010

Discussion

We found that the expansion of the dengue transmission zone is associated with temperature suitability, connectivity within the Brazilian urban network and urbanisation, and that the odds of future outbreaks significantly increase after both the vector and the virus have been introduced. This study builds on previous literature that showed the expansion of dengue across Brazil [6,7,17,26,43] and has updated the geographical barriers to transmission. The most recent epidemiological bulletins have shown that this expansion has continued in 2021 into previously unaffected parts of Acre, Amazonas, and further south into Paraná and Santa Catarina [44], highlighting the importance of monitoring the erosion of these barriers. To our knowledge, this is the first epidemiological modelling study to use the REGIC's levels of influence and show that there is an increased odds of dengue outbreaks in cities that are highly connected within the Brazilian urban network. However, this increase is not linear; regional capitals are considered less connected than metropoles but we found that the increase in odds were higher in these cities. Further investigation is needed to understand whether this is

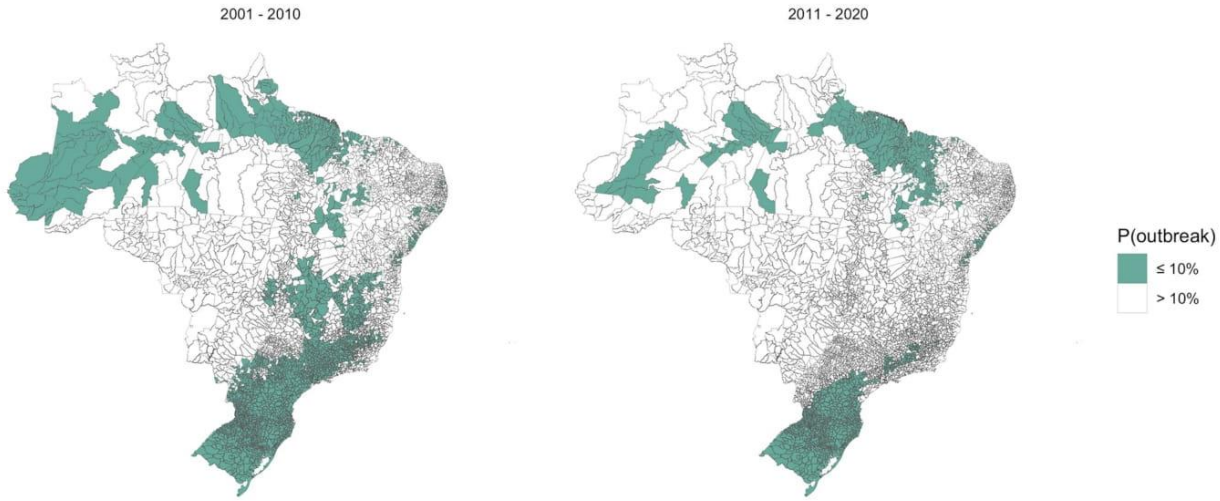


Fig 11. Geographical barriers to dengue transmission in a) 2001–2010 and b) 2011–2020. Maps showing areas where the probability of an outbreak was less than 10% on average in each decade of the 21st century. Between 2011–2020, only the 2 most southern states and the northern coast were fully protected from dengue transmission. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).

https://doi.org/10.1371/journal.pntd.0009773.g011

related to human movement, as people more often travel to regional capitals from smaller cities than metropoles [29], or differences in socioeconomic factors and health-seeking behaviour that we were unable to detect at the municipality level.

Although this study focuses on Brazil, there is evidence that similar patterns are emerging in other parts of South America. In Argentina, previously protected cities in temperate regions are experiencing regular outbreaks, partially related to increasing temperatures but also as a result of human movement importing cases from other parts of the continent [23,24]. Rural parts of the Amazon, which were previously isolated from infected hosts and vectors, are also experiencing outbreaks, thought to be associated with increased connectivity between rural areas and larger cities [13,17]. The introduction of dengue into Acre in the Brazilian Amazon has been linked to increased connectivity across the state following the construction of a highway between the two largest cities, Rio Branco and Cruzeiro do Sul [26]. The impact of this connection can be observed in the data as the outbreak appears to jump from Rio Branco in the south of Acre to Cruzeiro do Sul in the north in 2014 rather than spreading to neighbouring regions which appears to be the case in the South (Fig 5). The introduction of dengue into the Amazon is particularly worrying as it is the ideal environment for the virus to thrive: lower than average access to basic services such piped water and refuse collection, and the ideal climate conditions for large epidemics [17,45].

Although this study extends our understanding of the expansion of the dengue transmission zone in Brazil, there are several limitations. Dengue case data used in this study was taken from Brazil's passive surveillance system, which has been found to differ in accuracy between regions, and between epidemic and non-epidemic periods [31]. To reduce the impact of reporting bias in our model, we used an outbreak indicator rather than case data as a response variable. The outbreak indicator used was chosen as it reflects the Brazilian Ministry of Health's definition [37]. However, the threshold of an outbreak is likely to differ across the country. In regions that historically experienced little or no transmission, even a small number of cases may be viewed as an outbreak. The choice of such a high threshold is likely to produce more conservative estimates of the transmission zone. When our results were compared to a lower outbreak threshold of 100 cases per 100,000 residents, we found the model conclusions

were consistent with the higher threshold. We found that both models using a fixed threshold outperformed the model based on the 75th percentile based on the area under the ROC curve (S7 Fig) and the Brier score. The model failed to pick up some of the temporal trends in the data, which may be a result of using stationary indicators of urbanisation and connectivity measured every 10 years. Information collected at a finer temporal scale may provide more insights into the impact of sudden expansions such as the effect of improved infrastructure in the Amazon [26].

Our model used the level of influence extracted from the REGIC studies [28,29] to account for the level of connectivity between cities within Brazil as a proxy for human movement. However, this indicator may simplify the process and miss important patterns. The hierarchical model assumed by REGIC assumes each small city is linked to a higher-level urban centre, such as the regional capitals and metropoles. It is evident that large and warm cities may propagate epidemic waves and maintain dengue transmission in their hinterland, while temperate metropoles in the South (Porto Alegre, Curitiba and São Paulo) do not play a relevant role in dengue diffusion in their region. Previous studies have found that imported cases driven by human movement are responsible for dengue outbreaks in temperate cities [24,25]. The choice of spatial connectivity assumption and data can lead to very different results and the use of the REGIC levels of influence as a spatial covariate rather than including the direct links may miss some important patterns [27]. Future work will aim to incorporate the complex urban network from the REGIC studies into a statistical framework to account for direct and indirect links between metropoles and regional capitals, and smaller urban centres in their hinterland.

Despite these limitations, we have shown that the expansion of the dengue transmission zone has continued into the 21st century, driven by increased temperature suitability in the South, a network of highly connected cities, and high levels of urbanisation. The introduction of dengue outbreaks into an area more than doubles the odds of future outbreaks, which is particularly concerning given the expansion has continued into 2021. Given the dynamic nature of the growing dengue burden, the barriers identified here will be outdated very quickly. We have highlighted the importance of focusing control strategies in areas at risk of future outbreaks as well as those within the established dengue transmission zone.

Supporting information

S1 Text. Supplementary material. Additional information about the methods and materials used in this study and results of sensitivity analyses.
(DOCX)

S1 Alternative Language Abstract. Translation of the Abstract into Portuguese by Rafael de Castro Catão.
(DOCX)

S1 Fig. The organisation of Brazil into a) 5 geo-political regions, and b) 27 federal units. Abbreviations: AC = Acre, AL = Alagoas, AP = Amapá, AM = Amazonas, BA = Bahia, CE = Ceará, DF = Distrito Federal, ES = Espírito Santo, GO = Goiás, MA = Maranhão, MT = Mato Grosso, MS = Mato Grosso do Sul, MG = Minas Gerais, PA = Pará, PB = Paraíba, PR = Paraná, PR = Pernambuco, PI = Piauí, RJ = Rio de Janeiro, RN = Rio Grande do Norte, RS = Rio Grande do Sul, RO = Rondônia, RR = Roraima, SC = Santa Catarina, SP = São Paulo, SE = Sergipe, TO = Tocantins. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S2 Fig. Average monthly mean temperature (˚C) in each Brazilian state January 2001—December 2020.
(TIF)

S3 Fig. The average number of months suitable for dengue transmission per year a) 2001–2010, and b) 2011–2020. The average number of months with mean temperature between 16.2˚C and 34.5˚C aggregated to the two decades of data. Most of Brazil experiences suitable temperatures year-round apart from areas of South Brazil and areas of high altitude in the Southeast which experience cool winters. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S4 Fig. The percentage of residents living in urban areas of each municipality from the 2000 (a) and 2010 (b) censuses. Levels of urbanisation differ greatly across Brazil, with the majority of Southeast and South Brazil living in urban areas in comparison to the North and Northeast which has a larger rural population. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S5 Fig. The proportion of municipalities in each region of Brazil experiencing an outbreak per year 2001–2020. The proportion of municipalities affected by outbreak has increased since 2010 in every region of the country, although outbreaks in South Brazil are still focused on a small part of the region. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S6 Fig. Raincloud plots exploring the relationship between REGIC level of influence and a) urbanisation, b) access to piped water, and c) refuse collection. Metropoles and regional capitals have higher levels of urbanisation and access to basic services than municipalities that had lower levels of connectivity within the urban network. Local centres were more varied in terms of basic services and urban levels than the other levels and covered a wide range of city types.
(TIF)

S7 Fig. Receiver operating characteristic (ROC) curve for the final model (solid black line), the model using an outbreak threshold of over 100 cases per 100,000 residents (red dashed line), and the model using an outbreak threshold of over the 75th percentile (blue dashed line), compared to chance (black dashed line). The closer to the top-left corner, the better the predictive ability of a model. As the ROC curve lies above the dashed reference line, this model performs better than chance.
(TIF)

S8 Fig. The mean and 95% credible interval of the posterior distribution for each model covariate under different outbreak threshold definitions. Coefficient estimates using the outbreak indicator based on the 75th percentile were noticeably smaller than the fixed threshold alternatives. The fixed threshold models (where outbreaks were defined as a dengue incidence rate of over 100 or 300) produced similar estimates, however the odds of an outbreak in municipalities after a previous outbreak was higher for the DIR = 100 model.
(TIF)

S9 Fig. The probability of an outbreak estimated from the model for each year 2001–2020. The mean probability of an outbreak estimated by taking 1000 simulations from the posterior

distribution of the response and transforming the outcome using a probit function. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S10 Fig. Comparison of different risk thresholds to define current geographical barriers to dengue outbreaks. Municipalities were considered 'protected' if the probability of an outbreak was less than or equal to the threshold a) 0%, b) 5%, c) 10% or d) 15%. The threshold of 10% was chosen as it was the most comparable with previous studies. Maps were produced in R using the geobr package [32,35] (https://ipeagit.github.io/geobr/).
(TIF)

S1 Table. Distribution of municipalities at each level of influence in the urban network, 2007 [28] and 2018 [29]. The number of municipalities classified as metropoles (largest cities in Brazil, connected throughout the entire country), regional capitals (large cities connected regionally and to metropoles), sub-regional capitals (cities connected locally and to the three largest metropoles), zone centres (smaller cities generally connected only to their neighbours), and local centres (smallest cities typically disconnected from the urban network).
(DOCX)

S2 Table. Posterior mean and 95% credible interval (CI) estimates for linear effect parameters, shown on the adjusted odds ratio (aOR) scale, for alternative model formulations. Coefficient estimates for models assuming an outbreak threshold of over 100 cases per 100,000 (medium risk model), an outbreak threshold of over the 75th percentile of incidence rates, and using temperature suitability for *Aedes aegypti* only.
(DOCX)

S3 Table. Model comparison statistics. Area under the receiver operator curve and Brier scores for models assuming an outbreak threshold of over 300 cases per 100,000 residents (high risk model), over 100 cases per 100,000 (medium risk model), over the 75th percentile of incidence rates, and a model including the number of months considered extremely wet.
(DOCX)

Author Contributions

Conceptualization: Sophie A. Lee, Rachel Lowe.

Data curation: Sophie A. Lee.

Formal analysis: Sophie A. Lee.

Investigation: Sophie A. Lee, Rafael de Castro Catão, Christovam Barcellos, Rachel Lowe.

Methodology: Sophie A. Lee, Theodoros Economou, Rachel Lowe.

Software: Sophie A. Lee, Theodoros Economou, Rachel Lowe.

Supervision: Theodoros Economou, Rachel Lowe.

Validation: Sophie A. Lee, Rachel Lowe.

Visualization: Sophie A. Lee.

Writing – original draft: Sophie A. Lee.

Writing – review & editing: Sophie A. Lee, Theodoros Economou, Rafael de Castro Catão, Christovam Barcellos, Rachel Lowe.

References

1. Ten health issues WHO will tackle this year. [cited 10 Jun 2021]. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019
2. Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. Epidemic arboviral diseases: priorities for research and public health. Lancet Infect Dis. 2017; 17: e101–e106. https://doi.org/10.1016/S1473-3099(16)30518-7 PMID: 28011234
3. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. Lancet Infect Dis. 2016; 16: 712–723. https://doi.org/10.1016/S1473-3099(16)00026-8 PMID: 26874619
4. Gubler DJ. Dengue, Urbanization and Globalization: The Unholy Trinity of the 21st Century. Trop Med Health. 2011; 39: 3–11. https://doi.org/10.2149/tmh.2011-S05 PMID: 22500131
5. Fares RCG, Souza KPR, Añez G, Rios M. Epidemiological Scenario of Dengue in Brazil. BioMed Res Int. 2015; 2015: e321873. https://doi.org/10.1155/2015/321873 PMID: 26413514
6. Barcellos C, Lowe R. Expansion of the dengue transmission area in Brazil: the role of climate and cities. Trop Med Int Health. 2014; 19: 159–168. https://doi.org/10.1111/tmi.12227 PMID: 24286460
7. de Azevedo TS, Lorenz C, Chiaravalloti-Neto F. Spatiotemporal evolution of dengue outbreaks in Brazil. Trans R Soc Trop Med Hyg. 2020; 114: 593–602. https://doi.org/10.1093/trstmh/traa030 PMID: 32496548
8. Churakov M, Villabona-Arenas CJ, Kraemer MUG, Salje H, Cauchemez S. Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony. PLoS Negl Trop Dis. 2019; 13. https://doi.org/10.1371/journal.pntd.0007012 PMID: 31009460
9. Ferreira-de-Lima VH, Câmara DCP, Honório NA, Lima-Camara TN. The Asian tiger mosquito in Brazil: Observations on biology and ecological interactions since its first detection in 1986. Acta Trop. 2020; 205: 105386. https://doi.org/10.1016/j.actatropica.2020.105386 PMID: 32027837
10. Ministério da Saúde (BR). Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. Guia de vigilância em saúde: volume único [Internet]. 2019.
11. Rezende HR, Romano CM, Claro IM, Caleiro GS, Sabino EC, Felix AC, et al. First report of Aedes albopictus infected by Dengue and Zika virus in a rural outbreak in Brazil. PLOS ONE. 2020; 15: e0229847. https://doi.org/10.1371/journal.pone.0229847 PMID: 32163449
12. Powell JR, Tabachnick WJ. History of domestication and spread of Aedes aegypti—A Review. Mem Inst Oswaldo Cruz. 2013; 108: 11–17. https://doi.org/10.1590/0074-0276130395 PMID: 24473798
13. Guagliardo SA, Morrison AC, Barboza JL, Requena E, Astete H, Vazquez-Prokopec G, et al. River Boats Contribute to the Regional Spread of the Dengue Vector Aedes aegypti in the Peruvian Amazon. PLoS Negl Trop Dis. 2015; 9: e0003648. https://doi.org/10.1371/journal.pntd.0003648 PMID: 25860352
14. Pérez-Castro R, Castellanos JE, Olano VA, Matiz MI, Jaramillo JF, Vargas SL, et al. Detection of all four dengue serotypes in *Aedes aegypti* female mosquitoes collected in a rural area in Colombia. Mem Inst Oswaldo Cruz. 2016; 111: 233–240. https://doi.org/10.1590/0074-02760150363 PMID: 27074252
15. Ayllón T, Câmara DCP, Morone FC, Gonçalves L da S, de Barros FSM, Brasil P, et al. Dispersion and oviposition of Aedes albopictus in a Brazilian slum: Initial evidence of Asian tiger mosquito domiciliation in urban environments. PLOS ONE. 2018; 13: e0195014. https://doi.org/10.1371/journal.pone.0195014 PMID: 29684029
16. Kraemer MUG, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, et al. Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. Nat Microbiol. 2019; 4: 854–863. https://doi.org/10.1038/s41564-019-0376-y PMID: 30833735
17. Lowe R, Lee S, Lana RM, Codeço CT, Castro MC, Pascual M. Emerging arboviruses in the urbanized Amazon rainforest. BMJ. 2020; 371: m4385. https://doi.org/10.1136/bmj.m4385 PMID: 33187952
18. Lowe R, Lee SA, O'Reilly KM, Brady OJ, Bastos L, Carrasco-Escobar G, et al. Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. Lancet Planet Health. 2021; 5: e209–e219. https://doi.org/10.1016/S2542-5196(20)30292-8 PMID: 33838736
19. Lowe R, Gasparrini A, Meerbeeck CJV, Lippi CA, Mahon R, Trotman AR, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. PLOS Med. 2018; 15: e1002613. https://doi.org/10.1371/journal.pmed.1002613 PMID: 30016319
20. Reinhold JM, Lazzari CR, Lahondère C. Effects of the Environmental Temperature on Aedes aegypti and Aedes albopictus Mosquitoes: A Review. Insects. 2018; 9. https://doi.org/10.3390/insects9040158 PMID: 30404142

21. Mordecai EA, Cohen JM, Evans MV, Gudapati P, Johnson LR, Lippi CA, et al. Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. PLoS Negl Trop Dis. 2017; 11: e0005568. https://doi.org/10.1371/journal.pntd.0005568 PMID: 28448507

22. Mordecai EA, Caldwell JM, Grossman MK, Lippi CA, Johnson LR, Neira M, et al. Thermal biology of mosquito-borne disease. Ecol Lett. 2019; 22: 1690–1708. https://doi.org/10.1111/ele.13335 PMID: 31286630

23. Robert MA, Stewart-Ibarra AM, Estallo EL. Climate change and viral emergence: evidence from Aedes-borne arboviruses. Curr Opin Virol. 2020; 40: 41–47. https://doi.org/10.1016/j.coviro.2020.05.001 PMID: 32569752

24. Robert MA, Tinunin DT, Benitez EM, Ludueña-Almeida FF, Romero M, Stewart-Ibarra AM, et al. Arbovirus emergence in the temperate city of Córdoba, Argentina, 2009–2018. Sci Data. 2019; 6: 276. https://doi.org/10.1038/s41597-019-0295-z PMID: 31754110

25. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors. 2019; 12: 38. https://doi.org/10.1186/s13071-018-3280-z PMID: 30651125

26. Lana RM, Gomes MF da C, de Lima TFM, Honório NA, Codeço CT. The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: A network-based analysis. PLoS Negl Trop Dis. 2017; 11. https://doi.org/10.1371/journal.pntd.0006070 PMID: 29149175

27. Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. J R Soc Interface. 18: 20210096. https://doi.org/10.1098/rsif.2021.0096 PMID: 34034534

28. Estatística IB de G e. Regiões de influência das cidades 2007. IBGE Rio de Janeiro; 2008.

29. Estatística IB de G e. Regiões de influência das cidades 2018. IBGE Rio de Janeiro; 2020.

30. Population, total | Data. [cited 1 Nov 2021]. https://data.worldbank.org/indicator/SP.POP.TOTL?most_recent_value_desc=true

31. Silva MMO, Rodrigues MS, Paploski IAD, Kikuti M, Kasper AM, Cruz JS, et al. Accuracy of Dengue Reporting by National Surveillance System, Brazil. Emerg Infect Dis. 2016; 22: 336–339. https://doi.org/10.3201/eid2202.150495 PMID: 26812472

32. Lee S.A, Economou T., Catão R., Barcellos C., Lowe R. Data and R code to accompany "The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil" (version 1.0.0). 2021. https://github.com/sophie-a-lee/Dengue_expansion

33. Copernicus Climate Change Service. ERA5-Land monthly averaged data from 2001 to present. ECMWF; 2019.

34. Baston, Daniel. exactextractr: Fast Extraction from Raster Datasets using Polygons. 2020. https://CRAN.R-project.org/package=exactextractr

35. Pereira, R.H.M., Gonçalves, C.N. geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. 2019. https://github.com/ipeaGIT/geobr

36. IBGE. Arranjos populacionais e concentrações urbanas no Brasil. IBGE Rio de Janeiro; 2016.

37. Lowe R, Barcellos C, Coelho CAS, Bailey TC, Coelho GE, Graham R, et al. Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. Lancet Infect Dis. 2014; 14: 619–626. https://doi.org/10.1016/S1473-3099(14)70781-9 PMID: 24841859

38. Wood SN. Generalized additive models: an introduction with R. CRC press; 2017.

39. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950; 78: 1–3.

40. Nunes PCG, Daumas RP, Sánchez-Arcila JC, Nogueira RMR, Horta MAP, dos Santos FB. 30 years of fatal dengue cases in Brazil: a review. BMC Public Health. 2019; 19: 329. https://doi.org/10.1186/s12889-019-6641-4 PMID: 30898104

41. Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G. Köppen's climate classification map for Brazil. Meteorol Z. 2013; 711–728. https://doi.org/10.1127/0941-2948/2013/0507

42. Pliego Pliego E, Velázquez-Castro J, Fraguela Collar A. Seasonality on the life cycle of Aedes aegypti mosquito and its statistical relation with dengue outbreaks. Appl Math Model. 2017; 50: 484–496. https://doi.org/10.1016/j.apm.2017.06.003

43. Castro MC, Baeza A, Codeço CT, Cucunubá ZM, Dal'Asta AP, Leo GAD, et al. Development, environmental degradation, and disease spread in the Brazilian Amazon. PLOS Biol. 2019; 17: e3000526. https://doi.org/10.1371/journal.pbio.3000526 PMID: 31730640

44. Secretaria de Vigilância em Saúde. Boletim Epidemiológico—Monitoramento dos casos de arboviroses urbanas causados por vírus transmitidos pelo mosquito Aedes (dengue, chikungunya e zika), semanas

epidemiológicas 1 a 21, 2021. Ministério da Saúde; 2021. https://www.gov.br/saude/pt-br/media/pdf/2021/junho/07/boletim_epidemiologico_svs_21.pdf

45. Huber JH, Childs ML, Caldwell JM, Mordecai EA. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. Althouse B, editor. PLoS Negl Trop Dis. 2018; 12: e0006451. https://doi.org/10.1371/journal.pntd.0006451 PMID: 29746468

3. Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions

Bridging section

In Chapter 2, model results showed that the relationship between the level of influence of cities and dengue expansion was significant but nonlinear, with a greater odds of dengue outbreaks in regional capitals rather than the more influential metropoles. There are many reasons this could be the case, for example differences in healthcare investment and health-seeking behaviours that were not included in the model. Another potential reason is the Brazilian urban network structure that means people from cities of low influence often travel to metropoles via less influential cities such as the regional capitals. This means that although regional capitals are considered less influential, they may have many connections across Brazil due to their proximity in the urban network to metropoles that could not be captured in the model presented in Chapter 2[1]. Chapter 2 considered levels of influence in the hierarchical urban network as a proxy for human movement using a categorical variable rather than accounting for direct links between cities arising from human movement.

This chapter presents a systematic review that aimed to identify spatial models used to investigate the transmission of mosquito-borne diseases to humans, the spatial connectivity assumptions made by these models, and the data used to inform spatial models (Objective 2). I aimed to examine whether any statistical modelling frameworks existed that could incorporate human movement into the spatial structure of the model. Although statistical model frameworks are used in this thesis, the review included mechanistic and machine learning frameworks to gain insight into how other approaches account for complex networks of connectivity.

The chapter was published in the *Journal of the Royal Society Interface* in May 2021[2]. The published version of the paper is included below. Supplementary materials referred to in the paper can be found in Appendix C.

[1.] IBGE. 2020 *Regiões de influência das cidades 2018*.

[2.] Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021 Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**, 20210096. (doi:10.1098/rsif.2021.0096)

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1806358	**Title**	Ms
First Name(s)	Sophie Alice		
Surname/Family Name	Lee		
Thesis Title	Spatial Modelling of Emerging Infectious Diseases: Quantifying the Role of Climate, Cities and Connectivity on Dengue Expansion in Brazil		
Primary Supervisor	Dr Rachel Lowe		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Journal of the Royal Society Interface		
When was the work published?	26 May 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My contribution: defining the search strategy, carrying out literature search and screening of articles, data extraction, software, visualisation, analysis, and writing.

SECTION E

Student Signature	Sophie Lee
Date	08/09/2022

Supervisor Signature	Rachel Lowe
Date	08/09/2022

INTERFACE

royalsocietypublishing.org/journal/rsif

Review

Cite this article: Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021 Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**: 20210096. https://doi.org/10.1098/rsif.2021.0096

Received: 2 February 2021
Accepted: 26 April 2021

Subject Category:
Reviews

Subject Areas:
biomathematics, computational biology

Keywords:
spatial analysis, infectious disease dynamics, epidemiology, vector-borne disease, machine learning

Author for correspondence:
Sophie A. Lee
e-mail: sophie.a.lee10@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5427697.

THE ROYAL SOCIETY PUBLISHING

Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions

Sophie A. Lee[1,2,3], Christopher I. Jarvis[1,3], W. John Edmunds[1,3], Theodoros Economou[4] and Rachel Lowe[1,2,3]

[1]Centre for Mathematical Modelling of Infectious Diseases, [2]Centre on Climate Change and Planetary Health, and [3]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK
[4]Department of Mathematics, University of Exeter, Exeter, UK

SAL, 0000-0002-2049-9756; TE, 0000-0001-8697-1518; RL, 0000-0003-3939-7343

Spatial connectivity plays an important role in mosquito-borne disease transmission. Connectivity can arise for many reasons, including shared environments, vector ecology and human movement. This systematic review synthesizes the spatial methods used to model mosquito-borne diseases, their spatial connectivity assumptions and the data used to inform spatial model components. We identified 248 papers eligible for inclusion. Most used statistical models (84.2%), although mechanistic are increasingly used. We identified 17 spatial models which used one of four methods (spatial covariates, local regression, random effects/fields and movement matrices). Over 80% of studies assumed that connectivity was distance-based despite this approach ignoring distant connections and potentially oversimplifying the process of transmission. Studies were more likely to assume connectivity was driven by human movement if the disease was transmitted by an *Aedes* mosquito. Connectivity arising from human movement was more commonly assumed in studies using a mechanistic model, likely influenced by a lack of statistical models able to account for these connections. Although models have been increasing in complexity, it is important to select the most appropriate, parsimonious model available based on the research question, disease transmission process, the spatial scale and availability of data, and the way spatial connectivity is assumed to occur.

1. Introduction

The World Health Organization (WHO) estimates that over 80% of the world's population is now at risk of one or more vector-borne disease, accounting for 17% of the global burden of communicable diseases [1]. The past 50 years has seen an unprecedented emergence of mosquito-borne diseases, in particular dengue fever, chikungunya and Zika, linked to urbanization, globalization, international mobility and climate change [2,3]. Increased connectivity between geographical regions due to international air travel has led to these diseases invading previously naive populations where competent vectors exist, as seen in the introduction of chikungunya to Latin America and the Caribbean [4], and sporadic outbreaks of dengue fever in parts of Southern Europe [5]. Conversely, the global incidence of malaria has decreased over the past 20 years, with an increasing number of countries working towards eradication, although this trend has slowed in the past 5 years [6]. Spatial connectivity arising from human movement may pose a risk of re-introducing a pathogen into indigenous populations. Failure to account for this in modelling studies may negatively impact control and eradication campaigns [7].

Table 1. Search terms used to search Medline, Embase, Global Health and Web of Science related to mosquito-borne diseases, modelling and spatial connectivity.

mosquito-borne diseases	modelling	connectivity
mosquito[a,b] disease[b]	(math[b] OR statistic[b])[a] model[b]	(spati[b] OR cluster)[a] analysis
chikungunya	(gravity OR radiation)[a] model[b]	autocorrel[b] OR neighb[b] OR hierarch[b] OR adjacen[b] OR proximity OR network OR commut[b] OR connect[b]
dengue	(spati[b] OR Bayes[b])[a] model[b]	random[a] effect[b]
'Japanese encephalitis'	(ecolog[b] OR environment[b])[a] model[b]	(BYM OR 'Besag[b] Yorke and Mollie')[a] model[b]
malaria	(dynamic OR stochastic OR determinist[b] OR mechan[b] OR compartment[b])[a] model[b]	'conditional autoregress[b]' OR CAR
(Rift Valley)[a] (fever OR virus)	(regression OR general[b])[a] model[b]	human[a] (mobility OR movement OR travel)
sindbis	(SIR OR SEIR)[a] model[b]	spat[ba] depend[b]
('West Nile')[a] (fever OR disease[b] or virus)	patch[a] model[b]	metapopulation
'yellow fever'	(empirical OR correl[b] OR movement)[a] model[b]	spati[ba] (structure OR matrix)
Zika		
Aedes		
Anopheles		
Culex		

[a]Proximity searching was used, search terms had to be within three words of each other. ADJ3 was used for Embase, Medline and Global Health, NEAR/3 was used for Web of Science.
[b]Denotes truncation. MeSH terms related to terms above were also searched.

The inclusion of space within infectious disease epidemiology is not a new phenomenon; however, the introduction of Geographical Information Systems, improvements in computational power, and availability of spatial data have made spatial modelling more accessible [8]. Despite this, Reiner *et al.* [9] found that spatial modelling methods were underrepresented in their review of mathematical models for mosquito-borne diseases, and spatial connectivity was not explored in the majority of studies. Tobler's first law of geography states that 'everything is related to everything else, but near things are more related than distant things' [10]. However, when studying mosquito-borne diseases, long-distance movement of hosts and vectors may create connections between distant regions. Connectivity between geographical areas and observations can arise for a number of reasons, for example, shared characteristics such as human behaviour, vector-control programmes, levels of immunity within communities and human and vector movement. Although these issues are common among diseases, their impact and the assumption about how connectivity arises may differ due to mosquito behaviours and different geographical settings.

Spatial connectivity is an important driver of mosquito-borne disease, but to our knowledge, there are no systematic reviews of spatial modelling techniques that include statistical, machine learning and mechanistic frameworks. These three approaches are used to address different objectives and require different types of information. Mechanistic models are less dependent on extensive training datasets than statistical or machine learning approaches and can be parameterized using previous experiments. However, this requires an in-depth understanding of the underlying disease process and incorrect parameterization could lead to invalid inference [11]. Mechanistic models are useful for studying (re-)emerging diseases, where few data exist, and comparing potential control strategies [12]. By contrast, machine learning models are able to make predictions about complex biological processes, without prior knowledge of the underlying process, using algorithms that learn from rich, complex data [13]. Statistical models are able to explore relationships between variables, test hypotheses about the underlying transmission process and make predictions about an outcome of interest where adequate data are available.

This systematic review aims to identify spatial models used to investigate the transmission of mosquito-borne disease to humans, the assumptions made about how spatial connectivity arises and the data used to inform the spatial models. We provide detailed explanations of these methods, their assumptions, how they were used, and discuss their advantages and disadvantages.

2. Methods

2.1. Search strategy

The PRISMA guidelines for systematic reviews and meta-analyses were followed for this review [14]. Five online bibliographic databases were searched: Ovid/Medline, Web of Science, Embase, Global Health and Scopus. The final search was completed on 14 December 2020. The search strategy included relevant keywords and Medical Subject Headings (MeSH) related to mosquito-borne diseases and the mosquito species that transmit them, mathematical models used to model infectious diseases and spatial connectivity. Full details of the search strategy are provided in table 1. Mosquito-borne diseases listed on WHO and European

Centre for Disease Prevention and Control websites were considered: dengue fever, Zika, chikungunya, malaria, yellow fever, West Nile fever, Rift Valley fever, sindbis fever and Japanese encephalitis [15,16].

Results from database searches were combined and stored using EndNote referencing software; duplicates were removed manually. The titles and abstracts were screened and irrelevant articles excluded. Two reviewers screened full texts independently and disagreements were resolved by consensus. After relevant papers were identified, their references were screened to identify other relevant studies.

2.2. Inclusion and exclusion criteria

The inclusion criteria are as follows: articles must be peer-reviewed, published in English and contain a spatial model that investigates the transmission of mosquito-borne disease to humans. Spatial models are defined as those that explicitly account for connections between geographical areas or observations. There were no geographical or publishing date restrictions applied. Articles were excluded if they only modelled transmission to vectors or non-human hosts as these were outside the scope of this review and may require different assumptions of connectivity. Theoretical modelling studies that were fitted using simulated data were excluded unless they were validated using real data. Conference and workshop proceedings were excluded, as were review articles.

2.3. Data analysis

The following variables were extracted from eligible papers: title, first author, year of publication, disease studied, country/region studied, the spatial scale of the data, spatial model used, the spatial method used to account for connectivity, connectivity assumptions and the data used to inform the spatial element of the model.

Spatial models were classified as either statistical, machine learning or mechanistic. Statistical models assume that the data are a realization of a pre-specified probability distribution. These probability distributions are defined by a set of parameters which are estimated from the data using estimation, inference and sampling techniques, such as maximum likelihood, Markov chain Monte-Carlo and bootstrapping. The association between an outcome of interest and a set of covariates is determined by how these affect the probability distribution of the outcome. Statistical models were also classified as either fixed effect, where all parameters are treated as fixed, non-random values or mixed effect, which contain both fixed parameters and random parameters that account for unobserved heterogeneity or clustering within the data. Machine learning methods use algorithms to learn patterns from observed data without the need to specify a data model prior to analysis. This makes them a useful alternative to mechanistic or statistical models where underlying biological processes are not known [13]. Mechanistic models, sometimes referred to as mathematical models, aim to replicate the process of disease transmission through a population across time based on a simplified mathematical formulation of the underlying disease mechanisms. These models often simulate the movement of individuals through infectious stages, or compartments, known as compartmental models [11]. Mechanistic models can be parameterized using a combination of data, when available, and results from previous studies. This makes them particularly useful for studying novel pathogens where there are few empirical data or when comparing potential control measures [12]. Spatial assumptions were compared between diseases and mosquito species.

Analysis of the data and visualizations were carried out using R [17]. Data extracted from the studies included in this systematic review and code used to create figures and tables are available from https://github.com/sophie-a-lee/mbd_connectivity_review and archived in a permanent repository [18]. This study is registered with PROSPERO, CRD42019135872.

3. Results

3.1. General characteristics

We identified 248 studies published between 1999 and 2020 that were eligible for inclusion (figure 1). These studies used data from 164 countries across six continents (electronic supplementary material, figure S1). Almost half ($n = 118$, 47.6%) of the studies modelled malaria transmission, 99 (39.9%) modelled dengue fever (including two modelling dengue haemorrhagic fever, two which also modelled Zika, one that also modelled chikungunya and one that modelled dengue, chikungunya and Zika), 11 (4.4%) modelled just Zika and five (2%) just chikungunya, one modelled both. Seven (2.8%) modelled West Nile fever, five (2%) Japanese encephalitis, 1 (0.4%) Rift Valley fever and one (0.4%) yellow fever. No spatial modelling studies were identified for sindbis fever. The number of spatial modelling studies published has increased over time, with an average of one study published per year in 1999–2005, 5.8 per year 2006–2010, 14.2 per year 2011–2015 and 28.2 per year 2016–2020. The diversity of diseases studied using spatial modelling has also increased; until 2005, only malaria studies were identified whereas there have been six different diseases studied using these methods published in 2020 (electronic supplementary material, figure S2). Most studies ($n = 218$, 87.9%) used aggregated data to fit models, most often aggregated to administrative district- or country-level ($n = 169$, 68.1%) or clusters based on surveys or shared characteristics ($n = 25$, 10.1%). The remaining papers either separated their study area into a grid and aggregated data to these patches ($n = 24$, 9.7%) or fit data to individuals ($n = 30$, 12.1%). A full summary of data extracted from studies by disease is given in electronic supplementary material, table S1.

3.2. Spatial modelling methods

Most ($n = 209$, 84.2%) studies used a statistical modelling framework, in particular mixed effect models ($n = 155$, 62.5%). The first mechanistic model included in this review was published in 2012; mechanistic models are becoming more common with over half of those studies published since 2018 (figure 2). Newly emerging diseases (Zika and chikungunya) were more often modelled using mechanistic models rather than statistical, which were more commonly used for established diseases (e.g. malaria and dengue) (electronic supplementary material, table S1). There were two studies published in 2020 that used a combination of methods: one compared a mechanistic and machine learning approach to predicting dengue transmission [19], another used both a machine learning and statistical approaches to explore the relationship between risk factors and dengue outbreaks [20].

We identified 17 distinct models that incorporated spatial connectivity into their framework: nine statistical, four machine learning and four mechanistic models. Full descriptions of the 17 models identified in this review, including model structure, the method and data used to account for spatial connectivity, and a discussion about the advantages and disadvantages of each model are given in electronic supplementary material, technical appendix. Some models were

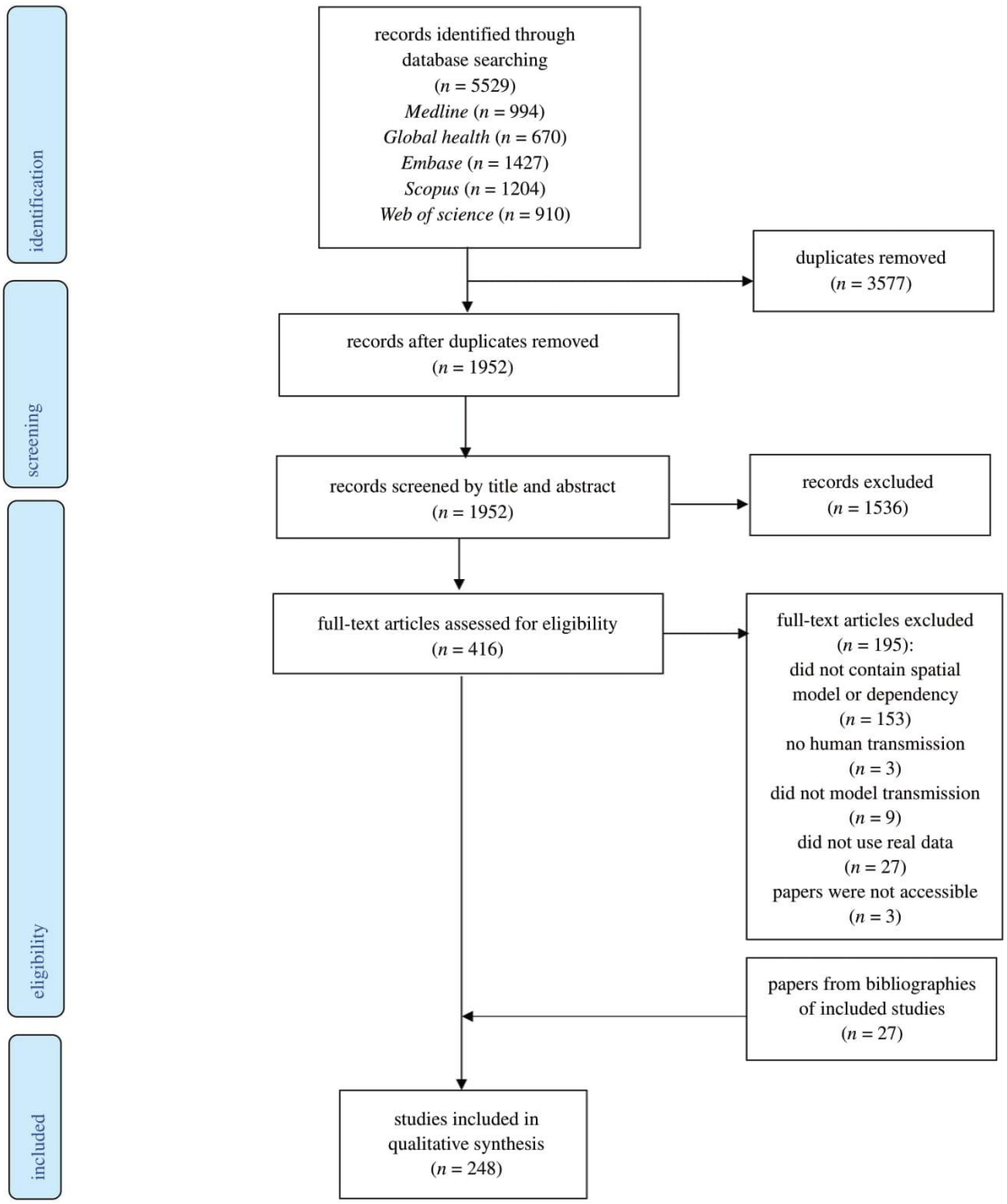


Figure 1. PRISMA flow diagram of the search and exclusion process.

specifically designed for spatial analysis, whereas others have been adapted or extended to incorporate this connectivity. This section gives an overview of the methods used to account for spatial connectivity for each type of model. Details and best practices are summarized in table 2.

3.2.1. Statistical models

All statistical models identified within this review were extensions of generalized linear or additive models (GLM/GAM). These models assume that all observations are independent after adjusting for the covariates, which is not always appropriate when considering spatial data. Although there were nine distinct statistical models, all of them used one of three methods to account for spatial connectivity: inclusion of spatial covariates as fixed effects, localized regression models or the inclusion of a spatially structured random effect or random field.

3.2.1.1. Spatial covariates

Of the 209 papers using statistical models, 25 (12%) included spatial covariates to account for spatial connectivity in the data. Spatial covariates are entered into the model in the same way as nonspatial covariates, but aim to account for connectivity within the model. Spatial covariates included the observed

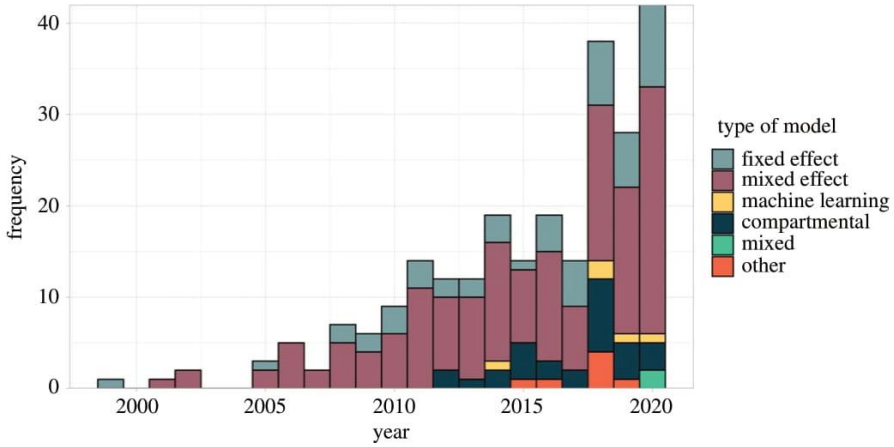


Figure 2. Number of spatial modelling studies published per year by model type. Statistical models were classified as a fixed effect if parameters were treated as fixed, non-random values or mixed effect if they also included random parameters to account for unobserved heterogeneity or clustering (also known as hierarchical or multilevel models). Machine learning models used algorithms to learn patterns from the data. Compartmental models were mechanistic models that simulated the movement of hosts and/or vectors through disease compartments. Models classified as 'other' did not fall into any of these categories, this included mechanistic models that did not explicitly model movement through compartments, or bespoke statistical models.

incidence in connected regions [21–30], the number of people moving between regions [20,31–35], the distance between regions [31,35–37], coordinates of the centroid of a region [38–40], the number of time spent commuting between regions [41] and spatial eigenvectors created using spatial filtering [42–44]. Spatial filtering creates spatial covariates by decomposing Moran's I (a measure of spatial correlation) into an eigenvector per region/observation [45]. Two studies applied a smoothing function to the spatial covariates within a GAM, allowing for a nonlinear relationship between the outcome and measure of connectivity [24,37]. Another study included spatial kernels, exponentially decaying correlation functions of the distance between cases' home and work addresses, estimated from public transport journeys, as spatial covariates when estimating the probability of cases being linked [46].

Spatial covariates are compatible with all statistical models identified in this review. If adequate data are available, this is a simple and efficient way to include connectivity information into a statistical model. Using information from connected regions also allows the model to 'borrow strength' from other parts of the data to increase the precision of estimates. Spatial covariates were the only method that allowed human movement data to be included in statistical models identified in this review; all other methods relied on a function of distance. However, the inclusion of a large number of spatial covariates risks overfitting the model to the data, meaning the model reflects the sample data too closely and is unable to make prediction or inferences about the wider population, or introducing multicollinearity. Most spatial covariates require 'connectivity' to be defined prior to model fitting, introducing a subjective element into the model and potentially oversimplifying the spatial structure. For example, models that included incidence from connected regions defined these as regions that share borders; this ignores potential dependency between distant regions which could still invalidate the independence assumption. The inclusion of spatial covariates as fixed effects assumes that the relationship between them and the outcome is stationary (the same across the whole spatial area) and linear which may not be appropriate across large areas.

3.2.1.2. Local regression models

Twenty papers used a geographically weighted regression (GWR) model [47–65] which fits local regression models to each observation or region rather than a single global model [66]. Each local model has different coefficients, estimated using information from connected observations that are weighted by a function of distance, such as the one shown in figure 3*c*. As with spatial covariates, GWR is a fairly simple and efficient method to account for connectivity and a useful exploratory tool to investigate how relationships differ across space. Estimating a different coefficient for each model overcomes the issue of stationarity which is present when using spatial covariates. GWR is not suitable for making inferences or predictions about the study area as a whole.

3.2.1.3. Spatially structured random effects and random fields

The final, and most common, method used to account for spatial connectivity in statistical methods was the inclusion of a spatially structured random effect or random field. Fixed effect statistical models assume that there is a true parameter value and that the only variation within the data, after accounting for covariates, is sampling error. Random effects and random fields explicitly allow additional spatial variation and/or correlation in the data to be incorporated directly into the model structure. The structure of the random effects or random fields must be specified prior to model fitting and should be informed by the spatial connectivity assumption. Most models identified in this review used a Gaussian process which assumes the spatial process at fixed locations follows a multivariate normal distribution, with a mean of 0 and a covariance structure based on distance or, when dealing with areal data, adjacency.

We identified 150 studies (150/209, 71.8%) that used a spatially structured random effect within their statistical model, 95 assumed a Markov random field structure based on adjacency [29,40,42,64,67–156] and 57 used a distance-based structure [141,157–212] (one used both [141]). A commonly used Markov random field is known as the conditional

Table 2. The advantages, disadvantages and uses of spatial modelling methods.

model type	spatial method	description	advantages	disadvantages	application
statistical or machine learning	spatial covariate	inclusion of a covariate that aims to describe spatial connectivity within a regression model. For example, incidence of surrounding regions, distance between observations, or number of people moving between regions. The covariate is treated as a fixed effect and included into a model as any other covariate	compatible with all statistical or machine learning methods relatively quick and simple to fit allows human and vector movement to be included interpretation of coefficients is often simpler than other methods allows models to 'borrow strength' from connected regions, improving precision	models assume that the relationship between the outcome and spatial covariates is stationary and isotropic inclusion of a large number of spatial covariates increases risk of overfitting and multicollinearity within a model user must specify which regions/observations are connected prior to model fitting which does not allow other connections to be explored	exploratory tool for statistical or machine learning studies carried out on a small scale where few spatial connections are expected. Statistical or machine learning modelling studies where spatial connectivity is assumed to arise from human movement
statistical	local regression models	local regression models are fitted to each region using data from nearby regions, weighted by distance. Also known as GWR. Coefficients are calculated separately for each regression model	relatively simple to carry out and interpret useful exploratory tool to understand how the relationship between covariates and the outcome differ across space does not assume these relationships are stationary	does not provide a global model to make interpretations about a region as a whole only allows distance-based spatial connectivity to be included	exploratory tool to generate hypotheses about how relationships differ across space. Cannot be used to make inferences about regions as a whole. Only appropriate when studying areal data
statistical	random effects and fields	random effects or fields with a spatially structured covariance function are included in a regression model to account for additional correlation or heterogeneity arising from spatial connectivity. Users must choose an appropriate spatial structure before fitting the model, usually assuming that regions are connected if and only if they are adjacent (areal data) or that connections decay exponentially as the distance between them increases (individual-level data)	relatively easy to obtain connectivity data (if using structure based on adjacency or distance) does not assume stationarity in the model allows connections between a large number of observations without issues of overfitting associated with other statistical methods increasing number of methods and software developed to make model-fitting process simpler	more complex to fit and interpret models than other statistical models random effects require an appropriate spatial structure defined before model is fitted structures identified in this review only allow models to account for connectivity between neighbours or close regions, other connectivity has not been explored	statistical models where spatial connectivity is expected to exist between nearby regions. Can be carried out in small- or large-scale studies. Recommended for established diseases rather than a newly emerging setting as requires large amounts of data for precise estimates

(*Continued*.)

Table 2. (*Continued.*)

model type	spatial method	description	advantages	disadvantages	application
machine learning	movement matrix	movement matrices reflecting the movement of humans around a network used to weight connections between hidden layers of a neural network	allows complex, dynamic connectivity structures to be explored allows human movement to be included in a machine learning framework	requires human movement data (or a representative proxy) to create which can be difficult to obtain inclusion of the matrix in the hidden layer of neural networks means the impact of this movement is difficult to observe computationally intensive	inclusion in a neural network where human mobility is known to drive transmission. Studies that require accurate predictions based on a large amount of data but quantifying this process is not the focus
mechanistic	spatial parameter	spatial parameters are included in mechanistic model equations, either to take account for a spatial process or to update populations within each disease compartment of the model. Examples include diffusion parameters allowing hosts and vectors to move across a region or mosquito abundance that borrows information from connected regions	models can be fitted with few data and used to make causal inferences parameters can borrow information from other regions about processes to take account of shared characteristics less computationally intensive to fit than other mechanistic approaches can be used within any mechanistic model	requires knowledge and information regarding the underlying process of transmission parameters assume that the impact of spatial coefficients on transmission is stationary within a compartmental model making them inappropriate on a large scale	models aiming to make causal inferences about the underlying process of transmission. Able to fit models where few data are available making it useful for newly emerging diseases or areas with low transmission. More appropriate in small-scale studies where stationarity can be assumed
mechanistic	movement matrix	movement matrices that reflect the movement of hosts and/or vectors around a network are included within a mechanistic model. These allow interaction between hosts and vectors in different locations and update the population at each node of the network	allows complex, dynamic connectivity structures to be explored results can be extrapolated beyond the data used to fit them and causal inferences can be made provides a more 'realistic' reflection of human and vector behaviour models can be fit with relatively few data	adequate movement data are difficult to obtain the complex nature of these models means computation can be difficult and time consuming inferences can only be made about the setting the model is parameterized to reflect requires the population being studied to be split into nodes in networks	models taking account of human and/or vector movement or other complex connectivity structures. Able to fit models where few data exist as well as large amounts, useful for newly emerging diseases. Able to study the process of transmission or causal structures. Works well with agent-based or metapopulation mechanistic models where the population is described using a network

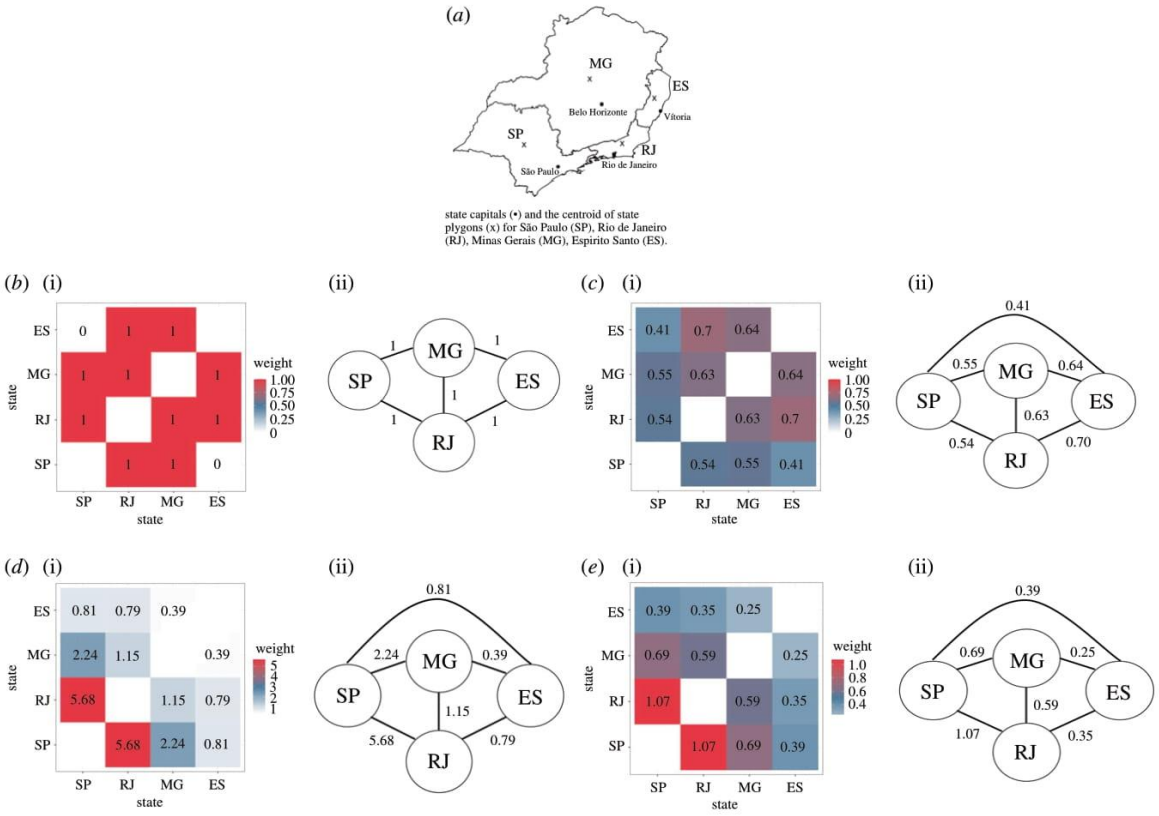


Figure 3. Comparison of spatial connectivity using different data sources and assumptions. The level of connectivity between regions represented in models can differ substantially depending on the assumptions made about how connectivity arises, and the data used to weight connections. The heat plots and connectivity matrices show the strength of connectivity between states in Southeast Brazil (*a*), represented by nodes in the matrices, using assumptions and methods identified in this review. Numbers within the heat plot and along edges of the connectivity matrix represent the weight of connections. These techniques were used to weight observations in GWR models, to structure random effects and random fields, or to weight movement matrices in neural networks, metapopulation models, and agent-based models. (*b*) Neighbourhood based: assumes states are connected if and only if they share a border. Application: to structure random effects in a CAR model. (*c*) Distance-based: assumes connectivity between states decays exponentially as distance between centroids (denoted x on the map) increases, where *weight = exp(dij /1000)* and *dij* is the distance between states *i* and *j*. Application: used to weight observations from neighbouring regions in a GWR model. (*d*) Human movement data: assumes connectivity between states arises due to human movement. In this case, based on the number of air travel passengers moving between capital cities of each state. Application: to weight hidden layers within a neural network. (*e*) Movement model: assumes connectivity between states arises due to human movement, estimated using a movement model (in this case, a gravity model). Application: used to weight movement between nodes in a metapopulation model.

autoregressive (CAR) model, which assumes that regions are connected if and only if they are neighbours [213], i.e. regions that share a border or, in one case, regions within a fixed distance [140]. The weighting matrix used to formulate this Markov random field is shown in figure 3*b*. Distance-based approaches identified in this review used the Matérn correlation function [214] to define the random effect covariance. This assumes that connectivity between points decays exponentially as the distance between them increases, as shown in figure 3*c*. There were 15 studies that included a spatially structured random field, a bi-dimensional smooth function in space over the coordinates of observations or the centroid of a region [40,215–227]. Bi-dimensional smooth functions are a type of Gaussian process, with a covariance structure defined by the distance between observations, for which connectivity is expected to decrease exponentially as distance increases [228] (figure 3*c*). One spatial model included a random field, estimated using a Markov random field [229], similarly to the CAR models above, assuming connectivity exists between neighbouring regions [228]. One study used an alternative way of accounting for residual spatial autocorrelation by fitting a separate regression model to the error terms of a non-spatial model. The observed outcomes from previous time points were included in the residual model as covariates. This model was fitted using an iterative process and was referred to as a vectorial autoregressive model [230]. Further details are given in electronic supplementary material, technical appendix.

Although random effects and random fields are more computationally intensive than the other statistical approaches, there are a number of statistical methods and programs built to fit these types of models which aim to overcome computational issues [228,231,232]. These models are able to account for dependency between a large number of regions or observations without overfitting or introducing multicollinearity that causes issues when using spatial covariates. The structure of random effects and random fields must be determined before the model-fitting process, potentially introducing subjectivity into the model-fitting process, although they can be visualized which can help generate hypotheses and identify additional factors that may not have been

accounted for within the original model. Within this review, we only identified two spatial structures that were used within these models: distance based and neighbourhood based. These structures are adequate if spatial connectivity exists between close observations but we did not identify structures that would allow for other assumptions, such as long-distance movement of hosts and vectors, to be incorporated into a statistical model.

3.2.2. Machine learning methods

We identified two methods that were used to account for spatial connectivity within machine learning models: the inclusion of spatial covariates, and the development of movement matrices that aim to replicate human movement behaviour.

3.2.2.1. Spatial covariates

Five papers included spatial covariates as inputs for their machine learning algorithms. These spatial covariates included cases from neighbouring regions [233–235], the number of people travelling between regions based on air travel [234], public transportation networks [20] or a gravity model that aimed to replicate human commuting behaviour [236], and the distance between countries [236]. The inclusion of spatial covariates as inputs is compatible with all machine learning models and, if the data are available, does not require any additional computation.

3.2.2.2. Movement matrices

We identified two papers that constructed a matrix reflecting the movement of people between districts using public transportation data [19,237]. Both papers used this matrix, similar to the one shown in figure 3*d*, to weight layers within a neural network model, allowing the algorithm to predict the number of dengue cases across the study area while accounting for connectivity arising from human mobility. Although both studies used public transportation information to create their matrices, they could be constructed using movement models that aim to replicate human commuting behaviour, such as gravity or radiation models [238] (figure 3*e*), or other proxies such as distance-based functions where data are not available (figure 3*c*).

3.2.3. Mechanistic models

There were two methods used to account for spatial connectivity in mechanistic models identified by this review: movement matrices and spatial parameters.

3.2.3.1. Movement matrices

There were 21 studies (21/34, 61.8%) included in the review that used a movement matrix within a mechanistic model to account for spatial connectivity [19,32,239–257]; all these studies assumed that connectivity arose from either host or vector movement. These models treated subgroups of the host and/or vector populations as nodes in a network with values of the matrix reflecting movement between those nodes. Examples of these matrices constructed using different assumptions and data are given in figure 3. Matrices were constructed using human movement data from Twitter [32,251,256], air travel [239,249,250] or public transportation [19], using movement models that aimed to replicate human commuting behaviour [32,241,243,244,246,248,254,255,257], distance [242] or using a fixed value based on the type of neighbourhood [252,253]. Two studies estimated people's home and work addresses using mobile phone data and simulated movement between those [245,247], and two simulated the short flight distance of mosquitoes by allowing movement into neighbouring cells [240,245].

3.2.3.2. Spatial parameters

Thirteen studies (13/34, 38.2%) included spatial parameters within the model equations that aimed to account for connectivity [67,258–269]. Unlike movement matrices, these were directly incorporated into the model equations to update the population within a given compartment, or as a proxy for another process. Spatial parameters included the force of infection calculated using a distance-based kernel [259,260] and mosquito abundance estimated using a GAM containing a spatial random field [258]. Some models updated the population within compartments based on spatial parameters, either using a fixed-distance dispersion value [264–266], or calculating the proportion leaving regions using mobile phone records [263], air travel [262] or movement models [262,269]. One study used a mechanistic model but estimated the number of infected people using a CAR model [67].

3.3. Spatial connectivity assumptions

We collected details on the assumptions that authors made about how spatial connectivity arises within the data, regardless of the model type or method used. Although the exact assumptions differed between studies, all could be grouped into one or more of the following categories:

1. distance based,
2. human movement,
3. vector movement.

This section presents the advantages, disadvantages and methods used to implement these assumptions. A summary of these points with guidance on their ideal uses are provided in table 3.

3.3.1. Distance based

There were 200 (200/248, 80.6%) studies that assumed connectivity existed between observations or regions if and only if they were close. Although this was by far the most common assumption observed in this review, it was not explicitly stated in many of the studies. Twenty-two studies stated that they used a distance-based assumption as close regions were more likely to share characteristics such as climate systems, protective behaviours (e.g. bed net use), socioeconomic and demographic factors, vector ecology and land use type.

The majority of studies making a distance-based assumption of connectivity used a statistical model, only five studies used a mechanistic model and three used machine learning. The most common method for including distance-based connectivity within a model was the inclusion of a random effect or random field with a covariance structure defined by distance or neighbours ($n = 162$). Other methods included using spatial covariates ($n = 16$), such as the incidence rate in neighbouring regions or distance between observations, and local regression models fitted using data from nearby regions, weighted by distance ($n = 20$).

Table 3. The advantages, disadvantages and application of connectivity assumptions.

connectivity assumption	advantages	disadvantages	application
distance based	easy to obtain data can be incorporated into all types of model can be used as a proxy for shared characteristics that cannot be observed	oversimplifies process of transmission misses connectivity between distant regions difficult to define how 'close' regions should be to be considered connected	small-scale studies where unobservable processes, such as shared behaviours, create spatial connectivity. Not appropriate where long-distance connections are expected to exist due to travel. Basis of most statistical approaches identified in this review, e.g. GWR and mixed effect models
human movement	shown to be an important part of disease transmission for mosquito-borne diseases can account for connectivity between distant observations as well as close	difficult to quantify and obtain data, often requiring a proxy such as distance to be used data often have a number of biases may not be necessary for malaria studies in small-scale studies of endemic areas	*Aedes* or *Culex*-borne diseases in endemic settings where commuting leads to increased exposure, studies in areas that are disease-naive or nearing elimination at risk of (re-)introduction from long-distance movement such as immigration. More popular in mechanistic approaches such as metapopulation or agent-based models that allow complex movement matrices to incorporated. Only spatial covariates were able to reflect this connectivity in statistical methods
vector movement	an important part of the disease transmission process for all mosquito-borne diseases	difficult or impossible to obtain data due to the short flight distances of most mosquitoes, would not be necessary if considering a large area or a short-term study	small-scale studies or long-term forecasts, particularly malaria studies where transmission generally occurs at night. Due to a lack of data, a proxy must be used such as distance based on known flight distances of mosquitoes. May be included to account for differences in exposure levels across space

One of the main advantages of making a distance-based assumption of connectivity is that measures of connectivity (either distance or contiguity) are easy to obtain from geographical data. Contiguity is usually defined with chess analogies: rook contiguity defines neighbours as those sharing a common edge or border, whereas queen contiguity also includes regions sharing a common vertex. Another advantage of using one of these approaches is that there are a number of well-established models (particularly in statistical analysis) that were designed or adapted to incorporate this information, such as GWR and CAR models.

The main drawback of assuming connectivity is solely based on distance is that it may oversimplify the process, particularly for mosquito-borne diseases which require interaction between a susceptible host and an infectious vector. One of the most common models based on the assumption that connectivity exists between neighbouring regions, the Besag, Yorke and Mollié model (one example of a CAR model), states that these assumptions are reasonable if the disease is non-contagious and rare, which is not the case for mosquito-borne diseases [273]. Although regions are more likely to share characteristics with close regions, it is hard to define where this 'closeness' ends and how similar places should be before they are considered connected. Most studies assumed that characteristics were shared between neighbours or within a set distance; however, applying the same rule for all shared characteristics may miss some heterogeneity or exaggerate connectivity.

3.3.2. Human movement

We identified 50 studies that assumed spatial connectivity was related to human movement; most used mechanistic models ($n = 28$, figure 4) which are able to include complex mobility matrices (see metapopulation and agent-based models in electronic supplementary material, technical appendix, and figure 3 for more details). Other methods used to account for human movement within models included spatial covariates based on the number of people moving between regions, random effects which assumed people were more likely to travel to neighbouring regions, and a bespoke statistical model which simulated home and work addresses based on public transport journeys [46].

Studies were more likely to assume spatial connectivity arose through human mobility if the disease was transmitted

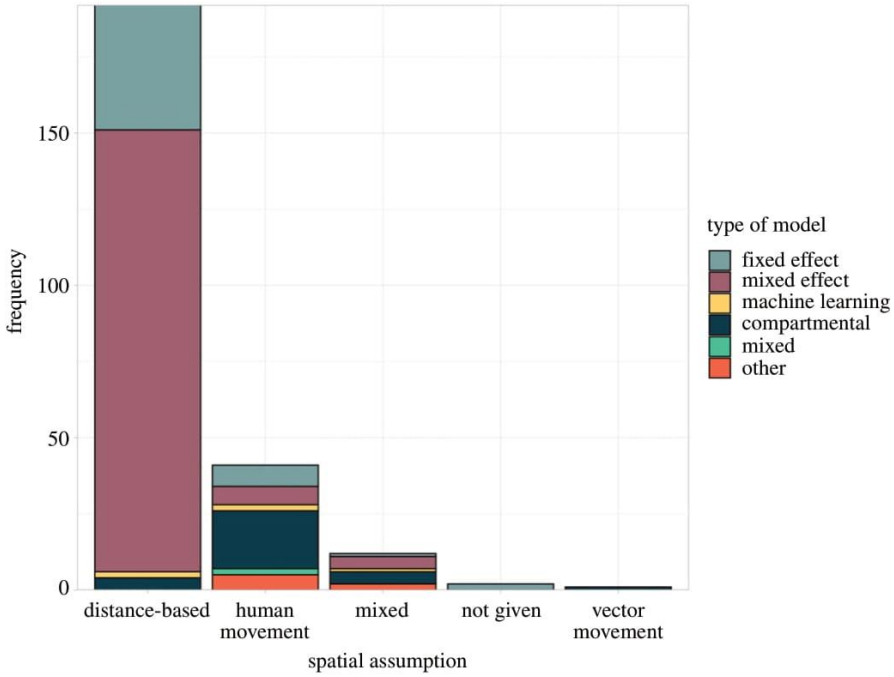


Figure 4. Connectivity assumption by model type. The number of spatial modelling studies that assumed connectivity is based on distance, human movement or vector movement (bars) separated by model type. The vast majority of statistical models (fixed and mixed effect models) assumed that connectivity was based on distance, whereas compartmental models were more likely to assume human movement drives connectivity.

by a mosquito of the *Aedes* genus (figure 5); this included dengue fever, chikungunya, yellow fever and Zika. *Aedes* mosquitoes are most active during the day, meaning interaction between host and vector is influenced by commuting behaviour [274], whereas *Anopheles* mosquitoes are night-biters and are more likely associated with vector movement or migration [275,276]. Less than half ($n = 22$) of the studies in this group used human mobility data to inform the spatial component of the model. Human mobility datasets included mobile phone GPS data, geo-located tweets, air travel information, public transportation networks and surveys. Other studies used a proxy such as distance or movement models, which replicate human commuting behaviours. The most common movement models were the gravity and radiation models. Both models assume that the movement of people is related to the population at each location and the distance between them; the radiation model also takes account of the population between locations under the assumption that people are less likely to commute to distant places when opportunities exist closer to home [238].

Unlike distance-based methods, the human mobility assumption allows for long-distance connections which may be important to the disease process, particularly in the region at risk of (re-)introduction of disease from imported cases. Prior studies have identified the importance of human mobility in the transmission of mosquito-borne diseases and found that failure to adequately account for this can lead to biased or invalid inferences [7,32,247,263,272,274,277]. However, human movement data can be difficult to obtain and may not be representative of all demographic and socioeconomic groups [272].

3.3.3. Vector movement

We identified 10 studies that explicitly stated they assumed spatial connectivity arose from vector movement; all these studies used a fixed distance or adjacency as a proxy for vector movement as adequate movement data was not available. One model included wind speed to account for vector movement as this extended the potential flight distance of mosquitoes, another weighted vector movement to adjacent tiles making this more likely if adjacent tiles contained humans or breeding grounds. There was only one study in this review that assumed all connectivity arose from vector movement, all others included other assumptions.

4. Discussion

This review provides the first comprehensive overview of spatial models, of any type, used to investigate the transmission of mosquito-borne pathogens, and the connectivity assumptions that underpin them. The last 10 years have seen a rapid increase in the number of spatial modelling studies of mosquito-borne diseases and the variety of approaches used. We identified 17 distinct spatial models that were used to explore the transmission of mosquito-borne pathogens to humans. These were classified as either statistical, machine learning or mechanistic; the choice of model should depend on the aim of the study, the type of data available and the information required from the modelling output. Statistical models are able to explore relationships between variables when sufficient data are available and can be used to make predictions or inferences about an outcome of interest. Unlike mechanistic models, they do not require an in-depth knowledge of the underlying biological process of the disease, although this can be used to improve the model. However, statistical models require a large amount of data to provide precise estimates, making them more suited to well-established diseases. They are able to make predictions within the scope of the data used

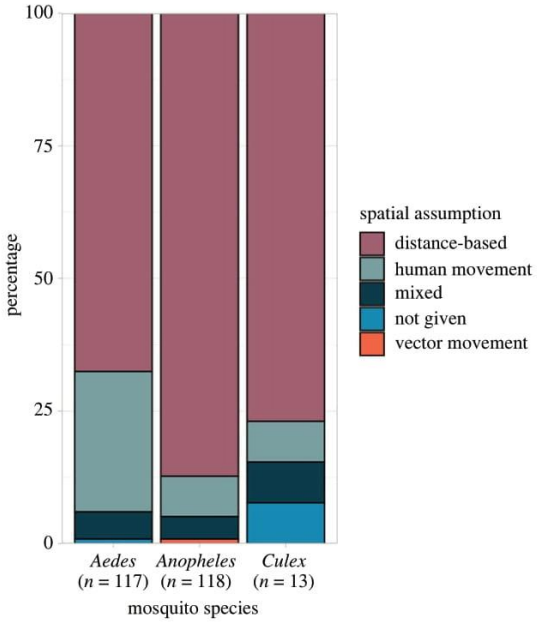


Figure 5. Connectivity assumptions by mosquito species. The percentage of studies modelling a disease transmitted by each mosquito species that assumed spatial connectivity is related to the distance between regions or observations (using a distance-based function or a neighbourhood structure), human movement or vector movement. Dengue fever, chikungunya, yellow fever and Zika were transmitted by mosquitoes of the *Aedes* genus; malaria was transmitted by mosquitoes of the *Anopheles* genus, and Japanese encephalitis, Rift Valley fever and West Nile fever were transmitted by mosquitoes of the *Culex* genus.

to fit them but are not recommended for causal investigations or extrapolation well beyond the data. Mechanistic models are more able to make causal inferences as they model the disease transmission process rather than the data itself; however, they are only able to do this within the specific setting for which they have been parameterized. Parameters can be taken from previous experiments where data are not available, making them particularly useful in settings where data are sparse or for newly (re-)emerging diseases. An example of this can be found in Zhang *et al*. [239] where parameters were 'borrowed' from other settings. Care should be taken when parameterizing mechanistic models in this way as processes may differ in ways that are not apparent at the model-fitting stage. By contrast, machine learning methods require a large amount of data but use flexible algorithms that allow them to learn patterns from rich, complex data. Although machine learning can be used to make inferences about data, most algorithms focus on making the most accurate predictions possible from available data rather than understanding underlying associations [270]. As with statistical models, they are inappropriate where there is a lack of data and are not recommended for making predictions or causal inferences well outside the range of data used to fit them [271].

Connectivity assumptions differed between mosquito species, indicating that authors consider mosquito behaviour and biting patterns when deciding which spatial model and assumptions are most appropriate. For example, dengue fever is transmitted by day-biting *Aedes* mosquitoes and is influenced by local movement or commuting [274], whereas *Anopheles*-borne malaria is transmitted by vectors most active between dusk and dawn so is influenced by proximity to vector breeding grounds and bed net use [275,276]. *Anopheles*-borne pathogens were more likely to be modelled assuming connectivity was driven by distance, potentially a proxy for vector movement because of the short flight span of vectors. *Aedes*- and *Culex*-borne pathogens were more likely modelled assuming human movement or proximity drives connectivity as this accounts for people commuting or moving to nearby regions/cities (figure 3). An alternative explanation could be that *Aedes*-borne emerging diseases (e.g. chikungunya and Zika) were more likely to be modelled using a mechanistic framework, allowing for the inclusion of complex movement matrices. The majority of statistical models within this review included a random effect to account for spatial connectivity, all of which used either a distance- or neighbourhood-based covariance structure. There were no random effect model structures that explicitly adjusted for connectivity arising from human movement.

Many studies included in this review did not explicitly state the assumptions they made about how connectivity arises. Often, assumptions had to be deduced from the data and spatial methods used in the studies. Although the vast majority of studies appeared to assume that regions were connected to neighbours or based on the distance between them, it is possible some used this as a proxy for another assumption, such as shared characteristics or human movement, where data were not available. Prior studies have discussed the difficulty of quantifying human behaviour when modelling infectious diseases [272]. Where mobility data are not available, movement models that aim to replicate commuting patterns, such as gravity and radiation models, were found to give similar results when modelling the spread of dengue fever compared to actual human movement data from geo-located Tweets [278]. These may help to avoid some of the issues surrounding privacy and bias when using a mobile phone or social media data to inform models, and where certain sections of the population, such as children and older adults, may be under represented. Some studies have suggested that radiation models are more accurate at representing commuting networks than mobile phone GPS data when compared to official census surveys in central locations [279].

This review provides a synthesis of the modelling approaches and spatial connectivity assumptions used to research mosquito-borne disease transmission to humans, but does not comment on the quality of these approaches. It is important to remember that more complex methods are not necessarily better and care should be taken to identify the most parsimonious method to address a studies' aim. Choice of the model should depend on the research question, the disease studied, the spatial scale and availability of the data and the way in which spatial connectivity is assumed to occur.

Data accessibility. Data extracted from the studies included in this systematic review are available from https://github.com/sophie-a-lee/mbd_connectivity_review and archived in a permanent repository on Zenodo (http://doi.org/10.5281/zenodo.4706866) [280].

Authors' contributions. All authors were involved in defining the search strategy. S.A.L. and C.I.J. performed the search and screening of articles. S.A.L. drafted the article with feedback, input and guidance

from C.I.J., T.E. and R.L. All authors read and approved the final manuscript.

Competing interests. The authors declare no conflicts of interest.

Funding. S.A.L. was supported by a Royal Society Research Grant for Research Fellows. C.I.J. receives funding from the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1). R.L. was supported by a Royal Society Dorothy Hodgkin Fellowship.

Acknowledgements. We would like to thank members of the Planetary Health Infectious Disease Lab at the London School of Hygiene & Tropical Medicine for their useful discussions and input.

References

1. World Health Organization. 2017 UNICEF. Global vector control response 2017–2030.
2. Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. 2017 Epidemic arboviral diseases: priorities for research and public health. *Lancet Infect. Dis.* **17**, e101–e106. (doi:10.1016/S1473-3099(16)30518-7)
3. Paixão ES, Teixeira MG, Rodrigues LC. 2018 Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases. *BMJ Glob. Health.* **3**(Suppl. 1), e000530. (doi:10.1136/bmjgh-2017-000530)
4. Khan K *et al.* 2014 Assessing the origin of and potential for international spread of chikungunya virus from the Caribbean. *PLoS Curr.* **6**. (doi:10.1371/currents.outbreaks.2134a0a7bf37fd8d388181539fea2da5)
5. Schaffner F, Mathis A. 2014 Dengue and dengue vectors in the WHO European region: past, present, and scenarios for the future. *Lancet Infect. Dis.* **14**, 1271–1280. (doi:10.1016/S1473-3099(14)70834-5)
6. World Health Organization. 2020 World malaria report 2020: 20 years of global progress and challenges.
7. Prothero RM. 1977 Disease and mobility: a neglected factor in epidemiology. *Int. J. Epidemiol.* **6**, 259–267. (doi:10.1093/ije/6.3.259)
8. Auchincloss AH, Gebreab SY, Mair C, Diez Roux AV. 2012 A review of spatial methods in epidemiology, 2000–2010. *Annu. Rev. Public Health* **33**, 107–122. (doi:10.1146/annurev-publhealth-031811-124655)
9. Reiner Jr RC *et al.* 2013 A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. *J. R Soc. Interface.* **10**, 20120921. (doi:10.1098/rsif.2012.0921)
10. Tobler WR. 1970 A computer movie simulating urban growth in the detroit region. *Econ. Geogr.* **46**(sup1), 234–240. (doi:10.2307/143141)
11. Lessler J, Azman AS, Grabowski MK, Salje H, Rodriguez-Barraquer I. 2016 Trends in the mechanistic and dynamic modeling of infectious diseases. *Curr. Epidemiol. Rep.* **3**, 212–222. (doi:10.1007/s40471-016-0078-4)
12. Lessler J, Cummings DAT. 2016 Mechanistic models of infectious disease and their impact on public health. *Am. J. Epidemiol.* **183**, 415–422. (doi:10.1093/aje/kww021)
13. Bzdok D, Krzywinski M, Altman N. 2017 Machine learning: a primer. *Nat. Methods* **14**, 1119–1120. (doi:10.1038/nmeth.4526)
14. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. 2009 Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* **6**, e1000097. (doi:10.1371/journal.pmed.1000097)
15. European Centre for Disease Prevention and Control. 2020 Mosquito-borne diseases. European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/en/mosquito-borne-diseases (accessed 21 July 2020).
16. World Health Organization. 2020 Mosquito-borne diseases. http://www.who.int/neglected_diseases/vector_ecology/mosquito-borne-diseases/en/ (accessed 21 July 2020).
17. R Core Team. 2019 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org.
18. Lee SA. 2021 A data and R code to accompany 'Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions' (version v1.0.0). Zenodo. (doi:10.5281/zenodo.4706866)
19. Bomfim R, Pei S, Shaman J, Yamana T, Makse HA, Andrade JS, Lima Neto AS, Furtado V. 2020 Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *J. R. Soc. Interface* **17**, 20200691. (doi:10.1098/rsif.2020.0691)
20. Chen Y, Yang Z, Jing Q, Huang J, Guo C, Yang K, Chen A, Lu J. 2020 Effects of natural and socioeconomic factors on dengue transmission in two cities of China from 2006 to 2017. *Sci. Total Environ.* **724**, 138200. (doi:10.1016/j.scitotenv.2020.138200)
21. Astutik S, Rahayudi B, Iskandar A, Fitriani R. 2013 Bayesian spatial-temporal autologistic regression model on dengue hemorrhagic fever in East Java, Indonesia. *Appl. Math Sci.* **7**, 435–443. (doi:10.12785/amis/072L08)
22. Wu PC, Lay JG, Guo HR, Lin CY, Lung SC, Su HJ. 2009 Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. *Sci. Total Environ.* **407**, 2224–2233. (doi:10.1016/j.scitotenv.2008.11.034)
23. DeGroote JP, Sugumaran R. 2012 National and regional associations between human West Nile virus incidence and demographic, landscape, and land use conditions in the coterminous United States. *Vector-Borne Zoonotic Dis.* **12**, 657–665. (doi:10.1089/vbz.2011.0786)
24. Jain R, Sontisirikit S, Iamsirithaworn S, Prendinger H. 2019 Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infect. Dis.* **19**, 272. (doi:10.1186/s12879-019-3874-x)
25. Impoinvil DE, Solomon T, Schluter WW, Rayamajhi A, Bichha RP, Shakya G, Caminade C, Baylis M. 2011 The spatial heterogeneity between Japanese encephalitis incidence distribution and environmental variables in Nepal. *PLoS ONE* **6**, e22192. (doi:10.1371/journal.pone.0022192)
26. Chuang TW, Ng KC, Nguyen TL, Chaves LF. 2018 Epidemiological characteristics and space-time analysis of the 2015 dengue outbreak in the metropolitan region of Tainan city, Taiwan. *Int. J. Environ. Res. Public Health.* **15**, 396. (doi:10.3390/ijerph15030396)
27. Wen T-H, Tsai C-T. 2016 Evaluating the role of disease importation in the spatiotemporal transmission of indigenous dengue outbreak. *Appl. Geogr.* **76**, 137–146. (doi:10.1016/j.apgeog.2016.09.020)
28. Xu Z, Bambrick H, Pongsumpun P, Tang IM, Yakob L, Devine G, Frentiu FD, Williams G, Hu W. 2020 Does Bangkok have a central role in the dengue dynamics of Thailand? *Parasit. Vectors* **13**, 1–9. (doi:10.1186/s13071-019-3862-4)
29. Gunderson AK *et al.* 2020 Malaria transmission and spillover across the Peru–Ecuador border: a spatiotemporal analysis. *Int. J. Environ. Res. Public Health* **17**, 7434. (doi:10.3390/ijerph17207434)
30. Ashmore P, Lindahl JF, Colón-González FJ, Sinh Nam V, Quang Tan D, Medley GF. 2020 Spatiotemporal and socioeconomic risk factors for dengue at the province level in Vietnam, 2013–2015: clustering analysis and regression model. *Trop. Med. Infect. Dis.* **5**, 81. (doi:10.3390/tropicalmed5020081)
31. Tao H, Wang K, Zhuo L, Li X, Li Q, Liu Y, Xu Y. 2019 A comprehensive framework for studying diffusion patterns of imported dengue with individual-based movement data. *Int. J. Geogr. Inf. Sci.* **34**, 604–624. (doi:10.1080/13658816.2019.1684497)
32. Kraemer MUG *et al.* 2018 Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore. *Pakistan. Epj Data Sci.* **7**, 1–7. (doi:10.1140/epjds/s13688-018-0144-x)
33. Salami D, Capinha C, Martins MdR, Sousa CA. 2020 Dengue importation into Europe: a network connectivity-based approach. *PLoS ONE* **15**, e0230274. (doi:10.1371/journal.pone.0230274)
34. Ramadona AL, Tozan Y, Lazuardi L, Rocklöv J. 2019 A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. *PLoS Negl. Trop. Dis.* **13**, e0007298. (doi:10.1371/journal.pntd.0007298)
35. Cauchemez S, Ledrans M, Poletto C, Quénel P, De Valk H, Colizza V, Boëlle PY. 2014 Local and regional spread of chikungunya fever in the Americas.

Eurosurveillance **19**, 20854. (doi:10.2807/1560-7917.ES2014.19.28.20854)

36. Kraemer MU *et al.* 2017 Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. *Lancet Infect. Dis.* **17**, 330–338. (doi:10.1016/S1473-3099(16)30513-8)
37. Da Silva-Nunes M, Codeço CT, Malafronte RS, Da Silva NS, Juncansen C, Muniz PT, Da Silva NS, Da Silva-Nunes M. 2008 Malaria on the Amazonian frontier: transmission dynamics, risk factors, spatial distribution, and prospects for control. *Am. J. Trop. Med. Hyg.* **79**, 624–635. (doi:10.4269/ajtmh.2008.79.624)
38. Parra MCP, Fávaro EA, Dibo MR, Mondini A, Eiras ÁE, Kroon EG, Teixeira MM, Nogueira ML, Chiaravalloti-Neto F. 2018 Using adult *Aedes aegypti* females to predict areas at risk for dengue transmission: a spatial case–control study. *Acta Trop.* **182**, 43–53. (doi:10.1016/j.actatropica.2018.02.018)
39. Wang X, Su L, Zhu H, Hu W, An J, Wang C, Qiannan E, Qi X, Zhuang G. 2020 Long-term epidemiological dynamics of Japanese encephalitis infection in Gansu Province, China: a spatial and temporal analysis. *Am. J. Trop. Med. Hyg.* **103**, 2065–2076. (doi:10.4269/ajtmh.20-0179)
40. Kazembe LN. 2007 Spatial modelling and risk factors of malaria incidence in northern Malawi. *Acta Trop.* **102**, 126–137. (doi:10.1016/j.actatropica.2007.04.012)
41. Gomes MFC, Codeço CT, Bastos LS, Lana RM. 2020 Measuring the contribution of human mobility to malaria persistence. *Malar. J.* **19**, 404. (doi:10.1186/s12936-020-03474-4)
42. Griffith DA. 2005 A comparison of six analytical disease mapping techniques as applied to West Nile virus in the coterminous United States. *Int. J. Health Geogr.* **4**, 18. (doi:10.1186/1476-072X-4-18)
43. Tevie J, Bohara A, Valdez RB. 2014 Examination of the geographical variation in human West Nile virus: a spatial filtering approach. *Epidemiol. Infect.* **142**, 2522–2529. (doi:10.1017/S0950268814000090)
44. Kim S, Kim Y. 2019 Spatially filtered multilevel analysis on spatial determinants for malaria occurrence in Korea. *Int. J. Environ. Res. Public Health* **16**, 11. (doi:10.3390/ijerph16071250)
45. Griffith DA. 2000 A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2**, 141–156. (doi:10.1007/PL00011451)
46. Prem K, Lau MSY, Tam CC, Ho MZJ, Ng LC, Cook AR. 2019 Inferring who-infected-whom-where in the 2016 Zika outbreak in Singapore—a spatio-temporal model. *J. R Soc. Interface* **16**, 20180604. (doi:10.1098/rsif.2018.0604)
47. Delmelle E, Hagenlocher M, Kienberger S, Casas I. 2016 A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta Trop.* **164**, 169–176. (doi:10.1016/j.actatropica.2016.08.028)
48. Robertson C, Pant DK, Joshi DD, Sharma M, Dahal M, Stephen C. 2013 Comparative spatial dynamics of Japanese encephalitis and acute encephalitis syndrome in Nepal. *PLoS ONE* **8**, e66168. (doi:10.1371/journal.pone.0066168)
49. Ehlkes L *et al.* 2014 Geographically weighted regression of land cover determinants of *Plasmodium falciparum* transmission in the Ashanti Region of Ghana. *Int. J. Health Geogr.* **13**, 1–11. (doi:10.1186/1476-072X-13-35)
50. Atique S, Chan TC, Chen CC, Hsu CY, Iqtidar S, Louis VR, Shabbir SA, Chuang T-W. 2018 Investigating spatio-temporal distribution and diffusion patterns of the dengue outbreak in Swat, Pakistan. *J. Infect. Public Health* **11**, 550–557. (doi:10.1016/j.jiph.2017.12.003)
51. Manyangadze T, Chimbari MJ, Macherera M, Mukaratirwa S. 2017 Micro-spatial distribution of malaria cases and control strategies at ward level in Gwanda district, Matabeleland South, Zimbabwe. *Malar. J.* **16**, 1–11. (doi:10.1186/s12936-017-2116-1)
52. Khormi HM, Kumar L. 2011 Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study. *Sci. Total Environ.* **409**, 4713–4719. (doi:10.1016/j.scitotenv.2011.08.028)
53. Acharya BK, Cao CX, Lakes T, Chen W, Naeem S, Pandit S. 2018 Modeling the spatially varying risk factors of dengue fever in Jhapa district, Nepal, using the semi-parametric geographically weighted regression model. *Int. J. Biometeorol.* **62**, 1973–1986. (doi:10.1007/s00484-018-1601-8)
54. Hasyim H, Nursafingi A, Haque U, Montag D, Groneberg DA, Dhimal M, Kuch U, Müller R. 2018 Spatial modelling of malaria cases associated with environmental factors in South Sumatra, Indonesia. *Malar. J.* **17**, 1–15. (doi:10.1186/s12936-017-2149-5)
55. Homan T *et al.* 2016 Spatially variable risk factors for malaria in a geographically heterogeneous landscape, western Kenya: an explorative study. *Malar. J.* **15**, 1. (doi:10.1186/s12936-015-1044-1)
56. Ren H, Wu W, Li T, Yang Z. 2019 Urban villages as transfer stations for dengue fever epidemic: a case study in the Guangzhou, China. *PLoS Negl. Trop. Dis.* **13**, e0007350. (doi:10.1371/journal.pntd.0007350)
57. Lin C-H, Wen T-H. 2011 Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. *Int. J. Environ. Res. Public Health* **8**, 2798–2815. (doi:10.3390/ijerph8072798)
58. Grillet ME, Barrera R, Martínez JE, Berti J, Fortin MJ. 2010 Disentangling the effect of local and global spatial variation on a mosquito-borne infection in a neotropical heterogeneous environment. *Am. J. Trop. Med. Hyg.* **82**, 194–201. (doi:10.4269/ajtmh.2010.09-0040)
59. Yang D, Xu C, Wang J, Zhao Y. 2017 Spatiotemporal epidemic characteristics and risk factor analysis of malaria in Yunnan Province, China. *BMC Public Health* **17**, 1–10. (doi:10.1186/s12889-016-3954-4)
60. Gopal S, Ma Y, Xin C, Pitts J, Were L. 2019 Characterizing the spatial determinants and prevention of malaria in Kenya. *Int. J. Environ. Res. Public Health.* **16**, 5078. (doi:10.3390/ijerph16245078)
61. de Oliveira Padilha MA, de Oliveira Melo J, Romano G, de Lima MVM, Alonso WJ, Sallum MAM, Laporta GZ. 2019 Comparison of malaria incidence rates and socioeconomic–environmental factors between the states of Acre and Rondonia: a spatio-temporal modelling study. *Malar. J.* **18**, 04. (doi:10.1186/s12936-019-2938-0)
62. Ren H, Zheng L, Li Q, Yuan W, Lu L. 2017 Exploring determinants of spatial variations in the dengue fever epidemic using geographically weighted regression model: a case study in the joint Guangzhou–Foshan area, China, 2014. *Int. J. Environ. Res. Public Health* **14**, 1518. (doi:10.3390/ijerph14121518)
63. Halimi M, Farajzadeh M, Delavari M, Takhtardeshir A, Moradi A. 2014 Modelling spatial relationship between climatic conditions and annual parasite incidence of malaria in southern part of Sistan&Balouchistan Province of Iran using spatial statistic models. *Asian Pac. J. Trop. Dis.* **4**(Suppl. 1), S167–S172. (doi:10.1016/S2222-1808(14)60434-5)
64. Ge Y, Song Y, Wang J, Liu W, Ren Z, Peng J, Lu B. 2017 Geographically weighted regression-based determinants of malaria incidences in northern China. *Trans. GIS* **21**, 934–953. (doi:10.1111/tgis.12259)
65. Anjos Rd, Nóbrega RS, Ferreira HdS, Lacerda Ad, Sousa-Neves Nd. 2020 Exploring local and global regression models to estimate the spatial variability of Zika and chikungunya cases in Recife, Brazil. *Rev. Soc. Bras. Med. Trop.* **53**, e20200027. (doi:10.1590/0037-8682-0027-2020)
66. Brunsdon C, Fotheringham S, Charlton M. 1998 Geographically weighted regression. *J. R. Stat. Soc. Ser. Stat.* **47**, 431–443. (doi:10.1111/1467-9884.00145)
67. Samat N, Percy D. 2012 Vector-borne infectious disease mapping with stochastic difference equations: an analysis of dengue disease in Malaysia. *J. Appl. Stat.* **39**, 2029–2046. (doi:10.1080/02664763.2012.700450)
68. Flórez-Lozano K *et al.* 2020 Spatial distribution of the relative risk of Zika virus disease in Colombia during the 2015–2016 epidemic from a Bayesian approach. *Int. J. Gynecol. Obstet.* **148**(S2), 55–60. (doi:10.1002/ijgo.13048)
69. Ferreira GS, Schmidt AM. 2006 Spatial modelling of the relative risk of dengue fever in Rio de Janeiro for the epidemic period between 2001 and 2002. *Braz. J. Probab. Stat.* **20**, 29–47.
70. Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, Alegana VA, Fall IS, Snow RW. 2014 The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000–10: A spatial and temporal analysis of transmission intensity. *Lancet* **383**, 1739–1747. (doi:10.1016/S0140-6736(13)62566-0)
71. Ssempiira J *et al.* 2018 The effect of case management and vector-control interventions on space-time patterns of malaria incidence in Uganda. *Malar. J.* 2018/04/14 ed. **17**, 162. (doi:10.1186/s12936-018-2312-7)

72. Wangdi K, Canavati SE, Ngo TD, Tran LK, Nguyen TM, Tran DT, Martin NJ, Clements ACA. 2018 Analysis of clinical malaria disease patterns and trends in Vietnam 2009–2015. *Malar. J.* **17**, 15. (doi:10.1186/s12936-018-2478-z)
73. Villalta D, Guenni L, Rubio-Palis Y, Ramírez Arbeláez R. 2013 Bayesian space-time modeling of malaria incidence in Sucre state, Venezuela. *AStA Adv. Stat. Anal.* **97**, 151–171. (doi:10.1007/s10182-012-0190-9)
74. Jaya I, Abdullah AS, Hermawan E, Ruchjana BN. 2016 Bayesian spatial modeling and mapping of dengue fever: a case study of dengue fever in the city of Bandung. *Indonesia. Int. J. Appl. Math Stat.* **54**, 94–103.
75. Thway AM, Rotejanaprasert C, Sattabongkot J, Lawawirojwong S, Thi A, Hlaing TM, Soe TM, Kaewkungwal J. 2018 Bayesian spatiotemporal analysis of malaria infection along an international border: Hlaingbwe Township in Myanmar and Tha-Song-Yang District in Thailand. *Malar. J.* **17**, 428. (doi:10.1186/s12936-018-2574-0)
76. Jaya I, Folmer H. 2020 Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia. *J. Geogr. Syst.* **22**, 105–142. (doi:10.1007/s10109-019-00311-4)
77. Rouamba T, Samadoulougou S, Tinto H, Alegana VA, Kirakoya-Samadoulougou F. 2020 Bayesian spatiotemporal modeling of routinely collected data to assess the effect of health programs in malaria incidence during pregnancy in Burkina Faso. *Sci. Rep.* **10**, 14. (doi:10.1038/s41598-019-56010-z)
78. Rotejanaprasert C, Ekapirat N, Areechokchai D, Maude RJ. 2020 Bayesian spatiotemporal modeling with sliding windows to correct reporting delays for real-time dengue surveillance in Thailand. *Int. J. Health Geogr.* **19**, 1–13. (doi:10.1186/s12942-020-00199-0)
79. Snow RW, Kibuchi E, Karuri SW, Sang G, Gitonga CW, Mwandawiro C, Bejon P, Noor AM. 2015 Changing malaria prevalence on the Kenyan coast since 1974: climate, drugs and vector control. *PLoS ONE* **10**, e0128792. (doi:10.1371/journal.pone.0128792)
80. Reid HL, Haque U, Roy S, Islam N, Clements ACA. 2012 Characterizing the spatial and temporal variation of malaria incidence in Bangladesh, 2007. *Malar. J.* **11**, 170. (doi:10.1186/1475-2875-11-170)
81. Aswi A, Cramb S, Duncan E, Hu W, White G, Mengersen K. 2020 Climate variability and dengue fever in Makassar, Indonesia: Bayesian spatio-temporal modelling. *Spat. Spatio-Temporal Epidemiol.* **33**, 100335. (doi:10.1016/j.sste.2020.100335)
82. Lowe R *et al.* 2014 Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. *Lancet Infect. Dis.* **14**, 619–626. (doi:10.1016/S1473-3099(14)70781-9)
83. Wang G, Minnis RB, Belant JL, Wax CL. 2010 Dry weather induces outbreaks of human West Nile virus infections. *BMC Infect. Dis.* **10**, 38. (doi:10.1186/1471-2334-10-38)
84. Wimberly MC, Hildreth MB, Boyte SP, Lindquist E, Kightlinger L. 2008 Ecological niche of the 2003 West Nile virus epidemic in the northern Great Plains of the United States. *PLoS ONE* **3**, e3744. (doi:10.1371/journal.pone.0003744)
85. Ouédraogo M, Rouamba T, Samadoulougou S, Kirakoya-Samadoulougou F. 2020 Effect of free healthcare policy for children under five years old on the incidence of reported malaria cases in burkina faso by bayesian modelling: 'not only the ears but also the head of the hippopotamus'. *Int. J. Environ. Res. Public Health* **17**, 417. (doi:10.3390/ijerph17020417)
86. Umer MF, Zofeen S, Majeed A, Hu W, Qi X, Zhuang G. 2019 Effects of socio-environmental factors on malaria infection in Pakistan: a Bayesian spatial analysis. *Int. J. Environ. Res. Public Health* **16**, 1365. (doi:10.3390/ijerph16081365)
87. Abd Naeeim NS, Rahman NA. 2017 Estimating relative risk for dengue disease in Peninsular Malaysia using INLA. *Malays. J. Fundam. Appl. Sci.* **13**, 721–727. (doi:10.11113/mjfas.v0n0.575)
88. Alegana VA, Atkinson PM, Wright JA, Kamwi R, Uusiku P, Katokele S, Snow RW, Noor AM. 2013 Estimation of malaria incidence in northern Namibia in 2009 using Bayesian conditional-autoregressive spatial-temporal models. *Spat. Spatio-Temporal Epidemiol.* **7**, 25–36. (doi:10.1016/j.sste.2013.09.001)
89. Mukhsar, Abapihi B, Sani A, Cahyono E, Adam P, Abdullah FA. 2016 Extended convolution model to bayesian spatio-temporal for diagnosing the DHF endemic locations. *J. Interdiscip. Math.* **19**, 233–244. (doi:10.1080/09720502.2015.1047591)
90. Husnina Z, Clements ACA, Wangdi K. 2019 Forest cover and climate as potential drivers for dengue fever in Sumatra and Kalimantan 2006–2016: a spatiotemporal analysis. *Trop. Med. Int. Health* **24**, 888–898. (doi:10.1111/tmi.13248)
91. Lekdee K, Ingsrisawang L. 2013 Generalized linear mixed models with spatial random effects for spatio-temporal data: an application to dengue fever mapping. *J. Math. Stat.* **9**, 137–143. (doi:10.3844/jmssp.2013.137.143)
92. Nkurunziza H, Gebhardt A, Pilz J. 2011 Geo-additive modelling of malaria in Burundi. *Malar. J.* **10**, 234. (doi:10.1186/1475-2875-10-234)
93. Zayeri F, Salehi M, Pirhosseini H. 2011 Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran. *Asian Pac. J. Trop. Med.* **4**, 985–992. (doi:10.1016/S1995-7645(11)60231-9)
94. Chien LC, Yu HL. 2014 Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. *Environ. Int.* **73**, 46–56. (doi:10.1016/j.envint.2014.06.018)
95. Ssempiira J, Kissa J, Nambuusi B, Mukooyo E, Opigo J, Makumbi F, Kasasa S, Vounatsou P. 2018 Interactions between climatic changes and intervention effects on malaria spatio-temporal dynamics in Uganda. *Parasit. Epidemiol. Control* **3**, e00070. (doi:10.1016/j.parepi.2018.e00070)
96. Zhao X, Cao M, Feng HH, Fan H, Chen F, Feng Z, Li X, Zhou X-H. 2014 Japanese encephalitis risk and contextual risk factors in Southwest China: a Bayesian hierarchical spatial and spatiotemporal analysis. *Int. J. Environ. Res. Public Health* **11**, 4201–4217. (doi:10.3390/ijerph110404201)
97. Martínez-Bello DA, López-Quílez A, Prieto AT. 2019 Joint estimation of relative risk for dengue and Zika infections, Colombia, 2015–2016. *Emerg. Infect. Dis.* **25**, 1118–1126. (doi:10.3201/eid2506.180392)
98. Zacarias OP, Andersson M. 2010 Mapping malaria incidence distribution that accounts for environmental factors in Maputo Province - Mozambique. *Malar. J.* **9**, 79. (doi:10.1186/1475-2875-9-79)
99. Huang F, Zhou S, Zhang S, Zhang H, Li W. 2011 Meteorological factors-based spatio-temporal mapping and predicting malaria in central China. *Am. J. Trop. Med. Hyg.* **85**, 560–567. (doi:10.4269/ajtmh.2011.11-0156)
100. Alegana VA, Wright JA, Nahzat SM, Butt W, Sediqi AW, Habib N, Snow RW, Atkinson PM, Noor AM. 2014 Modelling the incidence of *Plasmodium vivax* and *Plasmodium falciparum* malaria in Afghanistan 2006–2009. *PLoS ONE* **9**, e102304. (doi:10.1371/journal.pone.0102304)
101. Restrepo AC, Baker P, Clements ACA. 2014 National spatial and temporal patterns of notified dengue cases, Colombia 2007–2010. *Trop. Med. Int. Health* **19**, 863–871. (doi:10.1111/tmi.12325)
102. Santos-Vega M, Bouma MJ, Kohli V, Pascual M. 2016 Population density, climate variables and poverty synergistically structure spatial risk in Urban Malaria in India. *PLoS Negl. Trop. Dis.* **10**, e0005155. (doi:10.1371/journal.pntd.0005155)
103. Lowe R, Cazelles B, Paul R, Rodó X. 2016 Quantifying the added value of climate information in a spatio-temporal dengue model. *Stoch. Environ. Res. Risk Assess.* **30**, 2067–2078. (doi:10.1007/s00477-015-1053-1)
104. Lowe R, Chirombo J, Tompkins AM. 2013 Relative importance of climatic, geographic and socio-economic determinants of malaria in Malawi. *Malar. J.* **12**, 416. (doi:10.1186/1475-2875-12-416)
105. Sani A, Abapihi B, Mukhsar M, Kadir K. 2015 Relative risk analysis of dengue cases using convolution extended into spatio-temporal model. *J. Appl. Stat.* **42**, 2509–2519. (doi:10.1080/02664763.2015.1043863)
106. Samat NA, Pei Zhen W. 2017 Relative risk estimation for dengue disease mapping in Malaysia based on Besag, York and Mollié model. *Pertanika J. Sci. Technol.* **25**, 759–766.
107. Martínez-Bello DA, López-Quílez A, Torres Prieto A. 2017 Relative risk estimation of dengue disease at small spatial scale. *Int. J. Health Geogr.* **16**, 31. (doi:10.1186/s12942-017-0104-x)
108. Kristiani F, Yong B, Irawan R. 2016 Relative risk estimation of dengue disease in Bandung, Indonesia, using Poisson-gamma and bym models considering the severity level. *J. Teknol.* **78**, 57–64.
109. Rouamba T, Samadoulougou S, Tinto H, Alegana VA, Kirakoya-Samadoulougou F. 2020 Severe-malaria

infection and its outcomes among pregnant women in Burkina Faso health-districts: hierarchical Bayesian space-time models applied to routinely-collected data from 2013 to 2018. *Spat. Spatio-Temporal Epidemiol.* **33**, 100333. (doi:10.1016/j.sste.2020.100333)

110. Manh BH *et al.* 2011 Social and environmental determinants of malaria in space and time in Viet Nam. *Int. J. Parasitol.* **41**, 109–116. (doi:10.1016/j.ijpara.2010.08.005)
111. Clements AC, Barnett AG, Cheng ZW, Snow RW, Zhou HN. 2009 Space-time variation of malaria incidence in Yunnan province, China. *Malar. J.* 2009/08/04 ed. **8**, 180. (doi:10.1186/1475-2875-8-180)
112. de Almeida AS, de Andrade Medronho R, Ortiz Valencia LI. 2009 Spatial analysis of dengue and the socioeconomic context of the city of Rio de Janeiro (Southeastern Brazil). *Rev. Saude Publica* **43**, 666–673. (doi:10.1590/S0034-89102009000400013)
113. Honorato T, Lapa PPA, Sales CMM, Reis-Santos B, Tristão-Sá R, Bertolde AI, Maciel ELN. 2014 Spatial analysis of distribution of dengue cases in Espírito Santo, Brazil, in 2010: use of Bayesian model. *Rev. Bras. Epidemiol.* **17**, 150–159. (doi:10.1590/1809-4503201400060013)
114. Noh M, Lee Y, Oh S, Chu C, Gwack J, Youn SK, Cho SH, Lee WJ, Huh S. 2012 Spatial and temporal distribution of *Plasmodium vivax* malaria in Korea estimated with a hierarchical generalized linear model. *Osong Public Health Res. Perspect.* **3**, 192–198. (doi:10.1016/j.phrp.2012.11.003)
115. Phanitchat T *et al.* 2019 Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. *BMC Infect. Dis.* **19**, 743. (doi:10.1186/s12879-019-4379-3)
116. Wangdi K, Clements ACA, Du T, Nery SV. 2018 Spatial and temporal patterns of dengue infections in Timor-Leste, 2005–2013. *Parasit. Vectors* **11**, 9. (doi:10.1186/s13071-017-2588-4)
117. Ouédraogo M, Samadoulougou S, Rouamba T, Hien H, Sawadogo JEM, Tinto H, Alegana VA, Speybroeck N, Kirakoya-Samadoulougou F. 2018 Spatial distribution and determinants of asymptomatic malaria risk among children under 5 years in 24 districts in Burkina Faso. *Malar. J.* **17**, 460. (doi:10.1186/s12936-018-2606-9)
118. Kikuti M *et al.* 2015 Spatial distribution of dengue in a Brazilian Urban slum setting: role of socioeconomic gradient in disease risk. *PLoS Negl. Trop. Dis.* **9**, e0003937. (doi:10.1371/journal.pntd.0003937)
119. Costa JV, Donalisio MR, Silveira LdA. 2013 Spatial distribution of dengue incidence and socio-environmental conditions in Campinas, Sao Paulo State, Brazil, 2007. *Cad. Saude Publica* **29**, 1522–1532. (doi:10.1590/S0102-311X2013001200005)
120. Teixeira TdA, Cruz OG. 2011 Spatial modeling of dengue and socio-environmental indicators in the city of Rio de Janeiro, Brazil. *Cad. Saude Publica* **27**, 591–602. (doi:10.1590/S0102-311X2011000300019)
121. Hu WB, Clements A, Williams G, Tong SL, Mengersen K. 2012 Spatial patterns and socioecological drivers of dengue fever transmission in Queensland, Australia. *Environ. Health Perspect.* **120**, 260–266. (doi:10.1289/ehp.1003270)
122. Abellana R, Ascaso C, Aponte J, Saute F, Nhalungo D, Nhacolo A, Alonso P. 2008 Spatio-seasonal modeling of the incidence rate of malaria in Mozambique. *Malar. J.* **7**, 228. (doi:10.1186/1475-2875-7-228)
123. Mabaso MLH, Vounatsou P, Midzi S, Da Silva J, Smith T. 2006 Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. *Int. J. Health Geogr.* **5**, 20. (doi:10.1186/1476-072X-5-20)
124. Martínez-Bello DA, López-Quílez A, Torres Prieto A. 2018 Spatio-temporal modeling of Zika and dengue infections within Colombia. *Int. J. Environ. Res. Public Health* **15**, 1376. (doi:10.3390/ijerph15071376)
125. Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, Barcellos C. 2011 Spatio-temporal modelling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil. *Comput. Geosci.* **37**, 371–381. (doi:10.1016/j.cageo.2010.01.008)
126. Nobre AA, Schmidt AM, Lopes HF. 2005 Spatio-temporal models for mapping the incidence of malaria in Pará. *Environmetrics* **16**, 291–304. (doi:10.1002/env.704)
127. Bett B *et al.* 2019 Spatiotemporal analysis of historical records (2001–2012) on dengue fever in Vietnam and development of a statistical model for forecasting risk. *PLoS ONE* **14**, e0224353. (doi:10.1371/journal.pone.0224353)
128. McHale TC, Romero-Vivas CM, Fronterre C, Arango-Padilla P, Waterlow NR, Nix CD, Falconar AK, Cano J. 2019 Spatiotemporal heterogeneity in the distribution of chikungunya and Zika virus case incidences during their 2014 to 2016 epidemics in Barranquilla, Colombia. *Int. J. Environ. Res. Public Health* **16**, 1759. (doi:10.3390/ijerph16101759)
129. Martínez-Bello D, López-Quílez A, Prieto AT. 2018 Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stoch. Environ. Res. Risk Assess.* **32**, 1587–1601. (doi:10.1007/s00477-017-1461-5)
130. Chien L, Lin R, Liao Y, Sy F, Perez A. 2018 Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. *BMC Infect. Dis.* **18**, 180. (doi:10.1186/s12879-018-3085-x)
131. Lowe R, Bailey TC, Stephenson DB, Jupp TE, Graham RJ, Barcellos C, Carvalho M. 2013 The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Stat. Med.* **32**, 864–883. (doi:10.1002/sim.5549)
132. Wijayanti SPM, Porphyre T, Chase-Topping M, Rainey SM, McFarlane M, Schnettler E, Biek R, Kohl A. 2016 The importance of socio-economic versus environmental risk factors for reported dengue cases in Java, Indonesia. *PLoS Negl. Trop. Dis.* **10**, 15. (doi:10.1371/journal.pntd.0004964)
133. Mabaso MLH, Craig M, Vounatsou P, Smith T. 2005 Towards empirical description of malaria seasonality in southern Africa: the example of Zimbabwe. *Trop. Med. Int. Health* **10**, 909–918. (doi:10.1111/j.1365-3156.2005.01462.x)
134. Adin A, Martínez-Bello DA, López-Quílez A, Ugarte MD. 2018 Two-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia. *PLoS ONE* **13**, e0203382. (doi:10.1371/journal.pone.0203382)
135. Achcar JA, Martinez EZ, Souza AD, Tachibana VM, Flores EF. 2011 Use of Poisson spatiotemporal regression models for the Brazilian Amazon Forest: malaria count data. *Rev. Soc. Bras. Med. Trop.* 2012/01/11 ed. **44**, 749–754. (doi:10.1590/S0037-86822011000600019)
136. Hanandita W, Tampubolon G. 2016 Geography and social distribution of malaria in Indonesian Papua: a cross-sectional study. *Int. J. Health Geogr.* **15**, 13. (doi:10.1186/s12942-016-0043-y)
137. Zhang SB, Hu WB, Qi X, Zhuang GH. 2018 How socio-environmental factors are associated with Japanese encephalitis in Shaanxi, China—a Bayesian spatial analysis. *Int. J. Environ. Res. Public Health* **15**, 13. (doi:10.3390/ijerph15040608)
138. Battle KE *et al.* 2019 Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. *Lancet* **394**, 332–343. (doi:10.1016/S0140-6736(19)31096-7)
139. Yu HL, Lee CH, Chien LC. 2016 A spatiotemporal dengue fever early warning model accounting for nonlinear associations with hydrological factors: a Bayesian maximum entropy approach. *Stoch. Environ. Res. Risk Assess.* **30**, 2127–2141. (doi:10.1007/s00477-016-1328-1)
140. Maheu-Giroux M, Castro MC. 2013 Impact of community-based larviciding on the prevalence of malaria infection in Dar es Salaam, Tanzania. *PLoS ONE* **8**, e71638. (doi:10.1371/journal.pone.0071638)
141. Abd Naeeim NS, Rahman NA, Fahimi FAM. 2020 A spatial-temporal study of dengue in Peninsular Malaysia for the year 2017 in two different space–time model. *J. Appl. Stat.* **47**, 739–756. (doi:10.1080/02664763.2019.1648391)
142. Kleinschmidt I, Sharp B, Mueller I, Vounatsou P. 2002 Rise in malaria incidence rates in South Africa: a small-area spatial analysis of variation in time trends. *Am. J. Epidemiol.* **155**, 257–264. (doi:10.1093/aje/155.3.257)
143. Bisanzio D, Mutuku F, LaBeaud AD, Mungai PL, Muinde J, Busaidy H, Mukoko D, King CH, Kitron U. 2015 Use of prospective hospital surveillance data to define spatiotemporal heterogeneity of malaria risk in coastal Kenya. *Malar. J.* **14**, 482. (doi:10.1186/s12936-015-1006-7)
144. Nakhapakorn K, Sancharoen W, Mutchimwong A, Jirakajohnkool S, Onchang R, Rotejanaprasert C, Tantrakarnapa K, Paul R. 2020 Assessment of urban land surface temperature and vertical city associated with dengue incidences. *Remote Sens.* **12**, 3802. (doi:10.3390/rs12223802)
145. Semakula M, Niragire F, Faes C. 2020 Bayesian spatio-temporal modeling of malaria risk in Rwanda. *PLoS ONE* **15**, e0238504. (doi:10.1371/journal.pone.0238504)

146. Akter R, Hu W, Gatton M, Bambrick H, Cheng J, Tong S. 2020 Climate variability, socio-ecological factors and dengue transmission in tropical Queensland, Australia: a Bayesian spatial analysis. *Environ. Res.* **195**, 110285. (doi:10.1016/j.envres.2020.110285)
147. Kristiani F, Claudia Y, Yong B, Hilsdon A-M. 2020 A comparative analysis of frequentist and Bayesian approaches to estimate dengue disease transmission in Bandung-Indonesia. *J. Stat. Manag. Syst.* **23**, 1543–1559. (doi:10.1080/09720510.2020.1756049)
148. Ye J, Moreno-Madriñán MJ. 2020 Comparing different spatio-temporal modeling methods in dengue fever data analysis in Colombia during 2012–2015. *Spat. Spatio-Temporal Epidemiol.* **34**, 100360. (doi:10.1016/j.sste.2020.100360)
149. da Conceição Araújo D, Dos Santos AD, Lima SVMA, Vaez AC, Cunha JO, de Araújo KCGM. 2020 Determining the association between dengue and social inequality factors in north-eastern Brazil: a spatial modelling. *Geospat. Health* **15**, 71–80. (doi:10.4081/gh.2020.854)
150. Aswi A, Cramb S, Duncan E, Mengersen K. 2020 Evaluating the impact of a small number of areas on spatial estimation. *Int. J. Health Geogr.* **19**, 39. (doi:10.1186/s12942-020-00233-1)
151. Jaya IGNM, Folmer H. 2020 Identifying spatiotemporal clusters by means of agglomerative hierarchical clustering and Bayesian regression analysis with spatiotemporally varying coefficients: methodology and application to dengue disease in Bandung, Indonesia. *Geogr. Anal.* (doi:10.1111/gean.12264)
152. Tsheten T, Clements ACA, Gray DJ, Wangchuk S, Wangdi K. 2020 Spatial and temporal patterns of dengue incidence in Bhutan: a Bayesian analysis. *Emerg. Microbes Infect.* **9**, 1360–1371. (doi:10.1080/22221751.2020.1775497)
153. Wangdi K, Canavati SE, Ngo TD, Nguyen TM, Tran LK, Kelly GC, Martin NJ, Clements ACA. 2020 Spatial and temporal patterns of Malaria in Phu Yen Province, Vietnam, from 2005 to 2016. *Am. J. Trop. Med. Hyg.* **103**, 1540–1548. (doi:10.4269/ajtmh.20-0392)
154. Wangdi K, Xu Z, Suwannatrai AT, Kurscheid J, Lal A, Namgay R, Glass K, Gray DJ, Clements ACA. 2020 A spatio-temporal analysis to identify the drivers of malaria transmission in Bhutan. *Sci. Rep.* **10**, 7060. (doi:10.1038/s41598-020-63896-7)
155. Carabali M, Harper S, Lima Neto AS, Dos Santos de Sousa G, Caprara A, Restrepo BN, Kaufman JS. 2020 Spatiotemporal distribution and socioeconomic disparities of dengue, chikungunya and Zika in two Latin American cities from 2007 to 2017. *Trop. Med. Int. Health* **26**, 301–315. (doi:10.1111/tmi.13530)
156. Puggioni G, Couret J, Serman E, Akanda AS, Ginsberg HS. 2020 Spatiotemporal modeling of dengue fever risk in Puerto Rico. *Spat. Spatio-Temporal Epidemiol.* **35**, 100375. (doi:10.1016/j.sste.2020.100375)
157. Mallya S, Sander B, Roy-Gagnon MH, Taljaard M, Jolly A, Kulkarni MA. 2018 Factors associated with human West Nile virus infection in Ontario: a generalized linear mixed modelling approach. *BMC Infect. Dis.* **18**, 141. (doi:10.1186/s12879-018-3052-6)
158. Gething PW *et al.* 2012 A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl. Trop. Dis.* **6**, e1814. (doi:10.1371/journal.pntd.0001814)
159. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, Tatem AJ, Hay SI. 2011 A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, 1–16. (doi:10.1186/1475-2875-10-378)
160. Hay SI *et al.* 2009 A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med.* **6**, e1000048. (doi:10.1371/journal.pmed.1000048)
161. Alegana VA *et al.* 2016 Advances in mapping malaria for elimination: fine resolution modelling of *Plasmodium falciparum* incidence. *Sci. Rep.* **6**, 1–14. (doi:10.1038/s41598-016-0001-8)
162. Fornace KM *et al.* 2016 Association between landscape factors and spatial patterns of *Plasmodium knowlesi* infections in Sabah, Malaysia. *Emerg. Infect. Dis.* **22**, 201–208. (doi:10.3201/eid2202.150656)
163. Reid H *et al.* 2010 Baseline spatial distribution of malaria prior to an elimination programme in Vanuatu. *Malar. J.* **9**, 150. (doi:10.1186/1475-2875-9-150)
164. Stensgaard AS, Vounatsou P, Onapa AW, Simonsen PE, Pedersen EM, Rahbek C, Kristensen TK. 2011 Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: predictors of risk and geographical patterns of co-endemicity. *Malar. J.* **10**, 298. (doi:10.1186/1475-2875-10-298)
165. Amratia P, Psychas P, Abuaku B, Ahorlu C, Millar J, Oppong S, Koram K, Valle D. 2019 Characterizing local-scale heterogeneity of malaria risk: a case study in Bunkpurugu-Yunyoo district in northern Ghana. *Malar. J.* **18**, 81. (doi:10.1186/s12936-019-2703-4)
166. Diggle P, Moyeed R, Rowlingson B, Thomson M. 2002 Childhood malaria in the Gambia: a case-study in model-based geostatistics. *J. R. Stat. Soc. Ser. C Appl. Stat.* **51**, 493–506. (doi:10.1111/1467-9876.00283)
167. Giorgi E, Sesay SSS, Terlouw DJ, Diggle PJ. 2015 Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *J. R. Stat. Soc. Ser. C Stat. Soc.* **178**, 445–464. (doi:10.1111/rssa.12069)
168. Craig MH, Sharp BL, Mabaso ML, Kleinschmidt I. 2007 Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. *Int. J. Health Geogr.* **6**, 1–15. (doi:10.1186/1476-072X-6-44)
169. Giardina F, Kasasa S, Sié A, Utzinger J, Tanner M, Vounatsou P. 2014 Effects of vector-control interventions on changes in risk of malaria parasitaemia in sub-Saharan Africa: a spatial and temporal analysis. *Lancet Glob. Health* **2**, e601–e615. (doi:10.1016/S2214-109X(14)70300-6)
170. Ashton RA *et al.* 2015 Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys. *Am. J. Trop. Med. Hyg.* **93**, 168–177. (doi:10.4269/ajtmh.14-0620)
171. Giardina F, Franke J, Vounatsou P. 2015 Geostatistical modelling of the malaria risk in Mozambique: effect of the spatial resolution when using remotely-sensed imagery. *Geospat. Health* **10**, 232–238. (doi:10.4081/gh.2015.333)
172. Aimone AM, Brown P, Owusu-Agyei S, Zlotkin SH, Cole DC. 2017 Impact of iron fortification on the geospatial patterns of malaria and non-malaria infection risk among young children: a secondary spatial analysis of clinical trial data from Ghana. *BMJ Open* **7**, e013192. (doi:10.1136/bmjopen-2016-013192)
173. Noor AM, Alegana VA, Kamwi RN, Hansford CF, Ntomwa B, Katokele S, Snow RW. 2013 Malaria control and the intensity of *Plasmodium falciparum* transmission in Namibia 1969–1992. *PLoS ONE* 2013/05/15 ed. **8**, e63350. (doi:10.1371/journal.pone.0063350)
174. Reid H, Haque U, Clements AC, Tatem AJ, Vallely A, Ahmed SM, Islam A, Haque R. 2010 Mapping malaria risk in Bangladesh using Bayesian geostatistical models. *Am. J. Trop. Med. Hyg.* 2010/10/05 ed. **83**, 861–867. (doi:10.4269/ajtmh.2010.10-0154)
175. Nguyen M *et al.* 2020 Mapping malaria seasonality in Madagascar using health facility data. *BMC Med.* **18**, 1–11. (doi:10.1186/s12916-019-1443-1)
176. Arab A, Jackson MC, Kongoli C. 2014 Modelling the effects of weather and climate on malaria distributions in West Africa. *Malar. J.* **13**, 126. (doi:10.1186/1475-2875-13-126)
177. Samadoulougou S, Maheu-Giroux M, Kirakoya-Samadoulougou F, De Keukeleire M, Castro MC, Robert A. 2014 Multilevel and geo-statistical modeling of malaria risk in children of Burkina Faso. *Parasit. Vectors* **7**, 350. (doi:10.1186/1756-3305-7-350)
178. Elyazar IR, Gething PW, Patil AP, Rogayah H, Kusriastuti R, Wismarini DM, Tarmizi SN, Baird JK, Hay SI. 2011 *Plasmodium falciparum* malaria endemicity in Indonesia in 2010. *PLoS ONE* 2011/07/09 ed. **6**, e21315. (doi:10.1371/journal.pone.0021315)
179. Gething PW, Patil AP, Hay SI. 2010 Quantifying aggregated uncertainty in *Plasmodium falciparum* malaria prevalence and populations at risk via efficient space-time geostatistical joint simulation. *PLoS Comput. Biol.* **6**, e1000724. (doi:10.1371/journal.pcbi.1000724)
180. Hanks EM, Schliep EM, Hooten MB, Hoeting JA. 2015 Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26**, 243–254. (doi:10.1002/env.2331)
181. Chiaravalloti-Neto F *et al.* 2019 Seroprevalence for dengue virus in a hyperendemic area and associated socioeconomic and demographic factors using a cross-sectional design and a geostatistical approach,

state of Sao Paulo, Brazil. *BMC Infect. Dis.* **19**, 441. (doi:10.1186/s12879-019-4074-4)

182. Kazembe LN, Kleinschmidt I, Holtz TH, Sharp BL. 2006 Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. *Int. J. Health Geogr.* **5**, 41. (doi:10.1186/1476-072X-5-41)
183. Ayele DG, Zewotir TT, Mwambi HG. 2013 Spatial distribution of malaria problem in three regions of Ethiopia. *Malar. J.* **12**, 207. (doi:10.1186/1475-2875-12-207)
184. Noor AM, Clements AC, Gething PW, Moloney G, Borle M, Shewchuk T, Hay SI, Snow RW. 2008 Spatial prediction of *Plasmodium falciparum* prevalence in Somalia. *Malar. J.* 2008/08/23 ed. **7**, 159. (doi:10.1186/1475-2875-7-159)
185. Macharia PM, Giorgi E, Noor AM, Waqo E, Kiptui R, Okiro EA, Snow RW. 2018 Spatio-temporal analysis of *Plasmodium falciparum* prevalence to understand the past and chart the future of malaria control in Kenya. *Malar. J.* **17**, 1–13. (doi:10.1186/s12936-018-2489-9)
186. Kang SY *et al.* 2018 Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess *Plasmodium falciparum* endemicity trends between 2011 and 2016. *BMC Med.* **16**, 1–15. (doi:10.1186/s12916-017-0981-7)
187. Colborn KL, Giorgi E, Monaghan AJ, Gudo E, Candrinho B, Marrufo TJ, Colborn JM. 2018 Spatio-temporal modelling of weekly malaria incidence in children under 5 for early epidemic detection in Mozambique. *Sci. Rep.* **8**, 1–9. (doi:10.1038/s41598-018-27537-4)
188. Ssempiira J, Nambuusi B, Kissa J, Agaba B, Makumbi F, Kasasa S, Vounatsou P. 2017 The contribution of malaria control interventions on spatio-temporal changes of parasitaemia risk in Uganda during 2009–2014. *Parasit. Vectors* **10**, 450. (doi:10.1186/s13071-017-2393-0)
189. Noor AM *et al.* 2013 The receptive versus current risks of *Plasmodium falciparum* transmission in Northern Namibia: implications for elimination. *BMC Infect. Dis.* **13**, 1–10. (doi:10.1186/1471-2334-13-1)
190. Matthys B *et al.* 2006 Urban farming and malaria risk factors in a medium-sized town in Côte d'Ivoire. *Am. J. Trop. Med. Hyg.* **75**, 1223–1231. (doi:10.4269/ajtmh.2006.75.1223)
191. Kleinschmidt I, Sharp BL, Clarke GP, Curtis B, Fraser C. 2001 Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa. *Am. J. Epidemiol.* **153**, 1213–1221. (doi:10.1093/aje/153.12.1213)
192. Giorgi E, Osman AA, Hassan AH, Ali AA, Ibrahim F, Amran JG, Noor AM, Snow RW. 2018 Using non-exceedance probabilities of policy-relevant malaria prevalence thresholds to identify areas of low transmission in Somalia. *Malar. J.* **17**, 1–10. (doi:10.1186/s12936-018-2238-0)
193. Chirombo J, Lowe R, Kazembe L. 2014 Using structured additive regression models to estimate risk factors of malaria: analysis of 2010 Malawi malaria indicator survey data. *PLoS ONE* **9**, e101116. (doi:10.1371/journal.pone.0101116)
194. Guerra CA *et al.* 2019 Human mobility patterns and malaria importation on Bioko Island. *Nat. Commun.* **10**, 1–10. (doi:10.1038/s41467-019-10339-1)
195. Fornace KM *et al.* 2019 Local human movement patterns and land use impact exposure to zoonotic malaria in Malaysian Borneo. *Elife* **8**, e47602. (doi:10.7554/eLife.47602)
196. Raso G *et al.* 2012 Mapping malaria risk among children in Côte d'Ivoire using Bayesian geo-statistical models. *Malar. J.* **11**, 1–11. (doi:10.1186/1475-2875-11-160)
197. Gosoniu L, Vounatsou P. 2011 Non-stationary partition modeling of geostatistical data for malaria risk mapping. *J. Appl. Stat.* **38**, 3–13. (doi:10.1080/02664760903008961)
198. Cohen JM, Dlamini S, Novotny JM, Kandula D, Kunene S, Tatem AJ. 2013 Rapid case-based mapping of seasonal malaria transmission risk for strategic elimination planning in Swaziland. *Malar. J.* 2013/02/13 ed. **12**, 61. (doi:10.1186/1475-2875-12-61)
199. Valle D, Lima JMT. 2014 Large-scale drivers of malaria and priority areas for prevention and control in the Brazilian Amazon region using a novel multi-pathogen geospatial model. *Malar. J.* **13**, 13. (doi:10.1186/1475-2875-13-443)
200. Gosoniu L, Vounatsou P, Sogoba N, Smith T. 2006 Bayesian modelling of geostatistical malaria risk data. *Geospat. Health* 2008/08/08 ed. **1**, 127–139. (doi:10.4081/gh.2006.287)
201. Giardina F, Gosoniu L, Konate L, Diouf MB, Perry R, Gaye O, Faye O, Vounatsou P. 2012 Estimating the burden of malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 Data. *PLoS ONE* **7**, e32625. (doi:10.1371/journal.pone.0032625)
202. Janko M, Goel V, Emch M. 2019 Extending multilevel spatial models to include spatially varying coefficients. *Health Place* **60**, 102235. (doi:10.1016/j.healthplace.2019.102235)
203. Riedel N, Vounatsou P, Miller JM, Gosoniu L, Chizema-Kawesha E, Mukonka V, Steketee RW. 2010 Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modelling of the 2006 Zambia national malaria indicator survey (ZMIS). *Malar. J.* **9**, 37. (doi:10.1186/1475-2875-9-37)
204. Gosoniu L, Vounatsou P, Sogoba N, Maire N, Smith T. 2009 Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. *Comput. Stat. Data Anal.* **53**, 3358–3371. (doi:10.1016/j.csda.2009.02.022)
205. Salehi M, Mohammad K, Farahani MM, Zeraati H, Nourijelyani K, Zayeri F. 2008 Spatial modeling of malaria incidence rates in Sistan and Baluchistan province, Islamic Republic of Iran. *Saudi Med. J.* **29**, 1791–1796.
206. Raso G, Silué KD, Vounatsou P, Singer BH, Yapi A, Tanner M, Utzinger J, N'Goran EK. 2009 Spatial risk profiling of *Plasmodium falciparum* parasitaemia in a high endemicity area in Côte d'Ivoire. *Malar. J.* **8**, 252. (doi:10.1186/1475-2875-8-252)
207. Adegboye OA, Leung DHY, Wang YG. 2017 Analysis of spatial data with a nested correlation structure. *J. R. Stat. Soc. Ser. C Appl. Stat.* **67**, 329–354. (doi:10.1111/rssc.12230)
208. Sharmin S, Glass K, Viennet E, Harley D. 2018 Geostatistical mapping of the seasonal spread of under-reported dengue cases in Bangladesh. *PLoS Negl. Trop. Dis.* **12**, e0006947. (doi:10.1371/journal.pntd.0006947)
209. Sow A *et al.* 2020 Changes in the transmission dynamic of chikungunya virus in Southeastern Senegal. *Viruses* **12**, 196. (doi:10.3390/v12020196)
210. Ahmad H, Ali A, Fatima SH, Zaidi F, Khisroon M, Rasheed SB, Ullah I, Ullah S, Shakir M. 2020 Spatial modeling of dengue prevalence and kriging prediction of dengue outbreak in Khyber Pakhtunkhwa (Pakistan) using presence only data. *Stoch. Environ. Res. Risk Assess.* **34**, 1023–1036. (doi:10.1007/s00477-020-01818-9)
211. Routledge I *et al.* 2020 Tracking progress towards malaria elimination in China: individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. *PLoS Comput. Biol.* **16**, e1007707. (doi:10.1371/journal.pcbi.1007707)
212. Sedda L, Taylor BM, Eiras AE, Marques JT, Dillon RJ. 2020 Using the intrinsic growth rate of the mosquito population improves spatio-temporal dengue risk estimation. *Acta Trop.* **208**, 105519. (doi:10.1016/j.actatropica.2020.105519)
213. Rue H, Held L. 2005 *Gaussian Markov random fields: theory and applications*. Boca Raton, FL: CRC press.
214. Matérn B. 2013 *Spatial variation*, vol. 36. New York, NY: Springer Science & Business Media.
215. Chiaravalloti-Neto F, Pereira M, Fávaro EA, Dibo MR, Mondini A, Rodrigues-Junior AL, Chierotti AP, Nogueira M. 2015 Assessment of the relationship between entomologic indicators of *Aedes aegypti* and the epidemic occurrence of dengue virus 3 in a susceptible population, São José do Rio Preto, São Paulo, Brazil. *Acta Trop.* **142**, 167–177. (doi:10.1016/j.actatropica.2014.11.017)
216. Farinelli EC, Baquero OS, Stephan C, Chiaravalloti-Neto F. 2018 Low socioeconomic condition and the risk of dengue fever: a direct relationship. *Acta Trop.* **180**, 47–57. (doi:10.1016/j.actatropica.2018.01.005)
217. Charlwood JD, Tomas EVE, Braganca M, Cuamba N, Alifrangis M, Stanton M. 2015 Malaria prevalence and incidence in an isolated, meso-endemic area of Mozambique. *Peerj* **3**, 22. (doi:10.7717/peerj.1370)
218. Ugwu CLJ, Zewotir T. 2020 Evaluating the effects of climate and environmental factors on under-5 children malaria spatial distribution using generalized additive models (GAMs). *J. Epidemiol. Glob. Health* **10**, 304–314. (doi:10.2991/jegh.k.200814.001)
219. Mutucumarana CP *et al.* 2020 Geospatial analysis of dengue emergence in rural areas in the Southern Province of Sri Lanka. *Trans. R Soc. Trop. Med. Hyg.* **114**, 408–414. (doi:10.1093/trstmh/trz123)
220. Vazquez-Prokopec GM, Kitron U, Montgomery B, Horne P, Ritchie SA. 2010 Quantifying the spatial dimension of dengue virus epidemic spread within a tropical urban environment. *PLoS Negl. Trop. Dis.* **4**, e920. (doi:10.1371/journal.pntd.0000920)

221. Braga C, Luna CF, Martelli CM, de Souza WV, Cordeiro MT, Alexander N, Júnior JC, Marques ET. 2010 Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. *Acta Trop.* **113**, 234–240. (doi:10.1016/j.actatropica.2009.10.021)
222. Honório NA *et al.* 2009 Spatial evaluation and modeling of dengue seroprevalence and vector density in Rio de Janeiro, Brazil. *PLoS Negl. Trop. Dis.* **3**, e545. (doi:10.1371/journal.pntd.0000545)
223. Siqueira-Junior JB, Maciel IJ, Barcellos C, Souza WV, Carvalho MS, Nascimento NE, Oliveira RM, Morais-Neto O, Martelli CMT. 2008 Spatial point analysis based on dengue surveys at household level in central Brazil. *BMC Public Health* **8**, 361. (doi:10.1186/1471-2458-8-361)
224. Chien LC, Sy F, Pérez A. 2019 Identifying high risk areas of Zika virus infection by meteorological factors in Colombia. *BMC Infect. Dis.* **19**, 888. (doi:10.1186/s12879-019-4499-9)
225. Hundessa S *et al.* 2018 Projecting potential spatial and temporal changes in the distribution of *Plasmodium vivax* and *Plasmodium falciparum* malaria in China with climate change. *Sci. Total Environ.* **627**, 1285–1293. (doi:10.1016/j.scitotenv.2018.01.300)
226. Kazembe LN, Mathanga DP. 2016 Estimating risk factors of urban malaria in Blantyre, Malawi: a spatial regression analysis. *Asian Pac. J. Trop. Biomed.* **6**, 376–381. (doi:10.1016/j.apjtb.2016.03.011)
227. Cissoko M *et al.* 2020 Geo-epidemiology of malaria at the health area level, dire health district, Mali, 2013–2017. *Int. J. Environ. Res. Public Health* **17**, 3982. (doi:10.3390/ijerph17113982)
228. Wood SN. 2017 *Generalized additive models: an introduction with R*. Boca Raton, FL: CRC Press.
229. Watts MJ, Kotsila P, Mortyn PG, Sarto i Monteys V, Urzi Brancati C. 2020 Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico. *Int. J. Health Geogr.* **19**, 44. (doi:10.1186/s12942-020-00241-1)
230. Laguna F, Grillet ME, León JR, Ludeña C. 2017 Modelling malaria incidence by an autoregressive distributed lag model with spatial component. *Spat. Spatio-Temporal Epidemiol.* **22**, 27–37. (doi:10.1016/j.sste.2017.05.001)
231. Rue H, Martino S, Chopin N. 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 319–92. (doi:10.1111/j.1467-9868.2008.00700.x)
232. Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000 WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337. (doi:10.1023/A:1008929526011)
233. Halim S, Handojo A, Widodo I, Octavia T. 2020 Spatial multi-layer perceptron model for predicting dengue fever outbreaks in Surabaya. *Adv. Sci. Technol. Eng. Syst. J.* **5**, 103–108. (doi:10.25046/aj050514)
234. Akhtar M, Kraemer MUG, Gardner LM. 2019 A dynamic neural network model for predicting risk of Zika in real time. *BMC Med.* **17**, 171. (doi:10.1186/s12916-019-1389-3)
235. Haddawy P, Hasan AHMI, Kasantikul R, Lawpoolsri S, Sa-angchai P, Kaewkungwal J, Singhasivanon P. 2018 Spatiotemporal Bayesian networks for malaria prediction. *Artif. Intell. Med.* **84**, 127–138. (doi:10.1016/j.artmed.2017.12.002)
236. Rossi G, Karki S, Smith RL, Brown WM, Ruiz MO. 2018 The spread of mosquito-borne viruses in modern times: a spatio-temporal analysis of dengue and chikungunya. *Spat. Spatio-Temporal Epidemiol.* **26**, 113–125. (doi:10.1016/j.sste.2018.06.002)
237. Shi B, Liu J, Zhou X-N, Yang G-J. 2014 Inferring plasmodium vivax transmission networks from tempo-spatial surveillance data. *PLoS Negl. Trop. Dis.* **8**, e2682. (doi:10.1371/journal.pntd.0002682)
238. Simini F, González MC, Maritan A, Barabási A-L. 2012 A universal model for mobility and migration patterns. *Nature* **484**, 96–100. (doi:10.1038/nature10856)
239. Zhang Q *et al.* 2017 Spread of Zika virus in the Americas. *Proc. Natl Acad. Sci. USA* **114**, E4334–E4343. (doi:10.1073/pnas.1620161114)
240. Pizzitutti F, Pan W, Barbieri A, Miranda JJ, Feingold B, Guedes GR, Alarcon-Valenzuela J, Mena CF. 2015 A validated agent-based model to study the spatial and temporal heterogeneities of malaria incidence in the rainforest environment. *Malar. J.* **14**, 1030. (doi:10.1186/s12936-015-1030-7)
241. Zhu G *et al.* 2019 Effects of human mobility, temperature and mosquito control on the spatiotemporal transmission of dengue. *Sci. Total Environ.* **651**, 969–978. (doi:10.1016/j.scitotenv.2018.09.182)
242. Silal SP, Little F, Barnes KI, White LJ. 2015 Hitting a moving target: a model for malaria elimination in the presence of population movement. *PLoS ONE* **10**, e0144990. (doi:10.1371/journal.pone.0144990)
243. Senapati A, Sardar T, Ganguly KS, Ganguly KS, Chattopadhyay AK, Chattopadhyay J. 2019 Impact of adult mosquito control on dengue prevalence in a multi-patch setting: a case study in Kolkata (2014–2015). *J. Theor. Biol.* **478**, 139–152. (doi:10.1016/j.jtbi.2019.06.021)
244. Xue L, Scott HM, Cohnstaedt LW, Scoglio C. 2012 A network-based meta-population approach to model Rift Valley fever epidemics. *J. Theor. Biol.* **306**, 129–144. (doi:10.1016/j.jtbi.2012.04.029)
245. Karl S, Halder N, Kelso JK, Ritchie SA, Milne GJ. 2014 A spatial simulation model for dengue virus infection in urban areas. *BMC Infect. Dis.* **14**, 1–17. (doi:10.1186/1471-2334-14-447)
246. Moulay D, Pigné Y. 2013 A metapopulation model for chikungunya including populations mobility on a large-scale network. *J. Theor. Biol.* **318**, 129–139. (doi:10.1016/j.jtbi.2012.11.008)
247. Massaro E, Kondor D, Ratti C. 2019 Assessing the interplay between human mobility and mosquito borne diseases in urban environments. *Sci. Rep.* **9**, 1–13. (doi:10.1038/s41598-019-53127-z)
248. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, Engø-Monsen K, Buckee CO. 2015 Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl Acad. Sci.* **112**, 11 887–11 892. (doi:10.1073/pnas.1504964112)
249. Sun K *et al.* 2018 Quantifying the risk of local Zika virus transmission in the contiguous US during the 2015–2016 ZIKV epidemic. *BMC Med.* **16**, 195. (doi:10.1186/s12916-018-1185-5)
250. Gardner LM, Bota A, Gangavarapu K, Kraemer MUG, Grubaugh ND. 2018 Inferring the risk factors behind the geographical spread and transmission of Zika in the Americas. *PLoS Negl. Trop. Dis.* **12**, 25. (doi:10.1371/journal.pntd.0006194)
251. Seroussi I, Levy N, Yom-Tov E. 2020 Multi-season analysis reveals the spatial structure of disease spread. *Phys. Stat. Mech. Appl.* **547**, 124425. (doi:10.1016/j.physa.2020.124425)
252. Barrios E, Lee S, Vasilieva O. 2018 Assessing the effects of daily commuting in two-patch dengue dynamics: a case study of Cali, Colombia. *J. Theor. Biol.* **453**, 14–39. (doi:10.1016/j.jtbi.2018.05.015)
253. Stolerman LM, Coombs D, Boatto S. 2015 SIR-network model and its application to dengue fever. *SIAM J. Appl. Math.* **75**, 2581–2609. (doi:10.1137/140996148)
254. Zhu G *et al.* 2018 The spatiotemporal transmission of dengue and its driving mechanism: a case study on the 2014 dengue outbreak in Guangdong, China. *Sci. Total Environ.* **622–623**, 252–259. (doi:10.1016/j.scitotenv.2017.11.314)
255. Chadsuthi S, Althouse BM, Iamsirithaworn S, Triampo W, Grantz KH, Cummings DAT. 2018 Travel distance and human movement predict paths of emergence and spatial spread of chikungunya in Thailand. *Epidemiol. Infect.* **146**, 1654–1662. (doi:10.1017/S0950268818001917)
256. Kim M, Paini D, Jurdak R. 2019 Modeling stochastic processes in disease spread across a heterogeneous social system. *Proc. Natl Acad. Sci. USA* **116**, 401–406. (doi:10.1073/pnas.1801429116)
257. Zhu G, Liu J, Tan Q, Shi B. 2016 Inferring the spatio-temporal patterns of dengue transmission from surveillance data in Guangzhou, China. *PLoS Negl. Trop. Dis.* **10**, e0004633. (doi:10.1371/journal.pntd.0004633)
258. Li R, Xu L, Bjørnstad ON, Liu K, Song T, Chen A, Xu B, Liu Q, Stenseth NC. 2019 Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *Proc. Natl Acad. Sci. USA* **116**, 3624–3629. (doi:10.1073/pnas.1806094116)
259. Marini G, Guzzetta G, Marques Toledo CA, Teixeira M, Rosa R, Merler S. 2019 Effectiveness of ultra-low volume insecticide spraying to prevent dengue in a non-endemic metropolitan area of Brazil. *PLoS Comput. Biol.* **15**, e1006831. (doi:10.1371/journal.pcbi.1006831)
260. Guzzetta G, Marques-Toledo CA, Rosà R, Teixeira M, Merler S. 2018 Quantifying the spatial spread of dengue in a non-endemic Brazilian metropolis via transmission chain reconstruction. *Nat. Commun.* **9**, 2837. (doi:10.1038/s41467-018-05230-4)
261. Yu HL, Angulo JM, Cheng MH, Wu J, Christakos G. 2014 An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan).

Biom. J. **56**, 428–440. (doi:10.1002/bimj.201200270)

262. O'Reilly KM *et al.* 2018 Projecting the end of the Zika virus epidemic in Latin America: a modelling analysis. *BMC Med.* **16**, 180. (doi:10.1186/s12916-018-1158-8)
263. Ruktanonchai NW, DeLeenheer P, Tatem AJ, Alegana VA, Caughlin TT, Erbach-Schoenberg Ez, Lourenço C, Ruktanonchai CW, Smith DL. 2016 Identifying malaria transmission foci for elimination using human mobility data. *PLoS Comput. Biol.* **12**, e1004846. (doi:10.1371/journal.pcbi.1004846)
264. Fitzgibbon WE, Morgan JJ, Webb GF. 2017 An outbreak vector–host epidemic model with spatial structure: the 2015–2016 Zika outbreak in Rio De Janeiro. *Theor. Biol. Med. Model.* **14**, 7. (doi:10.1186/s12976-017-0051-z)
265. Miyaoka TY, Lenhart S, Meyer JFCA. 2019 Optimal control of vaccination in a vector-borne reaction–diffusion model applied to Zika virus. *J. Math. Biol.* **79**, 1077–1104. (doi:10.1007/s00285-019-01390-z)
266. Mukhtar AYA, Munyakazi JB, Ouifki R. 2020 Assessing the role of human mobility on malaria transmission. *Math. Biosci.* **320**, 108304. (doi:10.1016/j.mbs.2019.108304)
267. Gerardin J, Bever CA, Hamainza B, Miller JM, Eckhoff PA, Wenger EA. 2016 Optimal population-level infection detection strategies for malaria control and elimination in a spatial model of malaria transmission. *PLoS Comput. Biol.* **12**, e1004707. (doi:10.1371/journal.pcbi.1004707)
268. Núñez-López M, Alarcón Ramos L, Velasco-Hernández JX. 2021 Migration rate estimation in an epidemic network. *Appl. Math. Model* **89**, 1949–1964. (doi:10.1016/j.apm.2020.08.025)
269. Moore SM *et al.* 2018 Local and regional dynamics of chikungunya virus transmission in Colombia: the role of mismatched spatial heterogeneity. *BMC Med.* **16**, 1–6. (doi:10.1186/s12916-018-1127-2)
270. Jordan MI, Mitchell TM. 2015 Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260. (doi:10.1126/science.aaa8415)
271. Baker RE, Peña J-M, Jayamohan J, Jérusalem A. 2018 Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.* **14**, 20170660. (doi:10.1098/rsbl.2017.0660)
272. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Dis.* **3**, e481. (doi:10.1371/journal.pntd.0000481)
273. Besag J, York J, Mollié A. 1991 Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20. (doi:10.1007/BF00116466)
274. Stoddard ST *et al.* 2013 House-to-house human movement drives dengue virus transmission. *Proc. Natl Acad. Sci. USA* **110**, 994–999. (doi:10.1073/pnas.1213349110)
275. Dambach P, Schleicher M, Korir P, Ouedraogo S, Dambach J, Sié A, Dambach M, Becker N. 2018 Nightly biting cycles of *Anopheles* species in rural northwestern Burkina Faso. *J. Med. Entomol.* **55**, 1027–1034. (doi:10.1093/jme/tjy043)
276. Kabbale FG, Akol AM, Kaddu JB, Onapa AW. 2013 Biting patterns and seasonality of *Anopheles gambiae sensu lato* and *Anopheles funestus* mosquitoes in Kamuli District, Uganda. *Parasit. Vectors* **6**, 340. (doi:10.1186/1756-3305-6-340)
277. Pindolia DK, Garcia AJ, Wesolowski A, Smith DL, Buckee CO, Noor AM, Snow RW, Tatem AJ. 2012 Human movement data for malaria control and elimination strategic planning. *Malar. J.* **11**, 205. (doi:10.1186/1475-2875-11-205)
278. Kraemer MUG, Perkins TA, Cummings DA, Zakar R, Hay SI, Smith DL, Reiner Jr RC. 2015 Big city, small world: density, contact rates, and transmission of dengue across Pakistan. *J. R Soc. Interface* **12**, 20150468. (doi:10.1098/rsif.2015.0468)
279. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V. 2014 On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716. (doi:10.1371/journal.pcbi.1003716)
280. Lee SA *et al*. 2021 Data and R code to accompany 'Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions' (version v1.0.0). (Version V1.0.0). Zenodo. (http://doi.org/10.5281/zenodo.4706866)

4. A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping

Bridging section

In Chapter 3, I found that most spatial statistical models of mosquito-borne disease transmission assumed that spatial connectivity existed only because of distance, either between neighbouring regions or between close observations. The only statistical approach identified that could include human movement explicitly was the inclusion of spatial covariates, such as the number of people moving between areas, into a generalised linear model. However, these models require one covariate per connection which risks introducing multicollinearity, particularly in metropoles such as São Paulo, Brasilia and Rio de Janeiro which have connections across the whole of Brazil. There were no statistical approaches identified by the systematic review capable of accounting for multiple sources of spatial connectivity within the same model. In reality, spatial connectivity in dengue outbreak data likely arises due to multiple factors. For example, levels of immunity in the population, vector control measures, and climate variation which is likely shared between close areas, and long-distance connections created by people travelling between cities across Brazil.

In this chapter, I developed a novel statistical modelling framework capable of simultaneously accounting for multiple sources of spatial connectivity, including a complex human movement network (Objective 3). This model can be used to quantify the relative contribution of different sources of spatial connectivity (i.e., spatial autocorrelation between close areas and human movement) to the overall spatial structure of the outcome.

The chapter was published in the *Journal of the Royal Society Interface* in September 2022[1]. The published version of the paper is included below. Supplementary materials referred to in the paper can be found in Appendix D.

[1.] Lee SA, Economou T, Lowe R. 2022 A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping. *J. R. Soc. Interface* **19**, 20220440. (doi: 10.1098/rsif.2022.0440)

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1806358	**Title**	Ms
First Name(s)	Sophie Alice		
Surname/Family Name	Lee		
Thesis Title	Spatial Modelling of Emerging Infectious Diseases: Quantifying the Role of Climate, Cities and Connectivity on Dengue Expansion in Brazil		
Primary Supervisor	Dr Rachel Lowe		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Journal of the Royal Society Interface		
When was the work published?	21 September 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My contribution: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing – original draft, Writing – review & editing

SECTION E

Student Signature	Sophie Lee
Date	21/09/2022

Supervisor Signature	Rachel Lowe
Date	21/09/2022

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Lee SA, Economou T, Lowe R. 2022 A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping. *J. R. Soc. Interface* **19**: 20220440.
https://doi.org/10.1098/rsif.2022.0440

Received: 14 June 2022
Accepted: 31 August 2022

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
computational biology, biomathematics

Keywords:
infectious disease dynamics, vector-borne disease, spatial epidemiology, hierarchical modelling, spatial analysis, spatial connectivity

Author for correspondence:
Sophie A. Lee
e-mail: sophie.lee@lshtm.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6186135.

THE ROYAL SOCIETY PUBLISHING

A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping

Sophie A. Lee[1,2], Theodoros Economou[3] and Rachel Lowe[1,2,4,5]

[1]Centre for Mathematical Modelling of Infectious Diseases, and [2]Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, UK
[3]Climate and Atmosphere Research Centre, The Cyprus Institute, Nicosia, Cyprus
[4]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[5]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

SAL, 0000-0002-2049-9756; TE, 0000-0001-8697-1518; RL, 0000-0003-3939-7343

Spatial connectivity is an important consideration when modelling infectious disease data across a geographical region. Connectivity can arise for many reasons, including shared characteristics between regions and human or vector movement. Bayesian hierarchical models include structured random effects to account for spatial connectivity. However, conventional approaches require the spatial structure to be fully defined prior to model fitting. By applying penalized smoothing splines to coordinates, we create two-dimensional smooth surfaces describing the spatial structure of the data while making minimal assumptions about the structure. The result is a non-stationary surface which is setting specific. These surfaces can be incorporated into a hierarchical modelling framework and interpreted similarly to traditional random effects. Through simulation studies, we show that the splines can be applied to any symmetric continuous connectivity measure, including measures of human movement, and that the models can be extended to explore multiple sources of spatial structure in the data. Using Bayesian inference and simulation, the relative contribution of each spatial structure can be computed and used to generate hypotheses about the drivers of disease. These models were found to perform at least as well as existing modelling frameworks, while allowing for future extensions and multiple sources of spatial connectivity.

1. Introduction

When modelling infectious disease data across a geographical region, it is important to account for potential spatial connectivity between areas. For example, spatial connectivity may arise from human or vector movement contributing to the spread of a vector-borne disease, or unobservable climatic, behavioural, biological and socio-economic factors shared between areas. Conventionally, Bayesian hierarchical models aim to account for this spatial connectivity by including spatially structured random components within the model [1–3]. Fully Bayesian modelling approaches require the spatial structure of components to be defined prior to model fitting. However, the spatial structure of the data may not be fully known. A recent systematic review found that all Bayesian hierarchical models for mosquito-borne diseases used a distance-based spatial structure, assuming connectivity between regions only exists between neighbours or close observations [4].

Spatial autocorrelation in disease count data may be attributable to multiple sources of connectivity. For example, dengue incidence is associated with climate variation, vector control interventions and levels of immunity in the population which are likely to be shared between close regions [5]. However, dengue is

also influenced by human movement which creates links between distant regions that a distance-based spatial connectivity assumption would not capture [6,7]. Long-distance connections are particularly important when studying (re-)emerging diseases which are largely driven by connections between areas experiencing active disease transmission and disease-free areas [8–10]. In these examples, multiple random terms would be required within a Bayesian hierarchical model to capture the different sources of connectivity and quantify the relative importance of each to the disease transmission process.

In this paper, we present a Bayesian hierarchical modelling framework that uses penalized smoothing splines as a flexible method for structuring spatial model components. Smoothing splines use data to inform spatial components, given smoothing assumptions, rather than requiring the full specification of the spatial structure prior to model fitting [11,12]. The result is a non-stationary structure which is setting-specific and requires minimal user assumptions. This approach allows multiple spatially structured random components to be incorporated into the same model and can distinguish between these structures to quantify their relative contribution to the overall spatial structure. Although this study focuses on disease mapping models of count data, we also show that this method can be used for models of binary data.

2. Modelling approach

2.1. Disease mapping

Disease mapping is an important statistical tool used in epidemiology to explore spatial variation in disease incidence rates. Disease mapping models can generate and test hypotheses about associations between disease and a variety of potential explanatory variables, such as environmental and socio-economic factors [2,13]. Typically, disease counts, y_i ($i = 1, \ldots, n$), are collected across a study area separated into n contiguous areas. These counts are combined with an offset $\log(\xi_i)$ describing the underlying population at risk in each area i. For instance, y_i/ξ_i is the empirical incidence rate in i when ξ_i is population count. Where a disease is rare or areas within the study are small, estimates of the incidence are highly uncertain and thus unstable and inflated. To overcome this issue, Bayesian (hierarchical) modelling approaches have been developed to allow information from connected regions to be included in the rate estimation using random effects (data pooling). Conventionally, these models take the form

$$y_i \sim p(E(y_i), \psi)$$

and

$$\log(E(y_i)) = \log(\xi_i) + \alpha + S_i, \tag{2.1}$$

where p is a suitable count distribution (e.g. Poisson, negative binomial), $E(y_i)$ is the expected count, α is the intercept or baseline risk, S_i are spatially structured random components and ψ are hyperparameters of the distribution. The definition of S_i (which describes the spatial structure of $E(y_i)$ on the log scale, after correcting for ξ_i) depends on the disease of interest and the assumed spatial structure in the data. A recent systematic review found that spatial statistical models used to study mosquito-borne diseases only considered distance-based connectivity when defining the structure of such spatial random effects [4]. The most common spatial structure assumed connectivity between regions if and only if they share a border using a conditional autoregressive (CAR) model

$$S_i | S_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} W_{ij} S_j}{\sum_{j \neq i} W_{ij}}, \frac{\sigma_s^2}{\sum_{j \neq i} W_{ij}}\right), \tag{2.2}$$

where W_{ij} are proximity weights, often defined as $W_{ij} = 1$ if i and j share a border, and 0 otherwise. Although the conditional independence assumption intrinsic to neighbourhood-based spatial structures allows for efficient Bayesian computation [14], the nature of spatial connectivity is likely to be more complex and differ across settings. A smooth function with a structure defined using the data rather than prior to model fitting provides a flexible alternative and allows spatial dependency structures to be specific to each setting.

2.2. Penalized smoothing splines

Smoothing splines, or smooth functions, are used in generalized additive models to explore nonlinear relationships between a response variable and one or more covariate(s). Smoothing splines are constructed as a linear combination of basis functions, b_j (functions applied to the covariate(s) at given intervals, determined by the type of smoothing spline chosen), multiplied by regression coefficients, β_j [11]. For example,

$$f(x) = \sum_{j=1}^{K} \beta_j b_j(x). \tag{2.3}$$

Where f is a smooth function (the smoothing spline), x is the covariate of interest and K is the number of 'knots', or turning points, in the smooth function. The number of knots should be chosen to be large enough that the smooth function adequately describes the data, but not so large that they overfit or become 'overly wiggly'. To achieve this, a smoothing penalty parameter, λ, is introduced and estimated using the data to avoid overfitting when K is too large (e.g. as $\lambda \to \infty$, $f(x)$ becomes linear) [12].

Regression coefficients $\boldsymbol{\beta}$ are estimated using restricted maximum likelihood, which imposes a smoothing penalty on the coefficients of the form

$$\lambda \boldsymbol{\beta}^T P \boldsymbol{\beta}, \tag{2.4}$$

where λ is the penalty parameter introduced earlier and P is a penalty matrix computed prior to model fitting (based on the type of smoothing spline chosen) [11,12]. The penalty parameter, matrix and basis functions can be estimated efficiently using the mgcv package [15]. Although the mgcv package uses empirical methods to estimate the parameters defining smoothing splines, the results can be interpreted from a Bayesian perspective.

2.3. Bayesian interpretation of penalized smoothing splines

The assumption that smoothing functions f are more smooth than wiggly can be considered a prior belief on the values that the coefficients can take. This prior can be formalized and incorporated into Bayesian inference by assuming the regression coefficients β have the prior distribution

$$\beta \sim N\left(0, \frac{P^-}{\lambda}\right), \tag{2.5}$$

where P^-/λ is the covariance matrix [11,12]. However, the precision matrix $P\lambda$ is rank-deficient so is instead replaced by

$P_0\lambda_0 + P_1\lambda_1$, where the first term relates to a penalty on the null space of the smooth function and the second is the wiggliness penalty [16]. The interpretation of this is that the penalty matrix is separated into penalized components through P_1 (relating to wiggly behaviour) and non-penalized components through P_0. The splines $b_j(x)$ and penalty matrices can be efficiently generated using the jagam function in the mgcv package [16]. The definition of smoothing splines as linear combinations of (known) basis functions and (unknown) coefficients means that they can be entered into hierarchical models [17] and implemented using Bayesian inferential methods such as Markov chain Monte Carlo (MCMC). Under these conditions, the resulting penalized smoothing splines can be interpreted as random effects [11,18].

2.4. Spatial smoothing splines within Bayesian hierarchical models

In this study, we applied penalized smoothing splines to coordinates describing the relative 'connectivity' of regions (e.g. coordinates of the centroid of regions). This created two-dimensional smooth surfaces describing spatial patterns in the data. Thin plate regression splines are relatively efficient at estimating smooths over multiple variables and do not require a surface to be stationary. In addition, thin plate regression splines have low posterior correlation between parameters, which improves mixing when using MCMC methods [19,20]. If a coordinate system does not currently exist that describes the connectivity in question, this can be created from a symmetric continuous measure using multi-dimensional scaling (MDS). MDS translates a continuous measure of 'distance' or connectivity between observations onto an abstract Cartesian space and returns a set of coordinates [21]. For example, when connectivity is assumed to arise due to human movement, this could be defined as a continuous measure such as the number of air travel passengers, or an estimate from a movement model, such as a gravity or a radiation model [22,23], which assumes the number of people moving between areas is a function of population and distance. Note that MDS requires the measure of connectivity to be symmetric, for example, the number of people travelling to an area is assumed to be equal to the number returning.

Smooth surfaces were defined using splines and included in Bayesian hierarchical models of count data using the procedures detailed above. Models were implemented using NIMBLE [24,25], a flexible program that implements Bayesian models created in the BUGS language using MCMC methods within R [26]. The flexibility of this framework means that multiple spatially smooth surfaces can be included in the same model with different connectivity assumptions (e.g. distance-based and human movement). Interpreting the smooth surfaces over the various connectivity measures as random means the relative contribution of each spatial structure can be quantified by calculating the proportion of the overall variance of the random terms that is captured by each spatial term.

3. Simulation study 1: a single source of spatial structure

In this section, we present a simulation study in which we apply Bayesian spatial models to data generated from a distance-based spatial structure. We compare model performance between the penalized regression spline approach and a neighbourhood-based CAR model. A further simulation study assuming a single source of human movement-based connectivity is presented in the electronic supplementary material.

3.1. Data generation

Fictitious disease count data were generated from a Poisson distribution for each of the 1013 municipalities in South Brazil, the region used in the case study (§5), from model (2.1). The log of the population divided by 100 000, $\log(\xi_i)$, was included as an offset (electronic supplementary material, figure S1). The population of each municipality was taken from the Brazilian census and described in §5.1. The intercept term α was set to zero, while the term S_i was defined by

$$S_i = \sqrt{\phi} \cdot sm(x_i, z_i) + \sqrt{(1-\phi)} \cdot \varepsilon_i, \tag{3.1}$$

where ϕ is a mixing parameter, taking values between 0 and 1, which measures the contribution of each term (if we interpret $sm(x_i, z_i)$ as random and independent of ε_i) to the overall variance of S_i, and $\varepsilon_i \sim N(0, 1)$. $sm(x_i, z_i)$ is a continuous function applied to connectivity coordinates (x_i, z_i) to emulate a spatially structured surface (figure 1*b*, taken from [27]):

$$sm(x,z) = \pi\sigma_x\sigma_z(1.2\mathrm{e}^{-(x-0.2)^2/\sigma_x^2-(z-0.3)^2/\sigma_z^2} + 0.8\mathrm{e}^{-(x-0.7)^2/\sigma_x^2-(z-0.8)^2/\sigma_z^2}) \tag{3.2}$$

and

$$\sigma_x = 0.3, \quad \sigma_z = 0.4.$$

To create a distance-based spatial structure, the smooth function sm was applied to coordinates of the centroid of municipalities which were scaled to take values between 0 and 1. The function $sm(x_i, z_i)$ was centred at 0 by subtracting the overall mean from each value. Eleven simulated datasets were produced using equation (3.1), setting values of ϕ between 0 and 1 at intervals of 0.1 (figure 1).

3.2. Modelling approach

Two Poisson models containing spatially structured and unstructured random components were applied to each simulated dataset

$$y_i \sim \text{Poisson}(E(y_i))$$
$$\log(E(y_i)) = \log(\xi_i) + \alpha + u_i + v_i \tag{3.3}$$
$$\log(E(y_i)) = \log(\xi_i) + \alpha + \frac{1}{\tau}\left(\sqrt{\phi} \cdot u_{*i} + \sqrt{1-\phi} \cdot v_{*i}\right). \tag{3.4}$$

In model (3.3), u_i is a spatially structured term, constructed using a thin plate regression spline on the coordinates of the centroid of each municipality, and v_i is a spatially unstructured term, assumed to follow a zero-mean normal distribution, representing heterogeneity between regions. This spatially smooth model was compared with a more conventional random effect approach based on the BYM2 model (model (3.4)), which is often used to capture spatial structure in disease mapping [3,28,29]. In model (3.4), u_{*i} are spatially structured random effects assuming a CAR model with a binary neighbourhood matrix (see equation (2.2)), v_{*i} are unstructured normal random effects, and ϕ is a mixing parameter, measuring the contribution of each random effect to the marginal variance $(1/\tau^2)$ of the overall random effect [3,28]. Here, $\phi = 1$ represents

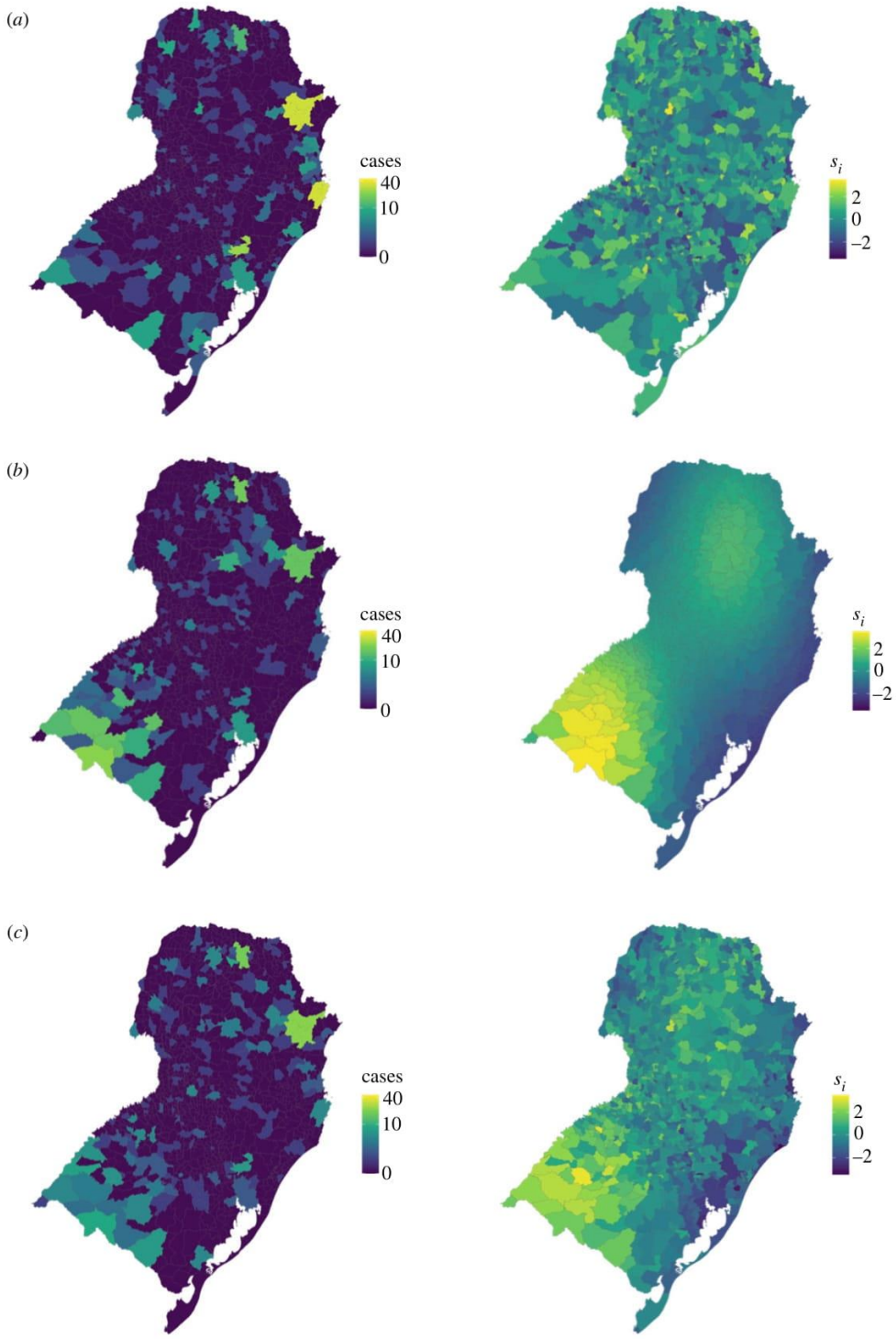


Figure 1. Simulated disease counts (left) and spatial random effects (right) under a distance-based structure using different spatial structure combinations. The number of cases simulated from a Poisson model and the underlying spatial structure where the data has (*a*) no spatial structure ($\phi = 0$), (*b*) a distance-based structure only ($\phi = 1$) and (*c*) equal contribution of both structures ($\phi = 0.5$).

a purely spatial model, equivalent to an intrinsic CAR model [30], and $\phi = 0$ indicates no spatial structure in the data. Spatially smooth models were fitted using MCMC simulations in R via the NIMBLE package [24]. Although the BYM2 model can be formulated and fitted using MCMC simulations [31], we found that most contemporary disease mapping studies use integrated nested Laplace approximations (INLA) for model fitting [32]. INLA is an approximate Bayesian inference approach which provides a more efficient alternative to MCMC and avoids issues with convergence [14,29]. We compared the spatially smooth model with a BYM2 model fitted using INLA to ensure we were comparing our results to the conventional approach. However, to ensure any differences were not a result of inferential methods, the BYM2 random

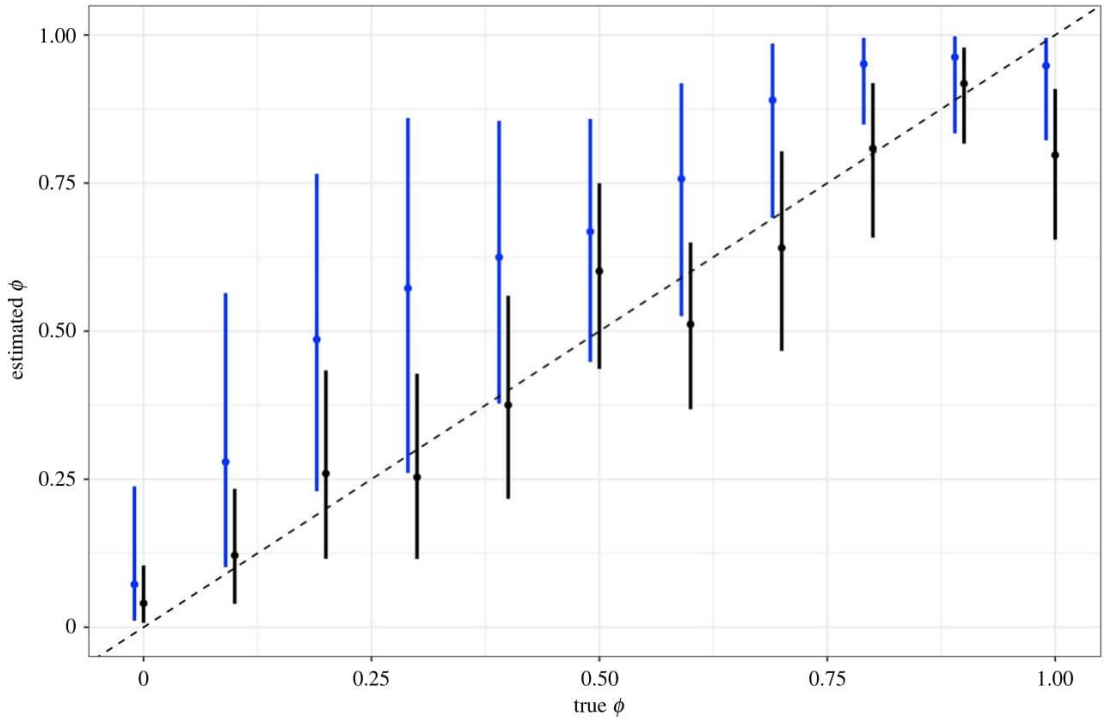


Figure 2. The mean and 95% credible interval of estimated ϕ values extracted from models including a smoothing spline (black) and BYM2 (blue) compared with the known value (dashed line). Estimated ϕ values for the smoothing spline model were calculated using the proportion of the random effect variance explained by the spatially structured term and were extracted from INLA output for the BYM2 model.

effect model was also fitted using MCMC simulations in NIMBLE and compared with the spatially smooth model. Results of this comparison are presented in the electronic supplementary material.

Model comparison was based on mean absolute error (MAE) and Watanabe–Akaike information criterion (WAIC), an information criterion used to assess the predictive accuracy of Bayesian models [33]. Lower values of MAE and WAIC are preferred. The relative contribution of the spatially structured term, u_i, to the overall random terms in the spatially smooth model was defined as the proportion of the overall random term variance explained by u ($\text{var}(u)/\text{var}(u+v)$). This was estimated using samples from the posterior distribution of u and v. We compared estimates of the ϕ hyperparameter from INLA, the relative contribution of u_i with the random effect variance from NIMBLE, and the known proportion of spatial variance used in the simulation. All analyses were carried out using R v. 4.1.1 [26]. The code used to simulate data and perform analyses is available here: https://doi.org/10.5281/zenodo.7054457 [34].

3.3. Results

We found that the spatial spline model estimates were closer to the true value of ϕ than the BYM2 model for most simulations (figure 2 and table 1), and that INLA's estimates of this parameter were not always consistent with the true value. This indicates that the spatial spline models were able to identify and quantify the relative contribution of this spatial structure within the data as well as (if not better than) INLA's BYM2 models.

MAEs and WAIC values show that model performance was similar between the smoothing spline and BYM2 models (table 1). The WAIC showed the smoothing spline model performed slightly better on all simulated datasets apart from one, although the MAE preferred the BYM2 models. When these approaches were compared with the BYM2 model fitted using MCMC (electronic supplementary material, S1), we found that some of these differences appear to be a result of fitting the model using INLA rather than model formulation itself. However, the objective of this comparison was not to show that the proposed smooth model outperforms these approaches, rather that it performs as well as the current standard. These results illustrate that the smoothing spline was able to detect spatial connectivity between neighbouring regions while being flexible enough to capture alternative structures. The 95% credible interval (CI) of the intercept coefficient estimate contained the true value 0 for all models for both approaches (electronic supplementary material, figure S2).

4. Simulation study 2: two sources of spatial structure

In this section, we present another simulation study in which we apply Bayesian spatial models to data generated with two sources of spatial connectivity: distance-based and human movement-based.

4.1. Data generation

An extension of the spatial term, S_i, in equation (3.1) was used to generate data with spatial connectivity arising from two different sources

$$S_i = \sqrt{\phi_1} \cdot sm(a_i, b_i) + \sqrt{\phi_2} \cdot sm(c_i, d_i) + \sqrt{\phi_3} \cdot \varepsilon_i \qquad (4.1)$$

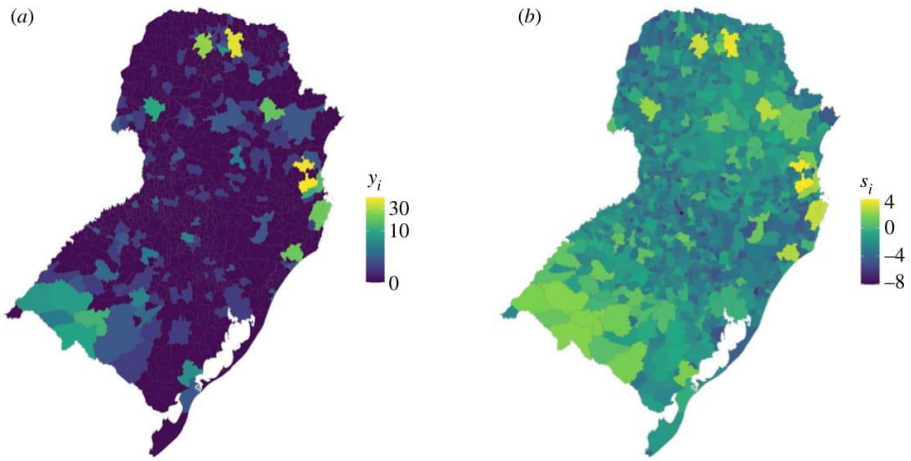


Figure 3. Simulated data containing two sources of spatial structure. Simulated disease counts, y_i (*a*) and spatial random terms, s_i (*b*), for South Brazil generated using equation (4.1), where $\phi_1 = 0.4$, $\phi_2 = 0.5$ and $\phi_3 = 0.1$.

Table 1. Model comparison statistics and mean estimates of the mixing parameter, ϕ, from the smoothing spline and INLA BYM2 models. Mean absolute error (MAE) and WAIC calculated for the spatial spline and BYM2 models for each simulated dataset. The lowest MAE and WAIC, and the ϕ estimate closest to the value used in each simulation are highlighted in italics.

	smoothing spline model			INLA BYM2 model		
ϕ	MAE	WAIC	ϕ estimate	MAE	WAIC	ϕ estimate
0	1.51	*996.94*	*0.041*	*1.04*	1005.79	0.072
0.1	1.54	*1030.64*	*0.121*	*1.11*	1034.29	0.279
0.2	1.33	932.42	*0.26*	*1.04*	*931.79*	0.486
0.3	1.27	*909.42*	*0.253*	*0.93*	912.5	0.572
0.4	1.39	*961.67*	*0.375*	*1.08*	976.12	0.625
0.5	1.54	*935.09*	*0.601*	*1.21*	954.34	0.668
0.6	1.5	*881.09*	*0.512*	*1.13*	973.61	0.757
0.7	1.45	*931.85*	*0.641*	*1.17*	989.24	0.89
0.8	1.63	*947.51*	*0.808*	*1.37*	983.96	0.951
0.9	1.59	*876.37*	*0.918*	*1.37*	922.29	0.963
1	1.48	*875.42*	0.797	*1.25*	924.14	*0.948*

and

$$\phi_1 + \phi_2 + \phi_3 = 1.$$

Where *sm* is a smooth function (equation (3.2)), applied to coordinates describing distance-based connectivity (a_i, b_i), and human movement-based connectivity (c_i, d_i). The coordinates of the centroid of municipalities were scaled to take values between 0 and 1 and used to describe distance-based connectivity (a_i, b_i). As a coordinate system describing connectivity arising from human movement does not exist, we applied MDS to an estimate of the number of people moving between municipalities, generated using a movement model described in the electronic supplementary material, to create coordinates c_i and d_i (electronic supplementary material, figure S3).

In this example, we used three scaling parameters, ϕ_1, ϕ_2 and ϕ_3, to describe the relative contribution of each random term to the marginal variance. We held ϕ_3 constant at 0.1, with ϕ_1 and ϕ_2 taking values between 0 and 0.9 at intervals of 0.1, creating 10 simulated datasets (figure 3).

4.2. Modelling approach

We applied a Poisson spatial model to each simulated dataset which contained three random terms

$$y_i \sim \text{Poisson}(E(y_i))$$
$$\log(E(y_i)) = \log(\xi_i) + \alpha + u_{1,i} + u_{2,i} + v_i. \qquad (4.2)$$

Where $u_{1,i}$ is constructed using a thin plate regression spline applied to coordinates of the centroids of municipalities, and $u_{2,i}$ is structured using a thin plate regression spline applied to human movement-based connectivity coordinates described previously. v_i is assumed to have no spatial structure and represents unobserved heterogeneity between municipalities.

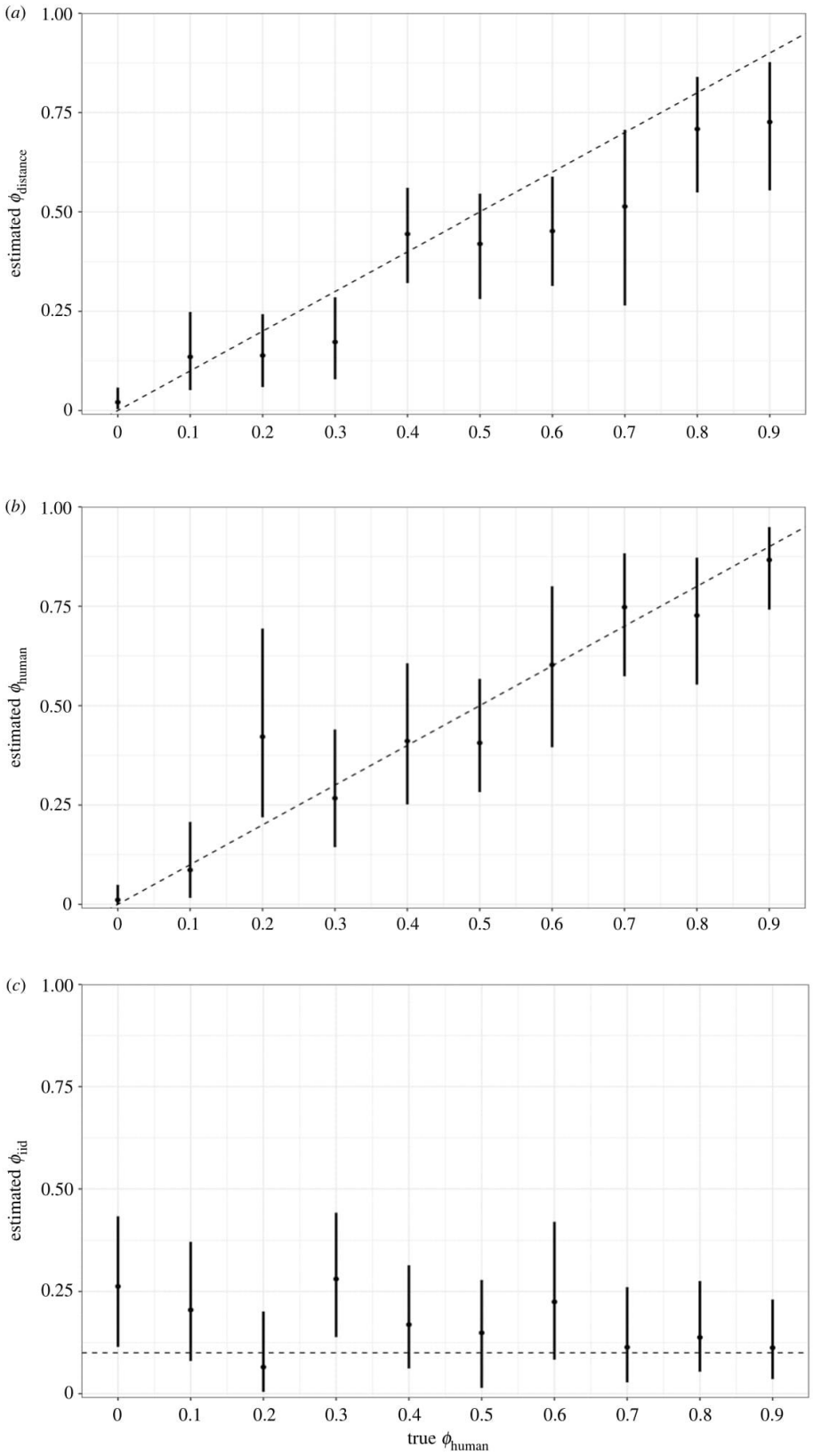


Figure 4. Mean and 95% credible interval of the proportion of variance of the random effects explained by (*a*) the distance-based structured term, (*b*) the human movement-based structured term and (*c*) unstructured random term. Dashed lines represent the true value from simulations.

We compared the proportion of the marginal variance explained by each random term and compared these with the known ϕ values used in data generation.

4.3. Results

We found that the models were able to accurately estimate the intercept coefficient value of 0 across most simulated datasets (electronic supplementary material, figure S4). Estimates of the relative contribution of each random term to the overall spatial structure were close to ϕ values used in simulations and were able to detect the increasing contributions of distance-based and human movement-based terms as the true value increased (figure 4).

5. Case study

This case study uses the Bayesian spatially smooth models introduced in previous sections to map the spatial patterns of dengue incidence in South Brazil between 2001 and 2020.

5.1. Data description

We obtained annual notified dengue cases for each of South Brazil's 1013 municipalities between 2001 and 2020 from Brazil's Notifiable Diseases Information System, freely available via the Health Information Department, DATASUS (https://datasus.saude.gov.br/informacoes-de-saude-tabnet/). To explore the pattern of disease over the whole period, we took the average annual number of cases over the period and rounded this to the nearest whole number. The annual population for each municipality was obtained from the Brazilian Institute of Statistics and Geography (IBGE) via DATASUS (http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/poptbr.def) over the same period and aggregated in the same way. We used the population divided by 100 000 as an offset to model the dengue incidence rate (DIR), a measure used by the Brazilian Ministry of Health to monitor dengue outbreaks. South Brazil was previously thought to be protected from dengue due to its temperate climate, with winter temperatures too low for the primary vector, *Aedes aegypti*, to breed and transmit the disease. However, recent studies have shown that the northern part of the South region now experiences outbreaks, thought to be due to increasing temperatures (figure 5, [35]). The data show a clear distance-based spatial pattern in this region. However, studies of other temperate regions of South America, such as Argentina, have hypothesized that increased outbreaks in cooler regions may be a result of human movement into previously protected cities [7,36]. Data used in this case study are available from https://doi.org/10.5281/zenodo.7054457 [34].

5.2. Modelling approach

We applied a negative binomial model to the average annual dengue cases, using the log of the population divided by 100 000 as an offset to explore the DIR in South Brazil. A negative binomial distribution was assumed to account for possible overdispersion in the dengue case count [5]. Model (4.2) was applied to the data, spatial random terms were structured by applying thin plate regression splines to the coordinates of the centroids of municipalities ($u_{1,i}$, assuming distance-based connectivity), and human movement-based connectivity coordinates described in §4 and the electronic supplementary material ($u_{2,i}$).

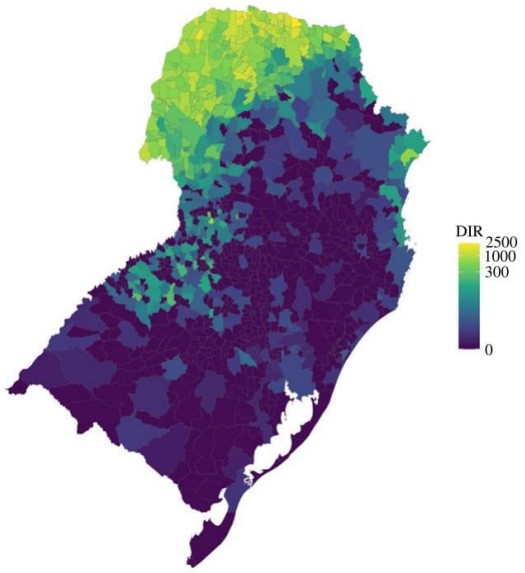


Figure 5. Average dengue incidence rate (DIR), 2001–2020 in South Brazil. The mean annual dengue incidence rate per 100 000 residents in South Brazil from 2001 to 2020. Data are shown on a log scale.

5.3. Results

The model found that human movement did not account for much of the spatial structure of the data in this region (ϕ_2 = 0.003, 95% CI: 0, 0.012), and most of the variation could be attributed to the distance-based random term (ϕ_1 =0.85, 95% CI: 0.823, 0.876, figure 6). The human movement data used to create these random effects were only able to capture movement between cities in South Brazil. However, outbreaks in temperate regions such as this are likely to be triggered by the movement of people from endemic regions elsewhere in Brazil into the South [7].

Estimates of each random term and the combined total were extracted and plotted to generate hypotheses about these patterns (figure 7). Most of the spatial structure came from the distance-based random term, which shows the highest risk was in the northwest and that the risk decreased to the south. This area of increased risk is the same region which was found to have an increase in the number of months per year with temperatures suitable for dengue transmission since 2010 in a previous study [35]. This model could be extended to include temperature and other variables known to influence dengue risk.

6. Discussion

In this paper, we have shown that penalized smoothing splines present a flexible alternative to CAR-based structures of spatial random effects that allow multiple sources of spatial connectivity to be considered within the same model. Smoothing splines allow the spatial structure to be derived from data as part of the model fitting process, producing a non-stationary spatial surface specific to the data being considered. This smooth surface can be extracted and plotted

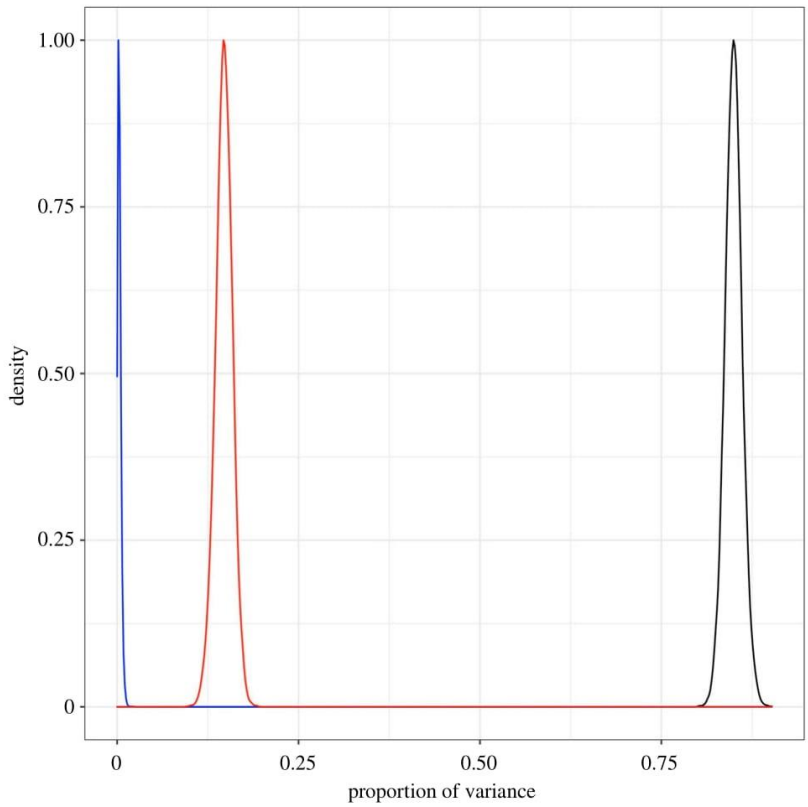


Figure 6. Estimates of the proportion of variance explained by distance-based (black), human movement-based (blue) and independent (red) random terms. Using simulations extracted from NIMBLE, the variance of each random term was calculated and divided by the variance of the combined random component, giving the relative contribution of each structure.

to generate hypotheses about the reasons for this spatial connectivity which may help identify potential drivers of disease. Although many disease mapping studies assume a distance-based structure of connectivity, the smooth spline approach used here can be applied to any symmetric continuous measure of connectivity, including human movement. Another benefit of the smoothing spline approach is that the model structure can be extended to include multiple sources of spatial connectivity and can produce parameters quantifying the relative contribution of each structure to the underlying variance of the data. Although this study has focused on disease mapping models of count data, we have shown this method is compatible with other models, such as logistic models for binary data (see the electronic supplementary material).

Formulating models in NIMBLE (or other similar coding languages) and implementing them using MCMC methods allows for flexibility and complexity in the model structure. However, these models are more likely to face issues with convergence than approximate methods such as INLA [14]. MCMC methods may also take longer than INLA to fit models if convergence is an issue, although this is not always the case when using NIMBLE [37].

One of the main benefits of using penalized smoothing splines over CAR-based priors is that they can be applied to any symmetric continuous measure of connectivity. However, the most appropriate measure may not always be clear or available. For example, human movement-based connectivity can be captured using data to describe regular, short-distant movement such as commuting within a city, or long-distance, long-term movement such as migration, which requires different assumptions [9]. Mobile phone data have potential to describe short-term movements at small spatial scales but may be difficult to obtain, and care must be taken in some settings where bias may arise [38]. Movement models, such as gravity and radiation models, assume that the number of people moving between areas can be described as a function of population and distance [22]. Movement models provide an alternative when data is unavailable or inappropriate and have been shown to replicate patterns of movement in large cities and European countries [23,39]. However, care must be taken when parametrizing these models, particularly in rural settings [40]. Although distance is recognized as an important driver of human movement [22], our simulation studies showed that this approach can distinguish between the relative contribution of both sources of connectivity to the overall spatial structure (see §4 and electronic supplementary material, S4).

One limitation of this method is that the measure of connectivity must be symmetric to produce a spatially smooth surface. This is often not realistic when considering human movement, as the number of people moving from smaller to larger cities is often different to those moving in the opposite direction [41]. In the examples presented in this study, we assumed that the number of people travelling between municipalities is equal to the number of people returning. Also, the models presented in this study only consider a single time point (or data summarized over a given time period); however, disease risk is likely to vary over time and models

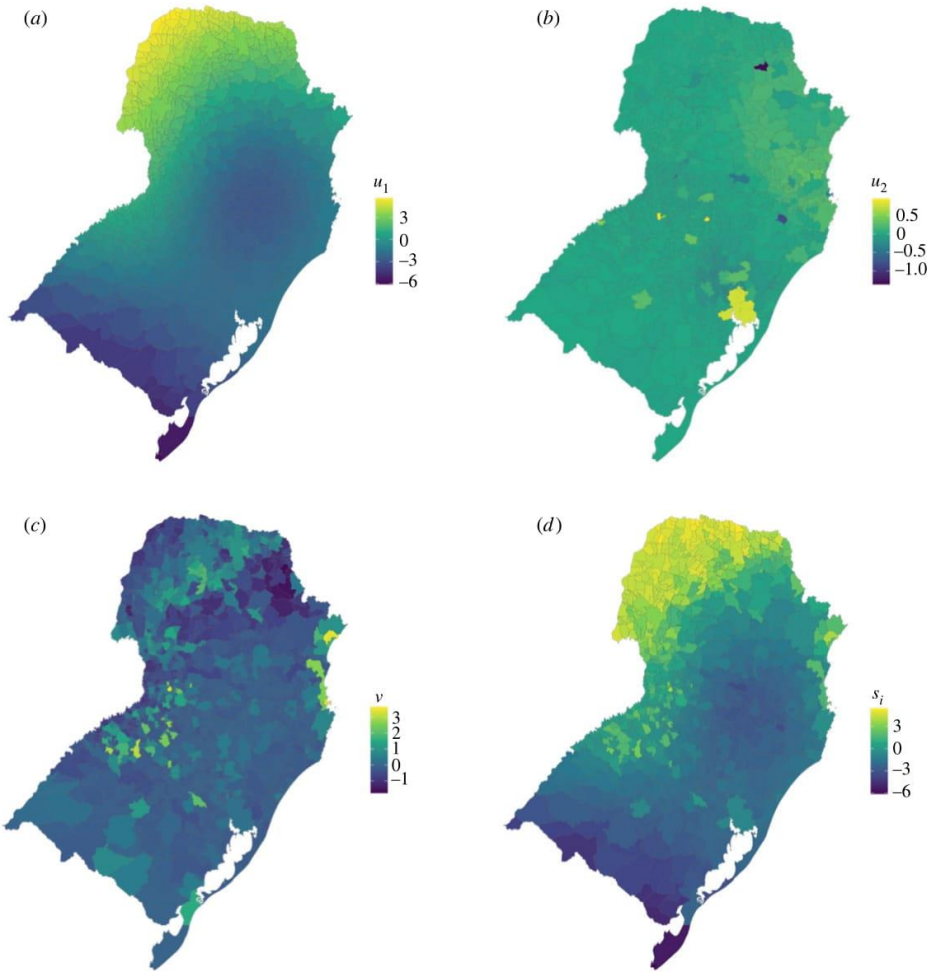


Figure 7. Mean estimates of the (*a*) distance-based, (*b*) human movement-based, (*c*) unstructured and (*d*) combined random terms.

may be required to account for inter-annual or seasonal variation. Data presented in the South Brazil study have been used elsewhere to show the expansion of dengue outbreaks into the region and the changes in spatial structure over the past 20 years [35,42]. The models presented here can be extended to include temporal covariates or random terms to account for seasonal and annual trends, and changing spatial connectivity surfaces to reflect changing patterns of movement. Tensor smooth functions, a type of smoothing spline which allows interaction between variables measured on different scales [27], may be incorporated to explore the interactions between time and connectivity. These structures can be explored to understand changing dynamics of diseases and generate hypotheses about drivers of change or highlight areas at risk. Covariates such as climate indicators can also be included into the models and random term estimates compared to highlight the relative variability in the disease risk explained by these covariates.

Penalized smoothing splines present a flexible alternative to conventional random effect structures when constructing Bayesian hierarchical models. They require minimal user assumptions beyond smoothness and can be applied to any symmetric continuous measure of connectivity. By taking a Bayesian view of these smoothing splines, we can incorporate multiple sources of spatial connectivity into a complex modelling framework efficiently and quantify their relative contribution to the overall spatial structure of the data. This is particularly useful in infectious disease epidemiology where the drivers of transmission may be complicated and not fully understood.

Data accessibility. All data used in this study are open access and available freely on the internet; see the methods section for more details. Data and code used to produce this analysis is available from Zenodo (https://doi.org/10.5281/zenodo.7054457) [34].

The data are provided in the electronic supplementary material [43].

Authors' contributions. S.A.L.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; T.E.: conceptualization, methodology, software, supervision, writing—review and editing; R.L.: conceptualization, methodology, software, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. S.A.L. was supported by a Royal Society Research Grant for Research Fellows. T.E. was funded by the European Union's Horizon 2020 research and innovation programme under Grant agreement no. 856612 and the Cyprus Government. R.L. was supported by a Royal Society Dorothy Hodgkin Fellowship.

References

1. Wakefield J. 2007 Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183. (doi:10.1093/biostatistics/kxl008)
2. Lee D. 2011 A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat. Spatio Temporal Epidemiol.* **2**, 79–89. (doi:10.1016/j.sste.2011.03.001)
3. Riebler A, Sørbye SH, Simpson D, Rue H. 2016 An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25**, 1145–1165. (doi:10.1177/0962280216660421)
4. Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021 Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**, 20210096. (doi:10.1098/rsif.2021.0096)
5. Lowe R *et al*. 2021 Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *Lancet Planet. Health* **5**, e209–e219. (doi:10.1016/S2542-5196(20)30292-8)
6. Lana RM, Gomes MF da C, de Lima TFM, Honório NA, Codeço CT. 2017 The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: a network-based analysis. *PLoS Negl. Trop. Dis.* **11**, e0006070. (doi:10.1371/journal.pntd.0006070)
7. Robert MA, Tinunin DT, Benitez EM, Ludueña-Almeida FF, Romero M, Stewart-Ibarra AM, Estallo EL. 2019 Arbovirus emergence in the temperate city of Córdoba, Argentina, 2009–2018. *Sci. Data* **6**, 276. (doi:10.1038/s41597-019-0295-z)
8. Kraemer MU *et al*. 2019 Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 1–11. (doi:10.1038/s41598-019-41192-3)
9. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Diseases* **3**, e481. (doi:10.1371/journal.pntd.0000481)
10. Findlater A, Bogoch II. 2018 Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends Parasitol.* **34**, 772–783. (doi:10.1016/j.pt.2018.07.004)
11. Wood SN. 2017 *Generalized additive models: an introduction with R*. New York, NY: CRC press.
12. Wood SN. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **73**, 3–36. (doi:10.1111/j.1467-9868.2010.00749.x)
13. MacNab YC. 2022 Bayesian disease mapping: past, present, and future. *Spat. Stat.* **50**, 100593. (doi:10.1016/j.spasta.2022.100593)
14. Rue H, Martino S, Chopin N. 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71**, 319–392. (doi:10.1111/j.1467-9868.2008.00700.x)
15. Wood S. 2015 Package 'mgcv'. R Package Version 1, 729.
16. Wood SN. 2016 Just another Gibbs additive modeler: interfacing JAGS and mgcv. *J. Stat. Softw.* **75**, 1–15. (doi:10.18637/jss.v075.i07)
17. Pedersen EJ, Miller DL, Simpson GL, Ross N. 2019 Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* **7**, e6876. (doi:10.7717/peerj.6876)
18. Wood SN. 2004 Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* **99**, 673–686. (doi:10.1198/016214504000000980)
19. Wood SN. 2003 Thin plate regression splines. *J. R. Stat. Soc. B* **65**, 95–114. (doi:10.1111/1467-9868.00374)
20. Crainiceanu C, Ruppert D, Wand MP. 2005 Bayesian analysis for penalized spline regression using WinBUGS. *J. Stat. Softw* **14** , 1–24. See https://ro.uow.edu.au/eispapers/2517.
21. Cox MA, Cox TF. 2008 Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347. Berlin, Germany: Springer.
22. Simini F, González MC, Maritan A, Barabási A-L. 2012 A universal model for mobility and migration patterns. *Nature* **484**, 96–100. (doi:10.1038/nature10856)
23. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V. 2014 On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716. (doi:10.1371/journal.pcbi.1003716)
24. De Valpine P *et al*. 2021 Nimble: MCMC, particle filtering, and programmable hierarchical modeling. *R Package Version 011* **1**.
25. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. 2017 Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.* **26**, 403–413. (doi:10.1080/10618600.2016.1172487)
26. R Core Team. 2021 *R: A language and environment for statistical computing. 4.1.1*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
27. Wood SN. 2006 Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**, 1025–1036. (doi:10.1111/j.1541-0420.2006.00574.x)
28. Simpson D, Rue H, Riebler AI, Martins TG, Sørbye SH. 2017 Penalising model component complexity: a principled, practical approach to constructing priors. *Stat. Sci.* **32**, 1–28. (doi:10.1214/16-STS576)
29. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. 2017 Bayesian computing with INLA: a review. *Annu. Rev. Stat. Appl.* **4**, 395–421. (doi:10.1146/annurev-statistics-060116-054045)
30. Besag J. 1974 Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B* **36**, 192–225. (doi:10.1111/j.2517-6161.1974.tb00999.x)
31. Lawson AB. 2020 NIMBLE for Bayesian disease mapping. *Spat. Spatio Temporal Epidemiol.* **33**, 100323. (doi:10.1016/j.sste.2020.100323)
32. Bakka H, Rue H, Fuglstad G, Riebler A, Bolin D, Illian J, Krainski E, Simpson D, Lindgren F. 2018 Spatial modeling with R-INLA: a review. *Wiley Interdiscip. Rev. Comput. Stat.* **10**, e1443. (doi:10.1002/wics.1443)
33. Gelman A, Hwang J, Vehtari A. 2014 Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016. (doi:10.1007/s11222-013-9416-2)
34. Lee SA *et al*. 2022 Data and R code to accompany 'A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping' (version v1.0.0). Zenodo. (doi:10.5281/zenodo.7054457)
35. Lee SA, Economou T, Catão R de C, Barcellos C, Lowe R. 2021 The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. *PLoS Negl. Trop. Dis.* **15**, e0009773. (doi:10.1371/journal.pntd.0009773)
36. Robert MA, Stewart-Ibarra AM, Estallo EL. 2020 Climate change and viral emergence: evidence from *Aedes*-borne arboviruses. *Curr. Opin. Virol.* **40**, 41–47. (doi:10.1016/j.coviro.2020.05.001)
37. Lawson AB. 2021 *Using R for Bayesian spatial and spatio-temporal health modeling*. New York, NY: CRC Press.
38. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. 2016 Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *J. Infect. Dis.* **214**, S414–S420. (doi:10.1093/infdis/jiw273)
39. Kraemer MUG *et al*. 2018 Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan. *EPJ Data Sci.* **7**, 17. (doi:10.1140/epjds/s13688-018-0144-x)
40. Meredith HR *et al*. 2021 Characterizing human mobility patterns in rural settings of sub-Saharan Africa. *eLife* **10**, e68441. (doi:10.7554/eLife.68441)
41. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106**, 21 484–21 489. (doi:10.1073/pnas.0906910106)
42. Codeco CT *et al*. 2022 Fast expansion of dengue in Brazil. *Lancet Reg. Health - Am.* **12**, 100274. (doi:10.1016/j.lana.2022.100274)
43. Lee SA, Economou T, Lowe R. 2022 Data from: A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping. Figshare. (doi:10.6084/m9.figshare.c.6186135)

5. The contribution of human movement to dengue expansion differs between regions in Brazil

5.1 Bridging section

The overall aim of this thesis was to understand the contribution of increasing temperatures in South Brazil and connectivity between cities arising from human movement to the expansion of the dengue transmission zone in Brazil. Model results from Chapter 2 showed that the odds of a dengue outbreak were significantly increased in municipalities with year-round temperatures suitable for dengue transmission, and that temperature suitability explained most spatiotemporal variation in dengue outbreaks in South Brazil. The model also found that the level of influence of cities was significantly associated with the odds of a dengue outbreak. However, this relationship was nonlinear and cities classified as regional capitals were found to have a higher odds of a dengue outbreak compared to the most influential cities in Brazil, metropoles. Although the level of influence indicator was used as a proxy for human movement, the model presented in Chapter 2 was not able to include direct connections between cities in Brazil arising due to human movement. Chapter 4 presented a statistical modelling framework that could incorporate explicit links between areas arising due to human movement and quantify the relative contribution of human movement to the overall spatial structure of the data.

In this chapter, I applied the statistical modelling framework presented in Chapter 4 to model the number of dengue outbreaks between 2001 – 2020 per municipality in Brazil, with spatially structured terms designed to capture spatial connectivity within the data. Simulations from this model were used to quantify the relative contribution of human movement based on regular commuting to the odds of a dengue outbreak across Brazil (Objective 4). The relative contribution of regular commuting to the spatial structure of dengue outbreaks was estimated for the whole of Brazil and each region in turn.

This paper is yet to be submitted to a journal. Supplementary materials referred to in this chapter can be found in Appendix E.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1806358	**Title**	Ms
First Name(s)	Sophie Alice		
Surname/Family Name	Lee		
Thesis Title	Spatial Modelling of Emerging Infectious Diseases: Quantifying the Role of Climate, Cities and Connectivity on Dengue Expansion in Brazil		
Primary Supervisor	Dr Rachel Lowe		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	eLife
Please list the paper's authors in the intended authorship order:	Sophie A. Lee, Theodoros Economou, Cláudia Torres Codeço, Raquel Martins Lana, Christovam Barcellos, Rafael de Catro Catão, Rachel Lowe

Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My contribution: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing – original draft, Writing – review & editing

SECTION E

Student Signature	Sophie Lee
Date	09/09/2022

Supervisor Signature	Rachel Lowe
Date	09/09/2022

Abstract

Dengue transmission has been expanding across Brazil since its re-introduction into areas that were previously thought to be protected. Previous studies have shown that geographical barriers to dengue transmission are being gradually eroded in South Brazil and the western Amazon. In this study, we quantify the relative contribution of regular commuting to the geographical expansion of dengue outbreaks between 2001 and 2020 in Brazil using a Bayesian hierarchical model. Spatially structured terms were included in the model and generated by applying penalised regression splines to coordinate systems that describe the relative 'connectedness' of municipalities. Spatial connectivity arising from regular commuting was described using coordinates generated from the 2010 census in Brazil, and connectivity between close regions due to unobserved shared characteristics e.g., climate type and socioeconomic conditions, were described using coordinates of the centroid of municipalities. We found that regular commuting contributed very little (1.3%) to the spatial structure of data when Brazil was considered as a whole. However, the relative contribution of commuting to dengue outbreaks was higher in the North and Northeast regions, particularly in the western Amazon. This supports previous findings that expansion in this region was a result of improved transportation infrastructure and increased human movement.

5.2 Introduction

Over the past 50 years, dengue has been expanding globally into previously unaffected areas. This has been attributed to climate change, urbanisation, and increased connectivity driven by human movement [1,2]. In Brazil, this expansion has taken place at an alarming rate since its re-emergence in the 1980s, resulting in an estimated 8.7 million new individuals at risk over the past 5 years [3]. One of the final frontiers of dengue transmission in Brazil was the western Amazon, a remote, mostly rural area with many communities reachable only by long boat journeys [3–5]. However, this barrier has been eroded over the past 20 years and there are now few municipalities in the area that have not experienced a dengue outbreak [4]. Evidence suggests that the introduction of dengue into the area was driven by increases in human movement to and within the region following improved road infrastructure and increased air travel [6]. Although dengue is typically an urban disease, rapid unplanned urbanisation has produced ideal habitats for the dengue vector, *Aedes* mosquitoes, to thrive. The humid, hot climate coupled with a lack of access to basic services, such as piped water, has seen dengue

vectors, and subsequently the viruses they transmit, move from larger urban centres into neighbouring peri-urban and rural areas within the Amazon [7].

A previous study found that the odds of a dengue outbreak was significantly higher in cities that were considered very connected within the Brazilian urban network compared to less influential centres based on their classification in the Regions of influence of cities ("Regiões de Influência das Cidades", REGIC) study [4,8]. However, this relationship was nonlinear and the cities considered most connected in Brazil, classified as metropoles, had lower odds of a dengue outbreak than those classified as regional capitals. This could be due to differences in infrastructure and health-seeking behaviour which could not be detected at the municipality level. Another potential reason for this nonlinear trend is that the level of influence indicator used in this study may not be an adequate proxy for connectivity arising due to human movement. The REGIC study found that small cities were connected to metropoles via less influential cities, such as regional capitals [8]. By not including the direct links between cities, we could be missing important connections within the data.

Spatial connectivity, including connections between close geographical areas due to unobserved shared characteristics and behaviours (referred to here as distance-based connectivity), and connections arising due to human movement between areas, is an important consideration when modelling infectious diseases. Most spatial models for mosquito-borne disease transmission assume spatial connectivity only exists between close areas, represented as a function of distance [9]. This distance-based connectivity is often used as a proxy to account for unobserved characteristics such as shared climatic and environmental factors, vector control measures, and levels of immunity within communities. However, these connections do not account for long-distance human movement, which has been identified as an important driver of infectious disease expansion [10–12]. In this study, we aim to understand how human movement around the Brazilian urban network has contributed to the expansion of dengue between 2001 – 2020. We include direct links between pairs of cities across Brazil arising from regular commuting into a Bayesian spatial model, which allows us to quantify the relative contribution of human movement on the expansion of dengue in Brazil [13]. A distance-based spatial connectivity structure is also included to account for unobserved shared characteristics, such as climate, between close areas. By comparing the relative contribution of regular commuting to the spatial structure of dengue outbreaks between regions of Brazil, we hope to better understand the recent changes to the dengue transmission zone.

5.3 Methods

5.3.1 Epidemiological data

Notified dengue case data was obtained from Brazil's Notifiable Diseases Information System (SINAN), freely available via the Health Information Department, DATASUS (https://datasus.saude.gov.br/informacoes-de-saude-tabnet/). This data is aggregated by month of first symptom (between January 2001 and December 2020) and the municipality of residence. Although there were 5,570 municipalities in 2020, these borders have changed over the period. To ensure municipalities were consistent over the entire period, we aggregated data to the 5,560 municipalities that were present in 2001 by combining the new municipalities' data with their 'parent municipality'. Data and code used to carry out the analysis are available from https://github.com/sophie-a-lee/dengue_human_movement_model.

5.3.2 Human movement data

The number of residents regularly commuting between municipalities for work or education was extracted from the 2010 Brazilian demographic census [14]. A random sample of residents were asked to give details about the country, state and municipality that they travel to for work or education. We excluded movements outside of Brazil and residents that did not provide any information about their destination (they responded that they leave the municipality for work or education but gave no further details about the destination). Some residents provided partial data, for example the state but not the municipality of their destination. These residents were assigned a destination proportionally based on complete answers of other residents from the same municipality (see [6] for more details).

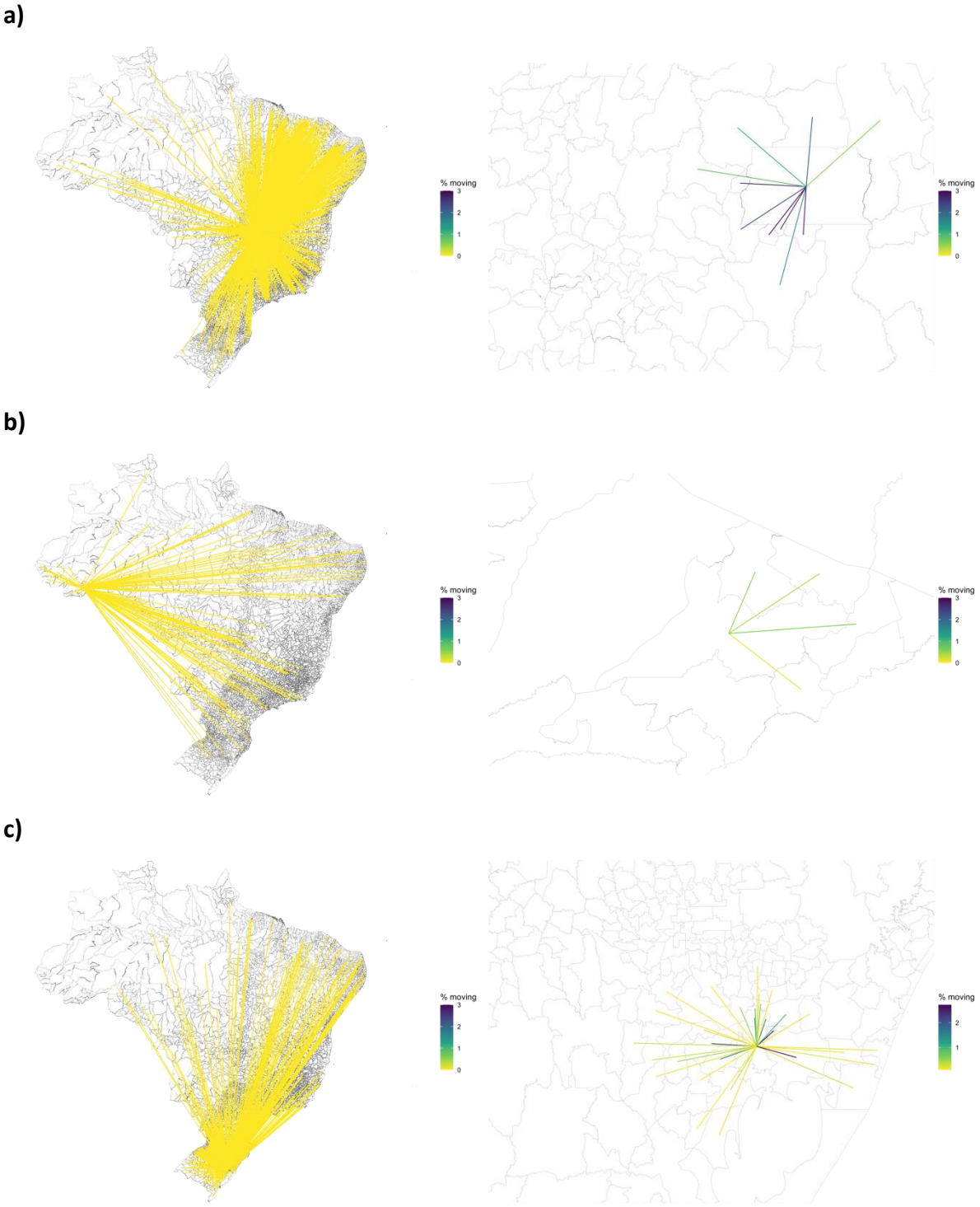


Figure 5.1: Percentage of residents regularly commuting between a) Brasilia, b) Rio Branco, and c) Porto Alegre. The percentage of residents commuting between cities in Brazil for work or education taken from the 2010 census, for all connections (left), and for connections where over 0.1% of the residents moved (right). Although connections exist across the country in each of these cities, these long-distance movements make up a small proportion compared to movements between neighbouring cities.

To measure the 'strength' of connection between municipalities based on the number of people commuting between them, the number of people was converted into the proportion of the source population using data from the 2010 Brazilian census obtained via DATASUS (http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/poptbr.def) (Figure 5.1). We applied multidimensional scaling (MDS) to these proportions to obtain an abstract cartesian coordinate system describing the relative connectivity between municipalities arising due to human movement [13,15] (Appendix E, Figure E1).

5.3.3 Modelling framework

To measure the expansion of the dengue transmission zone in Brazil between 2001 – 2020, we aggregated monthly dengue case data to annual dengue incidence rate (DIR), defined by the Brazilian Ministry of Health as the number of cases per 100,000 residents. The DIR was converted into a binary outbreak indicator using the Brazilian Ministry of Health's definition of 'high risk', over 300 cases per 100,000 residents, as a cut-off [16]. Although other outbreak definitions could be considered, our definition is consistent with Brazilian public health policy [16], and results from a previous study found that different outbreak thresholds produced similar conclusions [4].

We applied a Binomial spatial smooth model to the number of outbreaks per municipality between 2001 – 2020 using the spatial modelling framework outlined in [13]. Briefly, the model included 3 spatial terms: one assuming connectivity between municipalities as a function of distance (distance-based), one assuming connectivity between municipalities because of regular commuting (human movement-based), and another unstructured term to account for unobserved heterogeneity between municipalities. The final model equation was as follows:

$$y_i \sim binomial(p_i, 20)$$

$$log\left(\frac{p_i}{1 - p_i}\right) = \alpha + u_{1,i} + u_{2,i} + v_i$$

Where y_i is the number of outbreaks between 2001 – 2020 in municipality i $(i = 1, \ldots, 5{,}560)$, expected to follow a Binomial distribution defined by the probability of an outbreak, p_i. $u_{1,i}$ is a distance-based spatially structured term, created by applying a thin plate regression spline to

latitude-longitude coordinates of the centroid of municipalities. $u_{2,i}$ is a human movement-based spatially structured term, created by applying a thin plate regression spline to coordinates describing connectivity between municipalities arising due to regular commuting, described previously. v_i is a spatially unstructured term, assumed to follow a zero-mean Normal distribution. This unstructured term aims to capture heterogeneity between municipalities which is not spatially correlated. Spatially smooth terms were generated using the mgcv package [17] and extracted using the jagam function [18]. Model fit was carried out using Markov chain Monte-Carlo (MCMC) simulations in R via the NIMBLE package [19].

The relative contribution of each spatial term to the overall marginal variance was defined as the proportion of the overall random term variance explained (for example, the contribution of distance-based connectivity is calculated as $var(u_{1,i})/var(u_{1,i} + u_{2,i} + v_i)$) using simulations from the MCMC [13]. This was calculated for Brazil as a whole and then separately for each region of Brazil.

5.4 Results

Between 2001 and 2020, there were 1,322 (23.8%) municipalities that did not experience a dengue outbreak. Most of these (845) were in the South region and only 4 municipalities in the Centre-West region did not experience an outbreak in this period. The Goiânia and Aparecida de Goiânia, municipalities, both situated in the state of Goiás in Centre-West Brazil, each had an outbreak in 19 out of 20 years. There were no municipalities that experienced an outbreak every year between 2001 – 2020 (Figure 5.2). The Centre-West region of Brazil is currently the region in which most outbreaks occur. This pattern diverges from the early spatial distribution of dengue in the 1980s and 1990s, when transmission was most intense in warm, coastal metropoles. The Centre-West region has a relatively recent introduction of outbreaks, promoted by the expansion of agricultural areas, increasing urbanisation, and the construction of an intricate road transport network [20]. Alternatively, there remain some areas of Brazil with a virtual absence of dengue outbreaks in the past two decades, mainly located in the South region, mountainous regions, along the northern coast, and in isolated areas of the Amazon.

Figure 5.2: The number of years between 2001 and 2020 that each municipality in Brazil experienced an outbreak. The number of years between 2001 – 2020 that municipalities recorded a DIR of over 300 cases per 100,000 residents. Most municipalities that did not record an outbreak are located in South Brazil.

5.4.1 The contribution of human movement to dengue expansion in Brazil

Over the past 20 years, regular commuting for work or education was found to contribute very little to the spatial structure of dengue outbreaks in Brazil when it is considered as a whole (Figure 3, relative contribution: 0.013, 95% credible interval (CI): 0.009 – 0.017). The distance-based terms, included to account for spatial autocorrelation between close areas, contributed the most to the spatial structure in the number of dengue outbreaks, explaining 78.6% of the spatial variation in the data (relative contribution: 0.786, 95% CI: 0.769, 0.802).

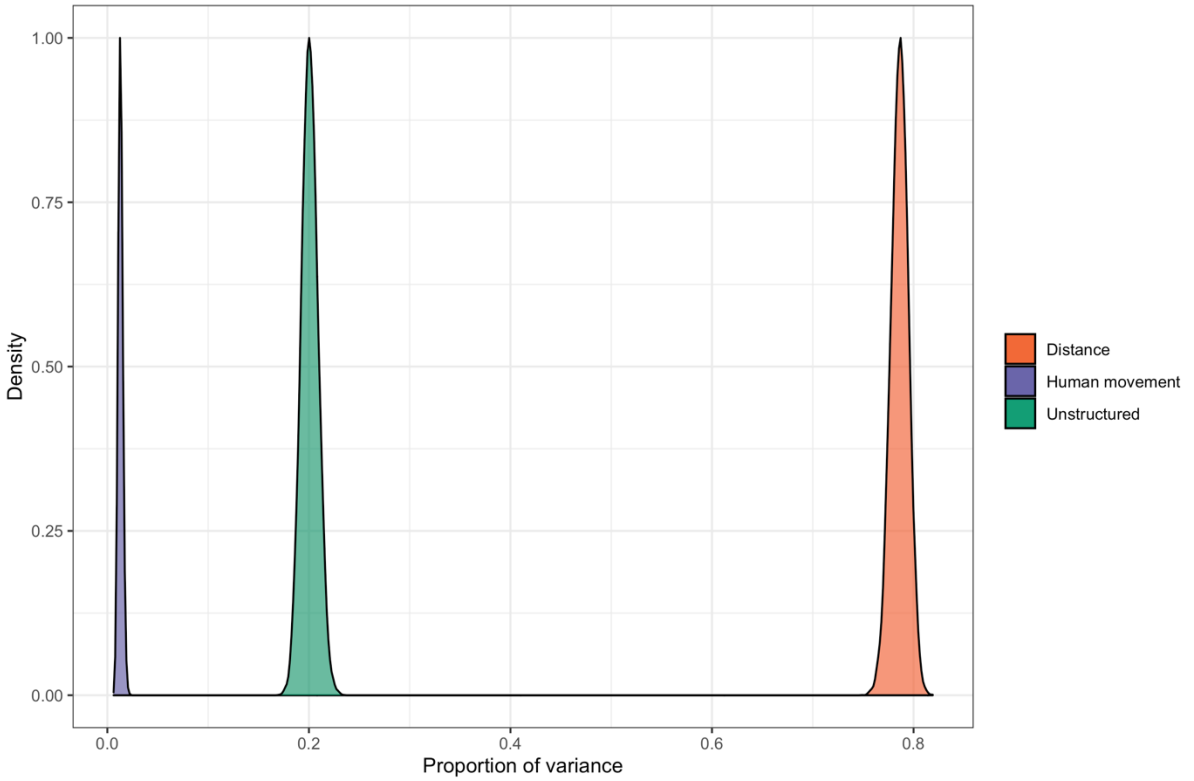


Figure 5.3: The relative contribution of each spatial random term to the overall random variance. The relative contribution is estimated using the proportion of the combined random term variance explained by each spatial random term. Note that the density curves have been rescaled to a maximum of 1 to aid interpretation.

However, when the relative contribution of spatial terms was calculated for each region separately, we found that South Brazil was mostly accounted for by distance (82.8%), whereas the unstructured term had the largest relative contribution to the spatial structure of the number of dengue outbreaks for all other regions (Figure 5.4). The contribution of the commuting term to the overall spatial variation in the model was found to be highest in the North and Northeast regions of Brazil (7.5% and 6.8% respectively). Although the contribution was relatively small, this finding suggests the expansion of outbreaks may be influenced by commuting between municipalities in these regions.

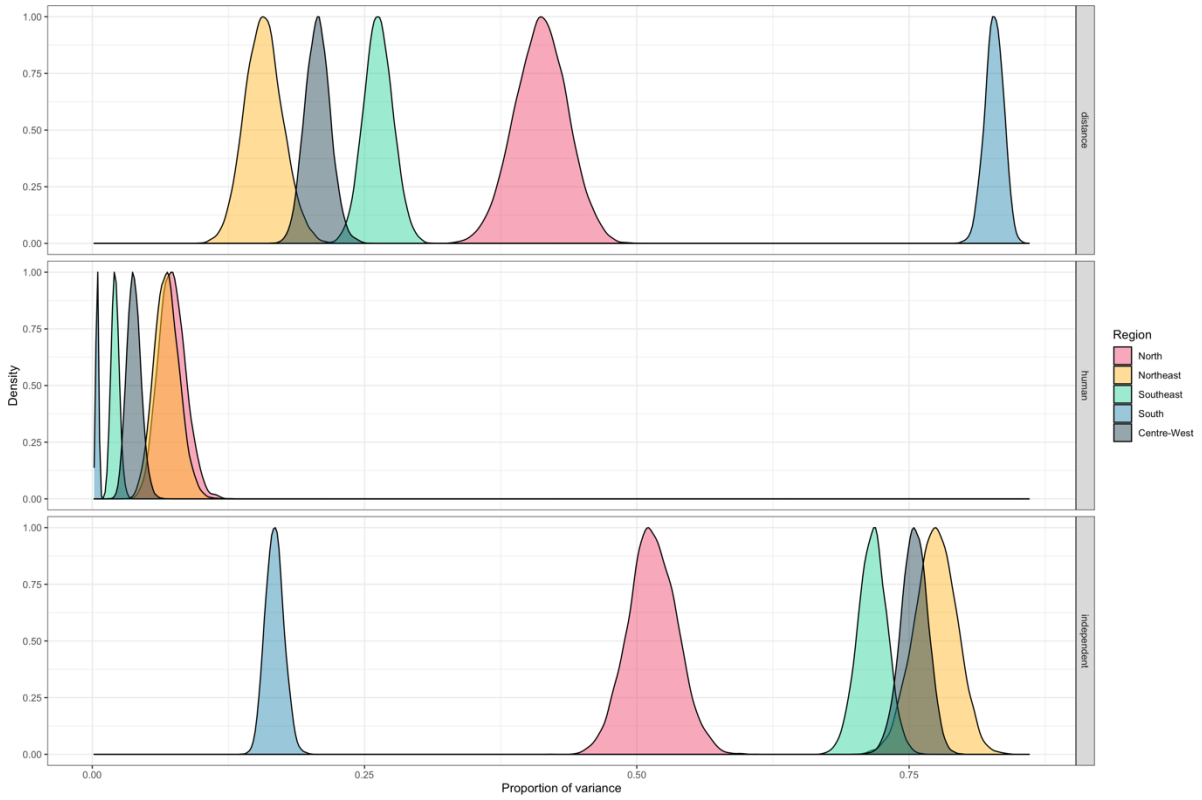


Figure 5.4: The relative contribution of each spatial random term to the overall random term variance calculated separately for each of the 5 geo-political regions of Brazil. The map is coloured according to Brazilian region. Each panel shows the probability density of the relative contribution of distance (top), regular commuting for work or education (middle) and unstructured heterogeneity (bottom) to the overall spatial structure of dengue outbreaks.

Although this model did not explicitly include covariates to describe potential drivers of dengue expansion (e.g., temperature or level of urbanisation), spatial patterns found using the random term estimates from the model can capture unmeasured or unexplained variation that may be attributable to variations in climate or socioeconomic factors across Brazil (Figure 5.5). For example, the distance-based random term estimate indicated that the odds of an outbreak in South Brazil was lower than the baseline average. This corresponds to the area of Brazil considered to be climate type C (temperate) according to the Köppen climate classification [21]. This climate type is characterised by strong seasonality and cold winter temperatures, suggesting that a temperature-based covariate might explain the difference in this area compared to the rest of Brazil (Fig 5.5a). In contrast, the regular commuting-based random term showed that the odds of an outbreak was higher in the western Amazon, where climate

conditions are favourable to dengue transmission year-round, compared to the baseline average, suggesting that human movement based on commuting patterns has a more determinant role in dengue virus diffusion in the northern region, and may have contributed to the expansion of dengue outbreaks in the region (Fig 5.5b).

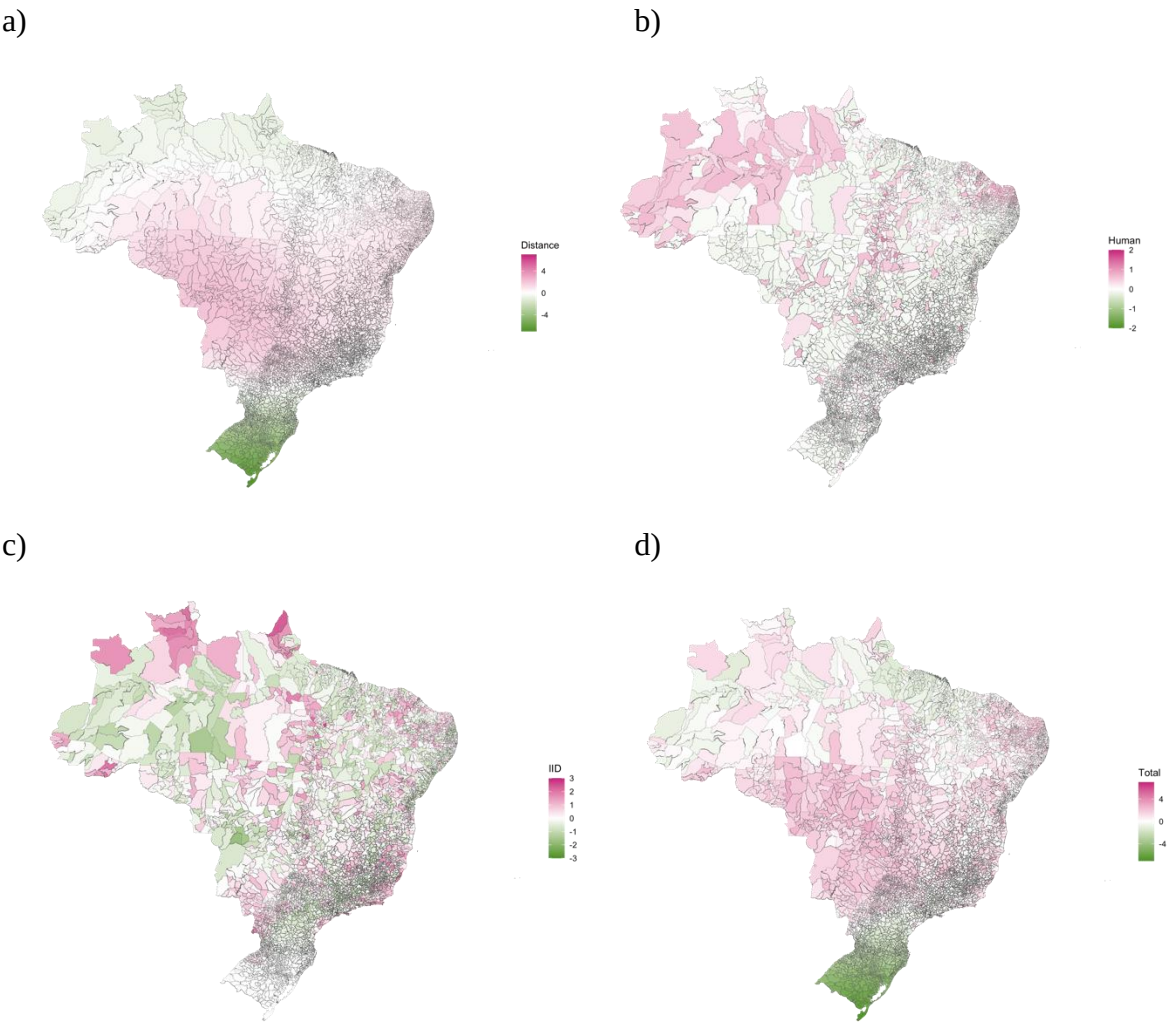


Figure 5.5: Mean estimates of the a) distance-based, b) human movement-based, c) unstructured, and d) combined random terms. Maps show regions where the odds of an outbreak was higher (lower) than the baseline average in pink (green).

5.5 Discussion

This study provides evidence that the contribution of regular commuting to the expansion of dengue differs substantially between regions of Brazil. When considering Brazil as a whole,

the model found minimal contribution of commuting to the spatial structure of dengue outbreaks between 2001 – 2020. However, when the contribution is calculated separately for each region, we can see that the relative contribution of different drivers of dengue expansion varies between regions (Figure 5.4). In particular, the increased odds of an outbreak in the western Amazon was captured by the commuting-based random term (Figure 5b), supporting our hypothesis that expansion into this region was influenced by human movement. In contrast, South Brazil had the strongest distance-based pattern of any region, indicating that other contextual factors, which are spatially structured, may play a more important role in the probability of outbreaks. The reduced odds of an outbreak in this region captured by the distance-based random term aligns with the area found to be protected due to regional temperatures, in particular, low winter temperatures [4].

There were some parts of Brazil where the odds of dengue were increased but not explained by spatially structured random terms and were captured by the independent unstructured term (Figure 5.5c). These areas of increased odds often correspond to areas close to Brazil's international borders such as the states of Roraima (bordering Venezuela and Guyana), Amapa (bordering French Guiana) and parts of Amazonas state (bordering Colombia). This suggests that dengue virus may have been imported into these regions of Brazil internationally, as is the case in other non-endemic countries such as Argentina, USA and Europe [22–24]. It is worth noting that the data used to inform the spatial random terms included in the model did not consider the international human movement nor distances to bordering international cities. Another municipality with a notable increase in odds captured by the unstructured term was Rio Branco, the capital of Acre state. Rio Branco was one of the first municipalities in the western Amazon to experience dengue transmission and has experienced explosive outbreaks since 2001. Although dengue was established in Rio Branco early in the 21st century, it took several years for the virus to affect surrounding municipalities. When this expansion did happen, outbreaks appeared to 'jump' between municipalities, potentially due to increased air travel and improvements to the road network within the state [6].

There are several limitations to this study. First, the regular commuting data was sourced from the 2010 census. This provides a snapshot of connectivity across Brazil and does not allow us to investigate the impact of changes in commuting patterns over the past 20 years on the dengue transmission zone. Although the model indicated that commuting for work or education has contributed to patterns of dengue outbreaks in the western Amazon, this commuting behaviour

was assumed to be constant between 2001 – 2020. A previous study investigating the introduction of dengue into the state of Acre in the western Amazon noted that the census data is most likely representative of the situation since 2009 in the region, following the completion of major maintenance works on the highway connecting the two largest cities in the state [6]. Another limitation of the human movement data is that it only contains information about residents that travel regularly for work or education [14] and therefore does not include irregular, long-distance connections that are less frequent but have been shown to be important when considering (re-)emergence of mosquito-borne diseases [25]. Despite this, the data did capture some long-distance travel, particularly between large cities in the Amazon region (Figure 1). Future work could consider an alternative model formulation that allows temporal trends to be included and alternative sources of human movement data that capture different types of human movement and the changing patterns of movement over the period.

Although this model does not explicitly include temperature or other explanatory variables, we are able to generate hypotheses about potential drivers of dengue outbreaks using the random term estimates. These hypotheses support previous research that showed South Brazil is protected due to its lower temperatures and North Brazil was previously protected due to its disconnection to the Brazilian urban network [4]. This work contributes to previous literature about the expansion of the dengue transmission zone in South America [4,5,23,26] and provides evidence that human movement plays a role in this expansion. Our results highlight the importance of considering different drivers that may be taking place across a large geographical region like Brazil, which varies enormously in terms of climatic, socioeconomic, and demographic factors.

References

1. Gubler DJ. 2011 Dengue, Urbanization and Globalization: The Unholy Trinity of the 21st Century. *Trop. Med. Health* **39**, 3–11. (doi:10.2149/tmh.2011-S05)

2. Kraemer, M.U., Reiner, R.C., Brady, O.J., Messina, J.P., Gilbert, M., Pigott, D.M., Yi, D., Johnson, K., Earl, L., Marczak, L.B., Shirude, S. 2019 Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nat. Microbiol.* **4**, 854–863. (doi:10.1038/s41564-019-0376-y)

3. Codeco, C.T., Oliveira, S.S., Ferreira, D.A., Riback, T.I., Bastos, L.S., Lana, R.M., Almeida, I., Godinho, V.B., Cruz, O.G., Coelho, F.C. 2022 Fast expansion of dengue in Brazil. *Lancet Reg. Health - Am.* **12**, 100274. (doi:10.1016/j.lana.2022.100274)

4. Lee SA, Economou T, Catão R de C, Barcellos C, Lowe R. 2021 The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil. *PLoS Negl. Trop. Dis.* **15**, e0009773. (doi:10.1371/journal.pntd.0009773)

5. Barcellos C, Lowe R. 2014 Expansion of the dengue transmission area in Brazil: the role of climate and cities. *Trop. Med. Int. Health* **19**, 159–168. (doi:10.1111/tmi.12227)

6. Lana RM, Gomes MF da C, de Lima TFM, Honório NA, Codeço CT. 2017 The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: A network-based analysis. *PLoS Negl. Trop. Dis.* **11**. (doi:10.1371/journal.pntd.0006070)

7. Lowe R, Lee S, Lana RM, Codeço CT, Castro MC, Pascual M. 2020 Emerging arboviruses in the urbanized Amazon rainforest. *BMJ* **371**, m4385. (doi:10.1136/bmj.m4385)

8. IBGE. 2020 *Regiões de influência das cidades 2018*.

9. Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021 Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**, 20210096. (doi:10.1098/rsif.2021.0096)

10. Findlater A, Bogoch II. 2018 Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends Parasitol.* **34**, 772–783.

11. Kraemer, M.U., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D.M., Ray, S.E., Brady, O.J., Brownstein, J.S., Faria, N.R., Cummings, D.A.T., Pybus, O.G. 2019 Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 1–11.

12. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Paz Soldan V, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Dis.* **3**, e481.

13. Lee SA, Economou T, Lowe R. 2022 A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping. *J. R. Soc. Interface* **19**, 20220440. (doi: 10.1098/rsif.2022.0440)

14. IBGE. 2012 Censo demográfico 2010: resultados gerais da Amostra.

15. Cox MA, Cox TF. 2008 Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347.

16. Lowe, R., Barcellos, C., Coelho, C.A., Bailey, T.C., Coelho, G.E., Graham, R., Jupp, T., Ramalho, W.M., Carvalho, M.S., Stephenson, D.B., Rodó, X. 2014 Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. *Lancet Infect. Dis.* **14**, 619–626. (doi:10.1016/S1473-3099(14)70781-9)

17. Wood S, Wood MS. 2015 Package 'mgcv'. *R Package Version* **1**, 729.

18. Wood SN. 2016 Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75**, 1–15. (doi:10.18637/jss.v075.i07)

19. De Valpine P *et al.* 2021 Nimble: MCMC, particle filtering, and programmable hierarchical modeling. *R Package Version 011* **1**.

20. Teixeira MG, Costa MDCN, Barreto F, Barreto ML. 2009 Dengue: Twenty-five years since reemergence in Brazil. *Cad. Saude Publica* **25**. (doi:10.1590/S0102-311X2009001300002)

21. Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G. 2013 Köppen's climate classification map for Brazil. *Meteorol. Z.* , 711–728. (doi:10.1127/0941-2948/2013/0507)

22. Gwee XWS, Chua PEY, Pang J. 2021 Global dengue importation: a systematic review. *BMC Infect. Dis.* **21**, 1078. (doi:10.1186/s12879-021-06740-1)

23. Robert MA, Tinunin DT, Benitez EM, Ludueña-Almeida FF, Romero M, Stewart-Ibarra AM, Estallo EL. 2019 Arbovirus emergence in the temperate city of Córdoba, Argentina, 2009–2018. *Sci. Data* **6**, 276. (doi:10.1038/s41597-019-0295-z)

24. Wilder-Smith A, Quam M, Sessions O, Rocklov J, Liu-Helmersson J, Franco L, Khan K. 2014 The 2012 dengue outbreak in Madeira: exploring the origins. *Eurosurveillance* **19**, 20718. (doi:10.2807/1560-7917.ES2014.19.8.20718)

25. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis* **3**, e481.

26. Churakov M, Villabona-Arenas CJ, Kraemer MUG, Salje H, Cauchemez S. 2019 Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony. *PLoS Negl. Trop. Dis.* **13**. (doi:10.1371/journal.pntd.0007012)

6. Discussion

6.1 Summary of findings

The primary aim of this thesis was to understand the complex, interacting drivers of dengue expansion in Brazil. This aim led to the following four objectives:

1. Explore the impact of temperature suitability, urbanisation, and connectivity of cities to the Brazilian urban network on the expansion of the dengue transmission zone in Brazil (Chapter 2)
2. Identify spatial modelling techniques currently used to study mosquito-borne disease transmission and the assumptions made in modelling studies about how spatial connectivity arises, and describe the data used to inform spatial models (Chapter 3)
3. Develop a statistical modelling framework capable of including multiple sources of spatial connectivity and quantifying the relative contribution of each source to the overall spatial structure of the data (Chapter 4)
4. Quantify the relative contribution of human movement to the expansion of dengue outbreaks in Brazil (Chapter 5)

In this section, I begin by summarising the main findings of previous chapters in relation to these objectives. Following this, I discuss the strengths and limitations of the research presented in this thesis. I then present future research opportunities that may arise from this project.

6.1.1 Objective 1: Explore the impact of temperature suitability, urbanisation, and connectivity of cities to the Brazilian urban network on the expansion of the dengue transmission zone in Brazil

In Chapter 2, a spatiotemporal generalised additive model (GAM) was applied to a binary dengue outbreak indicator, using an outbreak threshold of over 300 dengue cases per 100,000 residents. The model included hypothesised drivers of dengue expansion as fixed covariates, and spatial, temporal and spatiotemporal smooth terms to account for patterns in the data which were not captured by the covariates. Results from this model showed that the odds of an outbreak were significantly increased in highly urbanised, highly connected municipalities that experienced year-round suitable temperatures and had previously experienced an outbreak. A

comparison of the spatiotemporal smooth terms to a baseline model containing no covariates showed that temperature suitability explained most interannual and spatial variation in South Brazil, supporting the hypothesis that this region was protected due to low winter temperatures. Barriers to the dengue transmission zone were redrawn using the results from this model. Although a southern border still exists, this has shifted further south, and the western Amazon no longer has a clear barrier. Another barrier was identified along the northern coast of Brazil which was not explained by the covariates in this model.

The level of connectivity of cities was defined using the Regions of Influence of Cities (REGIC) study which categorised cities into 5 levels of influence from the most connected, metropoles, to least, zone centres [1]. Although connectivity within the Brazilian urban network was identified as a significant driver of dengue outbreaks, the relationship was nonlinear. Regional capitals had the highest increase in the odds of an outbreak but were considered less connected than metropoles. There are many potential reasons for this result, such as differences in socioeconomic factors, healthcare investment, or health seeking behaviours between metropoles and regional capitals that could not be captured at a municipality level. Another hypothesis was that this nonlinear trend was related to human movement which may not be fully captured within the REGIC connectivity indicators. Although the model presented in Chapter 2 included a spatially structured term, this was only able to capture spatial connectivity between close regions due to its smooth structure. As the current model was not able to incorporate human movement into the spatial structure of the random terms, or account for multiple (distance and human movement) sources of connectivity, identifying models that could became the basis for Objective 2.

6.1.2 Objective 2: Identify spatial modelling techniques currently used to study mosquito-borne disease transmission and the assumptions made in modelling studies about how spatial connectivity arises, and describe the data used to inform spatial models

The systematic review presented in Chapter 3 synthesised spatial modelling approaches described in the literature used to study the transmission of mosquito-borne diseases to humans, and the spatial connectivity assumptions that they made. Models were classified as statistical, mechanistic, machine learning or a combination of these approaches. Although this PhD

considers dengue expansion as a case study, the systematic review considered 9 mosquito-borne diseases as the issue of spatial connectivity and the assumption of how this arises is likely shared across diseases.

There were 248 published studies eligible for inclusion that used a spatial model to investigate the transmission of a mosquito-borne disease to humans. Of these, over 80% used a statistical model, most frequently a mixed effect model. All mixed effect models used a distance-based function to describe the relative 'connectedness' of areas or observations. The only statistical method identified that could include links arising due to human movement was the inclusion of spatial covariates (e.g., the number of people moving between areas) in a generalised linear model (GLM). Spatial covariates are a relatively quick and simple way to account for spatial connectivity in a model as they are added to a GLM in the same way as nonspatial covariates and can be interpreted in the same way. However, this approach required one covariate per connection in the data. In Brazil, metropoles such as Brasilia and São Paulo are connected across the entire country so would require thousands of covariates. The inclusion of many covariates in a GLM risks overfitting the data and introducing multicollinearity. GLMs also assume that the relationship between the outcome and spatial covariates is the same across time and space (stationarity). Given the size of Brazil and the diversity in the movement patterns across regions [1], the assumption of stationarity is not appropriate for this setting.

Only 50 of the studies included in the review assumed that spatial connectivity was related to human movement, despite it being recognised as an important driver of mosquito-borne disease transmission [2–5]. The assumption of human movement-based connectivity was more likely when studying *Aedes*-borne diseases and within a mechanistic model. This was likely because some mechanistic models (i.e., metapopulation and agent-based models) are designed to include complex networks that describe the movement of a population between nodes.

This systematic review found that there were no statistical modelling frameworks currently used for mosquito-borne disease transmission that were appropriate to quantify the role of human movement on the expansion of dengue in Brazil. Expanding the current statistical frameworks using ideas taken from network-based mechanistic models became the focus of Objective 3.

6.1.3 Objective 3: Develop a statistical modelling framework capable of including multiple sources of spatial connectivity and quantifying the relative contribution of each source to the overall spatial structure of the data

Chapter 4 presented a novel Bayesian hierarchical modelling framework that allows multiple sources of spatial connectivity, in this case distance and human movement, to be included in a single model. Spatially structured terms were constructed using penalised smoothing splines of coordinates that describe the relative connectedness of areas. This creates a 2-dimensional smooth surface describing the spatial structure of the data which can be incorporated into a hierarchical model and interpreted similarly to traditional random effects. Smoothing splines such as these can be applied to any symmetric continuous measure of connectivity, for example distance or the number of people moving between areas. These functions require minimal user assumptions about the spatial structure of the data beyond smoothness.

Using model inference and simulations, the proportion of the marginal variance explained by each spatial term can be computed from the proposed model. This result provides an estimate of the relative contribution of each spatial term to the overall structure, as demonstrated by simulation studies. This method could therefore be used to quantify the relative contribution of human movement to the expansion of dengue transmission in Brazil, the final objective of this thesis. A case study was carried out using dengue case data in South Brazil between 2001 – 2020 and human movement data generated using a movement model, assuming the total number of people moving between municipalities was a function of distance and population. The model found that human movement between municipalities did not account for a significant proportion of the spatial structure of the average dengue incidence rate between 2001 – 2020 in South Brazil.

6.1.4 Objective 4: Quantify the relative contribution of human movement to the expansion of dengue outbreaks in Brazil

The modelling framework presented in Chapter 4 was applied to the number of dengue outbreaks that occurred between 2001 – 2020 per municipality, defined as over 300 dengue cases per 100,000 residents, for the whole of Brazil. The model contained spatially structured random terms generated by applying penalised smooth splines to coordinates of the centroid of municipalities (aiming to capture spatial autocorrelation between close regions, referred to as

distance-based connectivity) and coordinates generated to describe the 'connectedness' of municipalities based on the number of people travelling between them for work or education taken from the 2010 census [6] (capturing spatial connectivity arising from human movement). The model also contained an unstructured, independent random term to capture the remaining heterogeneity in the number of dengue outbreaks between municipalities. The proportion of the marginal variance explained by each random term was estimated and used to describe the relative contribution of each term to the overall spatial structure of the data.

When Brazil was considered as a whole, regular commuting was found to contribute very little to the overall spatial structure of the number of dengue outbreaks. However, when this contribution was calculated for each geo-political region of Brazil separately, the relative contribution differed substantially across the country. The distance-based term contributed the most to the South region, but the unstructured random term had the largest relative contribution to the spatial structure of the data in all other regions of Brazil. The North and Northeast regions had the highest contribution of regular commuting which, although still relatively small, was significantly higher than South Brazil. When the random terms were extracted and visualised on a map, the commuting term was shown to capture the increased odds of an outbreak in the western Amazon compared to a baseline average. The distance-based terms captured the lower-than-average odds of an outbreak in South Brazil. This area corresponded to the 'protected' region identified in Chapter 2, which also aligned with the area classified as temperate by the Köppen climate classification [7,8]. This suggests that the spatial autocorrelation between municipalities in South Brazil captured by the distance-based term might be attributed to similar climate conditions across the region, particularly low winter temperatures, which are less suitable for dengue transmission than other parts of Brazil.

In summary, the results from this thesis show that the dengue outbreak zone has expanded between 2001 and 2020, and the current barriers to dengue transmission are determined by temperature, levels of urbanisation, and connectivity within the Brazilian urban network. Although the barriers to dengue transmission presented in Chapter 2 are likely already outdated due to the rapid expansion of dengue across Brazil [9], they highlight the importance of focusing control efforts on areas at risk of future outbreaks as well as those in the current transmission zone.

6.2 Strengths

Despite the importance of space in epidemiology being well recognised, spatial modelling studies have been underrepresented in mosquito-borne disease literature [10]. The systematic review presented in Chapter 3 found that the number of spatial modelling papers had increased rapidly over the past 10 years but that very few of these considered spatial connectivity arising due to human movement. Almost all spatial statistical models of mosquito-borne diseases assumed that spatial connectivity existed between observations if and only if they were 'close', defined using distance or adjacency. Many of these papers hypothesised that human movement was a driver of disease transmission but there were no statistical methods identified that could include additional sources of spatial connectivity, such as human movement, into a model to test this hypothesis. A major strength of this thesis is that it fills this important methodological gap. The modelling framework in Chapter 4 allows multiple sources of spatial connectivity to be include within a statistical model and can quantify their relative contribution to the overall spatial structure of the outcome.

Although the modelling framework presented in Chapter 4 requires technical statistical knowledge and skills, the approach was based on existing statistical methods that are well-documented and used throughout spatiotemporal epidemiology. Therefore, this approach should be accessible to any researcher familiar with conventional spatial modelling approaches and their interpretations, in particular Bayesian hierarchical models and generalised additive models. To aid this understanding, a detailed description of all models identified in the systematic review was provided as a technical appendix (Appendix C.1) which is freely available online [11]. These descriptions also include examples of how each model has been applied to mosquito-borne disease transmission. Care was taken throughout the entire thesis to provide full, detailed analysis methods in accessible language to ensure others could replicate these approaches. Where there was not sufficient space in a standard journal article full, detailed descriptions of analytical methods were provided as supplementary materials. All code contained within public repositories have detailed comments explaining each procedure to aid replicability.

Another strength of this thesis is the wide applicability of the methods used and developed. Although dengue expansion in Brazil was presented as a case study, spatial connectivity plays an important role in the transmission of other infectious diseases. Spatial connectivity due to

human movement is particularly important when considering (re-)emerging infectious diseases as links created by human movement between regions with active disease transmission and currently unaffected but susceptible populations allows these diseases to spread [4,12,13]. Although the assumption of how spatial connectivity arises (e.g., human movement) may be shared across diseases, the structure of this connectivity is likely to be different. Empirical Bayesian inferential methods using penalised regression splines, such as the ones used in this thesis, estimate the spatial structure of the data as part of the model fitting process rather than requiring it to be specified a priori [14,15]. This means that the same model fitting approach can be used across different settings but produce a spatial surface describing spatial connectivity which is setting-specific and requires no assumptions of this structure beyond smoothness. The simulation studies presented in Chapter 4 show that this approach is applicable to models for count data and binary outcomes, however it is flexible enough to be applied in any Bayesian hierarchical model where spatial connectivity exists within the data.

Great care has been taken to ensure that all research presented in this thesis is reproducible and as accessible as possible. There is some discussion about the true definition of 'reproducible' in scientific research. However, it is generally agreed that the minimum requirement is ensuring that all data and software used to carry out analysis are made available [16–18]. Each research chapter included in this thesis has its own public repository, published on Github, containing the analytical dataset and computer code used to carry out analyses. All analyses were carried out using the free, open-source statistical package R [19] to make reproduction as accessible as possible. In addition, all data used throughout the thesis are open-source and details of how to access these databases were provided within each chapter.

6.3 Limitations

All spatial models developed throughout this thesis have been applied to real-world, open-source data providing evidence of their applicability and usefulness in studying emerging infectious diseases. However, real data is often messy and contains inherent biases depending on how it is collected. A major limitation of this thesis is that the conclusions drawn from modelling results rely heavily on the quality of the data used to fit them. Dengue case data used in this thesis was obtained from the Brazilian Notifiable Diseases Information System (SINAN) which is passive and therefore likely to miss mild and asymptomatic cases where individuals

do not require medical assistance [20–22]. The accuracy of the surveillance system is also known to vary across the country, particularly between areas with established dengue transmission and previously disease-free areas [23]. As it is impossible to know the true dengue incidence, it is difficult to accurately account for these inherent biases within models.

Biases in dengue case data were likely further confounded by other infectious diseases co-circulating within Brazil at the same time. Zika and chikungunya are arboviruses transmitted to humans via the same mosquito vector as dengue, *Aedes aegypti*, that have been co-circulating in Brazil since approximately 2013 [24,25]. Dengue, Zika and chikungunya can produce similar clinical symptoms, including fever, joint pain and lethargy, making them difficult to distinguish between [26,27]. Although there are some differences in clinical manifestations, such as the severity and duration of symptoms, a study carried out in Northeast Brazil found that misclassification can occur, particularly in periods of intense simultaneous circulation of these arboviruses [28]. When suspected dengue cases are laboratory tested, this may not always overcome the issue of misclassification as cross-reaction between antibodies of the dengue and Zika viruses may lead to false positive serological tests [26].

To reduce the potential impact of differences in dengue reporting across Brazil and between epidemic and non-epidemic periods, I used a binary outbreak indicator as an outcome rather than incidence when modelling the data. Although the raw case data may not always be accurate, the surveillance system should be able to detect when an outbreak is occurring [29]. There are several different approaches to defining outbreak thresholds and the definition of an outbreak is likely to differ between endemic and disease-free areas. Throughout this thesis, I have chosen to use an outbreaks threshold of over 300 cases per 100,000 residents, defined as 'high risk' by the Brazilian Ministry of Health [30]. Sensitivity analyses in Chapter 2 confirmed that model results were similar when alternative outbreak thresholds were used (see Appendix B.1). Further investigation is necessary to assess the impact of arbovirus misclassification on the accuracy of results since the introduction of chikungunya and Zika.

A further limitation of this thesis arising from data was the lack of temporal resolution in socioeconomic variables such as urbanisation and access to piped water used in Chapter 2, and in the commuting data used in Chapter 5. These variables were taken from the Brazilian census which is usually carried out every 10 years. The 2020 Brazilian census was postponed due to the COVID-19 pandemic and is currently underway [31,32]. Therefore, census data was only

available from 2010 for this research, with no information about how these factors have changed over time. This severely limits the inferences that can be made about how changes in socioeconomic factors and connectivity arising from human movement have driven the expansion of the dengue transmission zone within Brazil. Future work could consider alternative sources or measures with a finer temporal resolution to explore these drivers further, for example by using land use data as a measure of changing urbanisation [33].

One of the main focuses of this thesis was the inclusion of human movement within a spatial model of emerging infectious diseases. However, human movement is a very broad term and can refer to many different types of connections depending on the setting and the spatial and temporal scales of the problem [4,34]. Another limitation of this thesis was that only one source of human movement was considered which included regular commuting trips. The human movement data used to inform the model presented in Chapter 5 was taken from the 2010 census and contains information about regular travel for work and education taken from a survey carried out on a random sample of the population [35]. This data fails to capture long-distance, irregular trips which are known to drive disease (re-)emergence [4,34] and is likely biased towards shorter connections. Despite this, some long-distance connections were present in the data, particularly between large cities and in the North region. This suggests that this data captured some of the regional differences in human movement behaviour in Brazil identified in the Regions of Influence of Cities (REGIC) studies [1,36]. For example, journeys in North Brazil are much longer on average due to the remoteness of some municipalities. This has led to the exclusion of the North region in some studies to minimise the biases which arise due to the differences in travel patterns [37]. By using survey data rather than assuming the same patterns of movement across the country, which is commonly assumed in movement models [38,39], I have been able to compare differences in the relative importance of human movement between different regions of Brazil while accounting for these differences. Future work could consider additional data sources such as air travel information and mobile phone data to account for irregular movements and provide information about changing patterns of movement over time.

6.4 Future work

There are several ways in which the work presented in this thesis could be extended and used to further understand (re-)emerging infectious disease transmission.

The first objective of this thesis was to understand how changes in climate, socioeconomic factors, and connectivity have impacted the expansion of the dengue transmission zone in Brazil. However, the extent to which this could be addressed was limited by data availability and model design. Socioeconomic variables such as levels of urbanisation and access to basic services were obtained via the Brazilian census which is carried out every 10 years. There was no information about how these factors changed in the intermediate years, nor about how they have changed since as the 2020 census was delayed [31]. Future work could explore alternative data sources, such as land use data that can be used to extract fine-scale information about changes in urbanisation over time [33]. The relative contribution of each term could then be compared over time to explore whether the rapid urbanisation of some previously protected areas has had a larger impact.

Increasing temperature was found to contribute to the expansion of dengue into South Brazil in Chapter 2. This was accounted for in the model using an indicator based on the temperatures *Aedes* mosquitoes can transmit the dengue virus to humans [40]. However, other climate variables such as precipitation, drought and humidity are also known to play a role in dengue transmission dynamics [41,42]. At present, there is no known optimal hydrometeorological conditions for dengue transmission, and the interaction between these conditions and socioeconomic factors, such as water storage practices, means this is likely to change across space [41]. An ‘extremely wet’ indicator was explored as part of the analysis carried out in Chapter 2 which was defined using the self-calibrating Palmer drought severity index (scPDSI), a measure of how wet or dry an area is relative to ‘normal’ conditions [43]. This indicator was not found to improve the model, however future work could be carried out to investigate other hydrometeorological indicators to investigate their role in dengue expansion.

The models presented in this thesis to explore the relative contribution of human movement on infectious disease expansion requires the data to be symmetric and stationary. However, this is not how human movement behaves [38,44]. Patterns of human movement in Brazil have changed over the past 20 years, as shown by the two latest REGIC studies [1,36]. The REGIC

study in 2018 showed that, although the North region still has fewer highly influential cities than other regions, the number of regional capitals and sub-regional centres had increased in the North and the number of connections between them has also increased (Appendix B.1, Figure E). This is likely a result of improved infrastructure and connectivity by air travel in the region since 2009 [45]. Future work is needed to explore how the spatial models presented in Chapters 4 and 5 can be extended to include temporal trends to explore the impact of changing human movement patterns on disease expansion and include more complex, realistic human movement networks. An interaction between spatial and temporal random terms, similar to those included in the model in Chapter 2, would allow the spatial patterns to change over time and allow hypotheses to be drawn about how changing patterns of human movement could lead to further expansion of the dengue transmission zone in Brazil in the future.

Finally, although this thesis uses dengue re-emergence and expansion in Brazil as a case study, spatial connectivity is an important driver of many infectious diseases. It would be interesting to apply the modelling framework developed in Chapter 4 to other infectious diseases and compare the relative importance of human movement. For example, Brazil was one of the most severely affected countries during the COVID-19 pandemic following the rapid spread of the virus across the country [46]. Human movement has been recognised as one of the major drivers of this spread, with São Paolo considered a 'super spreader' city due to its high level of influence and connectivity to the whole country [47]. In theory, a model with COVID-19 incidence as the outcome would have a far higher contribution of human movement to the spatial structure than a mosquito-borne disease such as dengue.

6.5 Concluding statement

In this thesis, I have explored the complex, interacting drivers of dengue expansion in Brazil since 2001. I have shown that increasing temperatures in South Brazil, high levels of urbanisation, and connections between cities arising from human movement have all played a role in the erosion of geographical barriers to the dengue transmission zone in Brazil. As part of this thesis, I have considered existing spatial modelling frameworks and presented a novel statistical approach to deal with the issue of complicated spatial connectivity systems that exist within infectious disease epidemiology. Given the increasing risk of future pandemic pathogens due to changes in climate and increased global connectivity [48,49], robust spatial

modelling tools are essential to gain better understanding of infectious disease emergence and identify areas at future risk of expansion.

References

1. IBGE. 2020 Regiões de influência das cidades 2018.

2. Stoddard, S.T., Forshey, B.M., Morrison, A.C., Paz-Soldan, V.A., Vazquez-Prokopec, G.M., Astete, H., Reiner Jr, R.C., Vilcarromero, S., Elder, J.P., Halsey, E.S., Kochel, T.J., 2013 House-to-house human movement drives dengue virus transmission. *Proc. Natl. Acad. Sci.* **110**, 994–999.

3. Kraemer, M.U., Bisanzio, D., Reiner, R.C., Zakar, R., Hawkins, J.B., Freifeld, C.C., Smith, D.L., Hay, S.I., Brownstein, J.S., Perkins, T.A. 2018 Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan. *Epj Data Sci.* **7**, 17.

4. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009 The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl Trop Dis* **3**, e481.

5. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. 2016 Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *J. Infect. Dis.* **214**, S414–S420.

6. IBGE. 2012 Censo demográfico 2010: resultados gerais da Amostra.

7. Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G. 2013 Köppen's climate classification map for Brazil. *Meteorol. Z.* , 711–728.

8. Dubreuil V, Fante K, Planchon O, Sant'anna Neto J. 2019 Climate change evidence in Brazil from Köppen's climate annual types frequency. *Int. J. Climatol.* **39**, 1446–1456.

9. Secretaria de Vigilância em Saúde. 2022 Monitoramento dos casos de arboviroses até a semana epidemiológica 18 de 2022.

10. Reiner Jr, R.C., Perkins, T.A., Barker, C.M., Niu, T., Chaves, L.F., Ellis, A.M., George, D.B., Le Menach, A., Pulliam, J.R., Bisanzio, D., Buckee, C. 2013 A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. *J. R. Soc. Interface* **10**, 20120921.

11. Lee SA, Jarvis CI, Edmunds WJ, Economou T, Lowe R. 2021. Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions. *J. R. Soc. Interface* **18**, 20210096.

12. Kraemer, M.U., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D.M., Ray, S.E., Brady, O.J., Brownstein, J.S., Faria, N.R., Cummings, D.A.T., Pybus, O.G. 2019 Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 1–11.

13. Findlater A, Bogoch II. 2018 Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends Parasitol.* **34**, 772–783.

14. Wood SN. 2017 *Generalized additive models: an introduction with R*.

15. Pedersen EJ, Miller DL, Simpson GL, Ross N. 2019 Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* **7**, e6876.

16. Peng RD, Hicks SC. 2020 Reproducible research: a retrospective. *ArXiv Prepr. ArXiv200712210*

17. Peng RD, Dominici F, Zeger SL. 2006 Reproducible epidemiologic research. *Am. J. Epidemiol.* **163**, 783–789.

18. Goodman SN, Fanelli D, Ioannidis JP. 2016 What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12-341ps12.

19. R Core Team. 2021 *R: A language and environment for statistical computing*. 4.1.1. Vienna, Austria. See https://www.R-project.org/.

20. Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F. 2013 The global distribution and burden of dengue. *Nature* **496**, 504–507. (doi:10.1038/nature12060)

21. Coelho GE, Leal PL, Cerroni M de P, Simplicio ACR, Siqueira Jr JB. 2016 Sensitivity of the dengue surveillance system in Brazil for detecting hospitalized cases. *PLoS Negl. Trop. Dis.* **10**, e0004705.

22. Angelo M, Ramalho WM, Gurgel H, Belle N, Pilot E. 2020 Dengue surveillance system in Brazil: A qualitative study in the federal district. *Int. J. Environ. Res. Public. Health* **17**, 2062.

23. Silva, M.M., Rodrigues, M.S., Paploski, I.A., Kikuti, M., Kasper, A.M., Cruz, J.S., Queiroz, T.L., Tavares, A.S., Santana, P.M., Araújo, J.M., Ko, A.I. 2016 Accuracy of Dengue Reporting by National Surveillance System, Brazil. *Emerg. Infect. Dis.* **22**, 336–339.

24. Lowe R, Barcellos C, Brasil P, Cruz OG, Honório NA, Kuper H, Carvalho MS. 2018 The Zika Virus Epidemic in Brazil: From Discovery to Future Implications. *Int. J. Environ. Res. Public. Health* **15**, 96.

25. Souza, T.M.L., Vieira, Y.R., Delatorre, E., Barbosa-Lima, G., Luiz, R.L.F., Vizzoni, A., Jain, K., Miranda, M.M., Bhuva, N., Gogarten, J.F., Ng, J. 2019 Emergence of the East-Central-South-African genotype of Chikungunya virus in Brazil and the city of Rio de Janeiro may have occurred years before surveillance detection. *Sci. Rep.* **9**, 2760.

26. Pan American Health Organization. 2017 Tool for the diagnosis and care of patients with suspected arboviral diseases.

27. Rodriguez-Morales AJ, Villamil-Gómez WE, Franco-Paredes C. 2016 The arboviral burden of disease caused by co-circulation and co-infection of dengue, chikungunya and Zika in the Americas. *Travel Med. Infect. Dis.* **14**, 177–179.

28. Silva, M.M., Tauro, L.B., Kikuti, M., Anjos, R.O., Santos, V.C., Gonçalves, T.S., Paploski, I.A., Moreira, P.S., Nascimento, L.C., Campos, G.S., Ko, A.I. 2019 Concomitant transmission of dengue, chikungunya, and Zika viruses in Brazil: clinical and epidemiological findings from surveillance for acute febrile illness. *Clin. Infect. Dis.* **69**, 1353–1359.

29. Barcellos C, Lowe R. 2014 Expansion of the dengue transmission area in Brazil: the role of climate and cities. *Trop. Med. Int. Health* **19**, 159–168.

30. Lowe, R., Barcellos, C., Coelho, C.A., Bailey, T.C., Coelho, G.E., Graham, R., Jupp, T., Ramalho, W.M., Carvalho, M.S., Stephenson, D.B., Rodó, X. 2014 Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. *Lancet Infect. Dis.* **14**, 619–626.

31. IBGE (Instituto Brasileiro de Geografia e Estatística). 2020 Censo 2020 adiado para 2021 | IBGE. See https://www.ibge.gov.br/novo-portal-destaques/27161-censo-2020-adiado-para-2021.html (accessed on 31 August 2022).

32. IBGE (Instituto Brasileiro de Geografia e Estatística). 2022 Censo 2022 já contou quase 60 milhões de pessoas no país | Agência de Notícias. *Agência Notícias - IBGE*. See https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/34728-censo-2022-ja-contou-quase-60-milhoes-de-pessoas-no-pais (accessed on 31 August 2022).

33. Liu, X., Huang, Y., Xu, X., Li, X., Li, X., Ciais, P., Lin, P., Gong, K., Ziegler, A.D., Chen, A., Gong, P. 2020 High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. *Nat. Sustain.* **3**, 564–570.

34. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V. 2014 On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716.

35. IBGE. 2010 Educação e deslocamento: resultados da amostra. *Rio Jan. IBGE*

36. IBGE. 2008 *Regiões de influência das cidades 2007*.

37. Freitas VL de S, Konstantyner TCR de O, Mendes JF, Sepetauskas CS do N, Santos LBL. 2020 The correspondence between the structure of the terrestrial mobility network and the spreading of COVID-19 in Brazil. *Cad. Saúde Pública* **36**, e00184820.

38. Simini F, González MC, Maritan A, Barabási A-L. 2012 A universal model for mobility and migration patterns. *Nature* **484**, 96–100.

39. Wesolowski A, O'Meara WP, Eagle N, Tatem AJ, Buckee CO. 2015 Evaluating Spatial Interaction Models for Regional Mobility in Sub-Saharan Africa. *PLOS Comput. Biol.* **11**, e1004267.

40. Mordecai, E.A., Caldwell, J.M., Grossman, M.K., Lippi, C.A., Johnson, L.R., Neira, M., Rohr, J.R., Ryan, S.J., Savage, V., Shocket, M.S., Sippy, R. 2019 Thermal biology of mosquito-borne disease. *Ecol. Lett.* **22**, 1690–1708.

41. Lowe, R., Lee, S.A., O'Reilly, K.M., Brady, O.J., Bastos, L., Carrasco-Escobar, G., de Castro Catão, R., Colón-González, F.J., Barcellos, C., Carvalho, M.S., Blangiardo, M. 2021 Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *Lancet Planet. Health* **5**, e209–e219.

42. Wu, X., Lang, L., Ma, W., Song, T., Kang, M., He, J., Zhang, Y., Lu, L., Lin, H., Ling, L. 2018 Non-linear effects of mean temperature and relative humidity on dengue incidence in Guangzhou, China. *Sci. Total Environ.* **628**, 766–771.

43. Wells N, Goddard S, Hayes MJ. 2004 A self-calibrating Palmer drought severity index. *J. Clim.* **17**, 2335–2351.

44. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci.* **106**, 21484–21489.

45. Lana RM, Gomes MF da C, de Lima TFM, Honório NA, Codeço CT. 2017 The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: A network-based analysis. *PLoS Negl. Trop. Dis.* **11**.

46. Castro, M.C., Kim, S., Barberia, L., Ribeiro, A.F., Gurzenda, S., Ribeiro, K.B., Abbott, E., Blossom, J., Rache, B., Singer, B.H. 2021 Spatiotemporal pattern of COVID-19 spread in Brazil. *Science* **372**, 821–826.

47. Nicolelis MA, Raimundo RL, Peixoto PS, Andreazzi CS. 2021 The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil. *Sci. Rep.* **11**, 1–12.

48. Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM, Eskew EA, Olival KJ, Ross N, Bansal S. 2022 Climate change increases cross-species viral transmission risk. *Nature* , 1–8.

49. Baker, R.E., Mahmud, A.S., Miller, I.F., Rajeev, M., Rasambainarivo, F., Rice, B.L., Takahashi, S., Tatem, A.J., Wagner, C.E., Wang, L.F., Wesolowski, A. 2022 Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205.

Appendix A: Spatial modelling with empirical Bayes

A.1 Introduction

Empirical Bayesian approaches use data to inform the prior distribution, with the prior being estimated as part of the model fitting process. This removes the requirement of having to fully specify prior beliefs before model fitting, making it particularly appealing in spatial statistical modelling where the spatial structure of the data may not be fully understood. In this section, I explain how penalised smoothing splines, in particular thin plate regression splines, can be used to produce a spatial smooth model, estimated using empirical Bayesian methods. I then show how these methods can be extended to more complex model structures and fitted using a fully Bayesian approach via Markov chain Monte Carlo (MCMC) methods.

A.2 Empirical Bayes with generalised additive models

Generalised additive models (GAMs) are statistical models that include smooth functions of covariates to allow flexibility in the nature of the relationships between an outcome and explanatory variables [1]. The original GAM was defined as:

$$\eta(y_i) = \alpha_o + \sum_{j=1}^{n} \alpha_j x_{ij} + \sum_{k=1}^{m} f_k(z_i),$$

where η is some link function determined by model choice, **y** are outcome variables, **x** and **z** are observed covariates, $\boldsymbol{\alpha}$ are unknown regression coefficients, and $\boldsymbol{f}$ are some smooth functions to be estimated. To introduce the concepts underlying GAMs and smoothing functions, I will first focus on a linear model containing a single, univariate smooth function:

$$y_i = f(z_i) + \varepsilon_i. \quad (1)$$

Expansions to models with multiple smooths or functions applied to multiple covariates will be introduced in later sections.

The inclusion of smooth functions allows for more flexible model specifications and the inclusion of nonlinear relationships between the outcome and explanatory variables. However, this additional flexibility comes at a cost: the structure of *f* must be determined and the degree

of ‘smoothness’ much be defined. The first assumption we make about this function is that it is smooth. One mathematical way to categorise this is that any function *f* is smoother than a function *g* if $\int f''(z)^2 dx < \int g''(z)^2 dz$. Therefore, we want a function that will minimise the integrated square second derivative (also known as a cubic spline penalty) [2]. This can be included in the model as a penalty. However, even with this penalty, the best fitting model will likely be one that interpolates the data. To avoid this, a smoothing penalty parameter is introduced and, using a penalised least squared approach, the function we seek to minimise is:

$$\|y - f(z)\|^2 + \lambda \int f''(z)^2 dz, \quad (2)$$

where λ is the smoothing penalty parameter aiming to control smoothness and ensure the smooth function is not too ‘wiggly’ (i.e. the function does not simply interpolate between data points) [2,3]. $\lambda \to \infty$ would produce a straight line and $\lambda = 0$ would lead to an unpenalized piecewise linear regression coefficient [4].

Many smooth functions used in GAMs are constructed using smoothing splines and can be described as linear combinations of (known) basis functions, b_j (functions applied to the covariate(s) at given intervals, determined by the type of smoothing spline chosen), multiplied by (unknown) regression coefficients, β_j [3]. When considering a univariate smoothing spline, $f(z)$, this can be expressed:

$$f(z) = \sum_{j=1}^{K} \beta_j b_j(z), \quad (3)$$

where *K* is the number of knots, or turning points, in the function. The number of knots should be large enough to capture patterns in the data but not so large it leads to excessive computational cost or overfits the data [2,3]. Assuming the number and value of basis functions are known and fixed, the estimation problem (2) can be rewritten as a function of the unknown coefficients, $\boldsymbol{\beta}$. Given $f(z) = \boldsymbol{\beta}^T \boldsymbol{b}(z)$, it follows that $f''(z) = \boldsymbol{\beta}^T \boldsymbol{b}''(z)$ and $f''(z)^2 = \boldsymbol{\beta}^T \boldsymbol{b}''(z) \boldsymbol{b}''(z)^T \boldsymbol{\beta}$. So, the smoothing penalty becomes $\int f''(z)^2 dx = \boldsymbol{\beta}^T \int \boldsymbol{b}''(z) \boldsymbol{b}''(z)^T dz \, \boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$ and the estimation problem (2) can be rewritten:

$$\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}, \quad (4)$$

where $\boldsymbol{Z} = \boldsymbol{b}(\boldsymbol{z})$ are the basis functions, $S_{ij} = \int b_i''(z) b_j''(z) dz$ and $\boldsymbol{S}$ can be thought of as a penalty matrix. By solving this expression, the penalised least squared estimator of the coefficients, $\widehat{\boldsymbol{\beta}}$, is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T \boldsymbol{Z} + \lambda \boldsymbol{S})^{-1} \boldsymbol{Z}^T \boldsymbol{y}. \quad (5)$$

The assumption that the spline is more smooth than wiggly can be viewed from a Bayesian perspective as a prior belief, with $\widehat{\boldsymbol{\beta}}$ representing the posterior mode of $\boldsymbol{\beta}$. An improper multivariate normal prior is used here as the penalised least squares estimate of $\boldsymbol{\beta}$ (Equation 5) is also the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}|\boldsymbol{y}$ where $\boldsymbol{\beta} \sim N(0, \sigma^2 \boldsymbol{S}^- / \lambda)$ and $\boldsymbol{S}^-$ is the generalised inverse of the penalty matrix [2,3]. The Bayesian posterior distribution of $\boldsymbol{\beta}$ is then:

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\widehat{\boldsymbol{\beta}}, (\boldsymbol{Z}^T \boldsymbol{Z} + \lambda \boldsymbol{S})^{-1} \sigma^2\right). \quad (6)$$

This interpretation gives the model the same structure as linear mixed models. Therefore, the resulting smooth functions can be interpreted in the same way as traditional random effects [3]. This also means that the parameters σ and λ can be estimated using restricted maximum likelihood (REML) [4]. As the prior and posterior distributions of $\boldsymbol{\beta}$ depend on the smoothing parameter λ which is estimated using the data, this approach is known as an empirical Bayesian approach and is justified by large sample approximation.

A.3 Thin plate regression splines

The empirical Bayesian approach introduced in the previous section assumes that basis functions defining the smooth spline $f(z)$ in Equation (3) are known and fixed. However in practice, the type of splines b_j, the number of them K and the position of the knots along the covariate range space need be defined prior to model fitting. There are many types of smoothing splines that can be used in GAMs to explore nonlinear relationships between an outcome and one or more explanatory variables (see [3] for some examples). The choice of spline depends on the nature of the relationship between the outcome and explanatory variables, and the number of explanatory variables being considered. Each spline has a

different definition of basis functions ($b_j(z)$ from Equation 3) and their associated penalty matrices, $\boldsymbol{S}$.

Thin plate splines are a general class of smoothing splines that are incredibly flexible and can be applied to multiple variables [3,5]. Thin plate splines can be used to estimate smooth functions $\boldsymbol{f}$ by minimising the equation:

$$\|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda J_{md}(\boldsymbol{f}), \tag{7}$$

where λ is the smoothing parameter introduced previously, $\boldsymbol{f}$ is a vector of basis functions applied to d covariate(s) ($\boldsymbol{z} = [\omega_1, \dots, \omega_d]^T$), $J_{md}(\boldsymbol{f})$ is a penalty function measuring the wiggliness of $\boldsymbol{f}$, m is the order of differentiation in this penalty and can be any integer satisfying $2m > d$. However, for visually smooth results it is preferable that $2m > d + 1$, and often m is set to the minimum that satisfies this condition [3]. The wiggliness penalty is defined as

$$\boldsymbol{J}_{md} = \int \dots \int \sum_{\nu_1 + \dots + \nu_d = m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\frac{d^m \boldsymbol{f}}{d\omega_1^{\nu_1} \dots d\omega_d^{\nu_d}} \right)^2 d\omega_1 \dots d\omega_d.$$

Note that the penalty matrix in Equation (2) is an example of a thin plate spline applied to a single covariate ($d = 1$) where $m = 2$. One of the major benefits of using thin plate splines is that knot positions and basis functions arise naturally due to the mathematical properties of the smoothing penalty and do not have to be specified by the user (see [3] and [5] for full specification of these basis functions). If we take the two-dimensional case ($d = 2$, covariates ω_1 and ω_2) and the minimum value of m to satisfy the visually smooth condition ($m = 2$), then the smoothing penalty becomes:

$$\boldsymbol{J}_{22} = \int \int \left(\frac{d^2 f}{d\omega_1^2} \right)^2 + 2 \left(\frac{d^2 f}{d\omega_1 d\omega_2} \right)^2 + \left(\frac{d^2 f}{d\omega_2^2} \right)^2 d\omega_1 d\omega_2.$$

In this case, the basis function functions minimising Equation (7) has the form

$$\hat{f}(\boldsymbol{z}) = \sum_{i=1}^{n} \delta_i \gamma(\|\boldsymbol{z} - \boldsymbol{z}_i\|) + \sum_{j=1}^{3} \alpha_j \phi_j(\boldsymbol{z}),$$

where δ_i and α_j are coefficients to be estimated, $\phi_j(\boldsymbol{z})$ are linearly independent polynomial functions, and $\gamma(r) = \frac{-1}{8\pi} r^2 \log(r)$. The ϕ_j functions span the null space of the $\boldsymbol{J}_{md}$ penalty and are considered completely smooth (where $d = m = 2,\ \phi_1(\boldsymbol{z}) = 1,\ \phi_2(\boldsymbol{z}) = \omega_1$ and $\phi_3(\boldsymbol{z}) = \omega_2$). By defining the matrix $\boldsymbol{E}$ as $E_{ij} = \gamma(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ and $\boldsymbol{T}$ as $T_{ij} = \phi_j(\boldsymbol{z}_i)$, the spline fitting problem (7) requires the following equation to be minimised:

$$\|\boldsymbol{z} - \boldsymbol{E\delta} - \boldsymbol{T\alpha}\|^2 + \lambda \boldsymbol{\delta}' \boldsymbol{E\delta}. \tag{8}$$

The main drawback of using full thin plate splines is that they are extremely computationally intensive (they require one parameter per observation plus an additional smoothing parameter λ). To overcome this, thin plate regression splines were developed as truncated versions of full thin plate splines which reduce the computational cost of model fitting by substituting $\boldsymbol{E}$ in Equation (8) with a rank deficient approximation [3,5]. It is important to note that thin plate regression splines are isotropic in nature, i.e. they smooth equally with respect to each covariate [3,5]. This makes them inappropriate where covariates included in the smoother are measured on a different scale (e.g. when considering spatio-temporal relationships). For these instances, another choice of smoother would be required.

A.4 Bayesian spatially smoothed generalised additive models

When considering spatial data, thin plate regression splines can be applied to geographical coordinates (longitude, latitude) to produce a spatially smoothed surface. These can be incorporated into generalised additive models as an alternative to traditional random effect terms where spatial autocorrelation between close areas is present in the data. This spatially smoothed model takes the form:

$$\eta(y_i) = \alpha_o + \sum_{j=1}^{n} \alpha_j x_{ij} + f_{spat}(a_i, b_i), \tag{9}$$

where η is a link function determined by model choice, **y** are outcome variables, **x** are observed covariates, $\boldsymbol{\alpha}$ are unknown regression coefficients, and $f_{spat}(\boldsymbol{a}, \boldsymbol{b})$ is a thin plate regression spline applied to coordinates $\boldsymbol{a}, \boldsymbol{b}$.

This model can be fit using the R package `mgcv` [3,6]. By default, this package uses thin plate regression splines to define smooth functions (although alternatives are available) and estimates the degree of spline smoothness (defined by the smoothing parameter λ) using generalised cross-validation by default. The empirical Bayesian interpretation of smoothing splines as random effects introduced earlier require the smoothing parameter to be estimated using REML, this can be achieved by adding the argument `method = "REML"` to the smooth function definition.

Users must specify an upper bound for the number of basis dimensions (k from Equation 3). Although there is no set rule for how to define this upper limit, it should be large enough to ensure the smooth represents the data adequately but small enough to ensure computational efficiency [3]. Often this choice is arbitrary and in practice, unless the dimensions are set restrictively small, this choice will only have a small impact on the model fit (the actual flexibility of the smooth is mainly controlled by the λ parameter). Informal checks performed using the `gam.check` function in the `mgcv` package help to determine whether the basis dimension is adequate [6]. This function estimates the residual deviance along the covariates of the smooth and compares it to close values. If there is little or no difference, this suggests that the function is too smooth and the maximum basis dimensions should be increased [3]. The output includes a p-value testing the difference between residual deviance estimates, generated using simulations by randomly resampling from the model results at different covariate values. If the p-value is low, this indicates a small difference, and the maximum number of basis dimensions should be increased.

Given the smoothing parameter has been estimated using REML, this model can be viewed from an empirical Bayesian perspective. Following from Equation 6, the posterior distribution of $\boldsymbol{\beta}$ is assumed to follow a multivariate Normal distribution with mean $\widehat{\boldsymbol{\beta}}$ and precision matrix proportional to the smoothing penalty $\boldsymbol{S}_{\lambda}$ (assuming the prior distribution of $\boldsymbol{\beta}$ takes a zero-mean multivariate Gaussian prior). As these values are estimated as part of the model fitting process, this distribution is fully defined, and simulations can be performed to generate estimates or credible intervals of the coefficients via bootstrapping. Estimates of the coefficients can be combined with the basis functions used to generate the smooth surface and can be visualised to show the spatial structure of the data. An example of this approach can be found in Chapter 2.

A.5 Multidimensional scaling

Multidimensional scaling (MDS) is a set of mathematical procedures that aim to convert measures of pairwise 'distances' or dissimilarity between data points into a set of points mapped onto an abstract coordinate space [7]. One common example of MDS is principal component analysis (PCA).

Metric or classical scaling aims to return coordinates x_i, x_j based on dissimilarities d_{ij}. Here, let $\boldsymbol{X}$ be a matrix containing the coordinate values, and $\boldsymbol{D}$ be a matrix of squared dissimilarities (where $D_{ij} = d_{ij}^2$). If $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$, then $d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs}$. Therefore, given matrix $\boldsymbol{D}$ is known, this process can be inverted and $\boldsymbol{X}$ can be calculated by factorising the matrix $\boldsymbol{B}$ [8]. Note that one of the major assumptions of MDS is that the measure of dissimilarities d_{ij} are symmetrical ($d_{ij} = d_{ji}$).

In Chapter 5 of this thesis, I sought to project the measure of 'connectivity' between municipalities based on regular commuting onto a two-dimensional abstract cartesian space. As MDS requires a measure of dissimilarity d_{ij}, the measure of connectivity (the proportion of residents of municipality *i* regularly travelling to municipality *j* for work or education, c_{ij}) was converted using $d_{ij} = 1 - c_{ij}$.

A.6 Fully Bayesian simulations of generalised additive models

The empirical Bayesian approach introduced in Section A.4 allows simulations from the posterior distribution of the smooth functions which can be used to generate credible intervals (via bootstrapping) and estimates of the spatial smooth function. However, there are limits to the complexity of the models that can be generated within the `mgcv` package. For example, it would not be possible to extend the spatial smooth model in Equation 9 to include an unstructured random effect that accounts for unobserved heterogeneity between areas. The `jagam` function has been developed to allow models with complex smooth structures (as can be defined in `mgcv`) and random structures (as can be defined in BUGS-based programmes) to be fit using a fully Bayesian MCMC approach [9]. This combines the flexibility of model specification available in BUGS language with the flexible structures available in `mgcv`.

To illustrate this approach, we will consider the model presented in Chapters 4 and 5 of this thesis that contain two spatially smoothed random terms and an unstructured random effect to account for spatial heterogeneity. The spatially smoothed terms are created by applying thin plate regression splines to sets of coordinates describing the relative 'connectedness' of areas. The first assumes connectivity based on distance, i.e. areas close together are more similar in terms of the outcome, and is defined using geographical lat-lon coordinates. The second assumes connectivity based on the number of people regularly commuting between areas and uses coordinates created using MDS (see Section A.5):

$$\eta(y_i) = \alpha_o + \sum_{j=1}^{n} \alpha_j x_{ij} + f_{spat}(a_i, b_i) + f_{spat}(c_i, d_i) + v_i.$$

The thin plate regression splines are generated using the `jagam` function in `mgcv` which produces estimates of the basis functions, basis coefficients and the smoothing penalties. Note that the maximum value of k must still be defined in these functions and should be checked as described in Section A.4. These values are then extracted and used to define the prior distributions of the basis coefficients in the MCMC, assuming improper zero mean Normal distributions.

A.7 Conclusion

The empirical Bayesian approach presented in this section is a flexible, computationally efficient alternative to fully Bayesian approaches such as MCMC where sufficient data are available. Prior distributions are determined indirectly through the specification of each spline rather than having to choose a spatial structure subjectively. With splines, the spatial structure is estimated from the data objectively which is particularly appealing for spatial analysis where the spatial structure of the data may not be fully understood. Estimates of the smooth functions can be used to inform future analysis and generate hypotheses about underlying spatial structures in the data. This approach reduces some of the subjectivity involved in Bayesian analysis as users are not required to pre-specify the prior distribution. In addition, the empirical Bayesian approach in Section A.4 does not suffer issues with convergence and reduces the computational burden which can be an issue for simulation-based approaches, such as MCMC.

However, an empirical Bayesian approach is not fully Bayesian as the choice of prior distributions is restricted and the inferential approximations underlying the computation of the

posterior distributions relies on large sample approximations. The approach presented in Section A.4 is also limited in terms of the complexity of the model structure that can be specified. Models with complex spatial structures involving multiple sources of spatial connectivity require MCMC to fit the models, increasing the computational cost of the model fitting process.

Despite these limitations, empirical Bayesian approaches offer a flexible approach where the underlying structure of the data are not known. Future work could explore alternative specifications of the posterior distribution beyond multivariate normal. Additionally, the sensitivity analysis presented in Chapter 4 showed that spatial models fitted using the approach in Section A.6 performed as well as the current conventional approach fitted using MCMC or integrated nested Laplace approximations (INLA).

References

1. Hastie, T, Tibshirani, R. 1986 Generalized Additive Models. *Stat. Sci.* **1**, 297–318.
2. Wood SN. 2020 Inference and computation with generalized additive models and their extensions. *Test* **29**, 307–339.
3. Wood SN. 2017 *Generalized additive models: an introduction with R*.
4. Wood SN. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36. (doi:10.1111/j.1467-9868.2010.00749.x)
5. Wood SN. 2003 Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**, 95–114.
6. Wood S, Wood MS. 2015 Package 'mgcv'. *R Package Version* **1**, 729.
7. Cox MA, Cox TF. 2008 Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347.
8. Mead A. 1992 Review of the development of multidimensional scaling methods. *J. R. Stat. Soc. Ser. Stat.* **41**, 27–39.
9. Wood SN. 2016 Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75**, 1–15. (doi:10.18637/jss.v075.i07)

Appendix B: Supplementary Material Chapter 2

B.1 Supplementary text

Supplementary information to support Chapter 2: *The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil*. Contains additional information about methods and materials used in the manuscript and the results of sensitivity analyses. Taken from https://doi.org/10.1371/journal.pntd.0009773.

B.1.2 Methods and materials

B.1.2.1 Dengue surveillance and outbreak definitions in Brazil

Monthly dengue case data are freely available from Brazil's Notifiable Diseases Information System (SINAN), via the Health Information Department, DATASUS (https://datasus.saude.gov.br/informacoes-de-saude-tabnet/). Although notification of a suspected dengue case is mandatory in Brazil, the surveillance system is predominantly passive, which means that many mild and asymptomatic cases may be missed. One investigation of the Brazilian dengue surveillance system estimated that there were 12 actual infections per reported case overall, which rose to over 17 in periods of high incidence [1].

Rather than use dengue case data which differs in accuracy between regions, and between epidemic and non-epidemic periods, we aggregated the cases by year and converted them into a binary outbreak indicator where cases exceeded some outbreak threshold. Several methods have been used to define outbreak thresholds Brazil, including a monthly moving average, where historical data are used to estimate the expected number of cases within a region [2,3], and a fixed threshold based on the dengue incidence rate (DIR), defined by the Brazilian Ministry of Health as the number of cases per 100,000 residents [4]. We chose to use a fixed threshold approach as the mean incidence was heavily influenced by outbreak years, making the probability of detecting an outbreak inconsistent between municipalities. Our primary analysis used an outbreak threshold of more than 300 cases per 100,000 residents, defined as 'high risk' by the Brazilian Ministry of Health. We also tested a 'medium risk' indicator, defined as more than 100 cases per 100,000 residents [4].

The annual DIR was calculated using estimates of the annual population for each municipality obtained from the Brazilian Institute of Statistics and Geography (IBGE) via DATASUS (https://datasus.saude.gov.br/populacao-residente). As an alternative, we used the 75th percentile of the DIR per municipality with a minimum threshold value equivalent to 5 cases per year to avoid very low cases triggering an outbreak in 'protected' areas. The 75th percentile of the DIR was calculated using all available data from 2001 - 2020 for each municipality. Many municipalities in previously 'protected' areas such as South Brazil and the western Amazon had lower thresholds using this method than the fixed thresholds used by the Brazilian Ministry of Health (Fig B1). However, the threshold was much higher (up to a maximum of DIR = 3275) in regions which had experienced high levels of dengue transmission in the past.

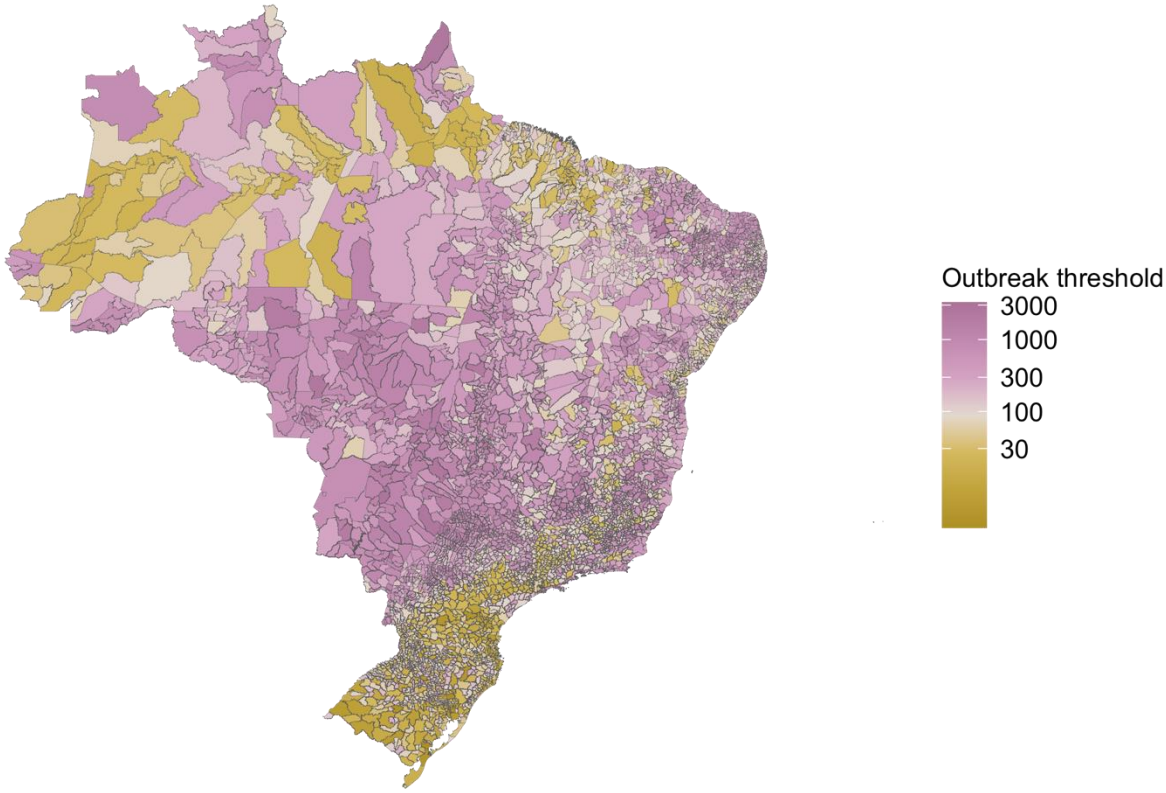


Fig B1: The outbreak threshold for each municipality based on the 75th percentile of dengue incidence rates between 2001 - 2020. Regions with historically low dengue transmission, such as South Brazil, had thresholds below 100 (shown in gold), whilst areas with sustained high transmission such as the Centre-West had much higher thresholds, up to a maximum of 3275. Maps were produced in R using the geobr package [5,6] (hsttps://ipeagit.github.io/geobr/).

B.1.2.2 Hydrometeorological factors

In addition to temperature suitability, hydrometeorological conditions such as precipitation and drought have been linked to dengue transmission. Prior studies have found that the risk of dengue increases immediately following extremely wet conditions [7,8], however the level of precipitation considered extreme varies greatly across Brazil between climate systems. To measure the relative wetness of municipalities, we used the self-calibrating Palmer Drought Severity Index (scPDSI). The scPDSI was obtained from the Climate Research Unit gridded Time Series (v4.05) [9,10] for the period January 2001 - December 2020, at a spatial resolution of 0.5° x 0.5°. The PDSI is a widely used measure of meteorological drought ranging from -

10 (dry) to 10 (wet) compared to 'normal conditions', with values below -4 and above 4 considered extreme [11,12]. The scPDSI calibrates this index to the 'normal conditions' for each location of interest separately, providing a more spatially comparable measure [10,13]. The scPDSI was aggregated to each municipality using the exactextractr package [14] in R (version 4.0.3) by calculating the mean of the grid boxes lying within each municipality. Grid boxes partially covered by a municipality were weighted by the percentage of area that lay within the municipality.

Most states, particularly those in North Brazil, have experienced increasingly severe drought conditions in recent years. However, there have been several extremely wet events, particularly in the Southeast of the country (Fig B2). To understand the relationship between wet conditions and dengue outbreaks, we calculated the number of months per year each municipality took an scPDSI value of 4 and above, considered 'extremely wet' by the scPDSI [10,12]. On average, the number of months considered extremely wet has increased in parts of South Brazil and in Pará, North Brazil, and has reduced in the Amazon and Southeast Brazil (Fig B3).

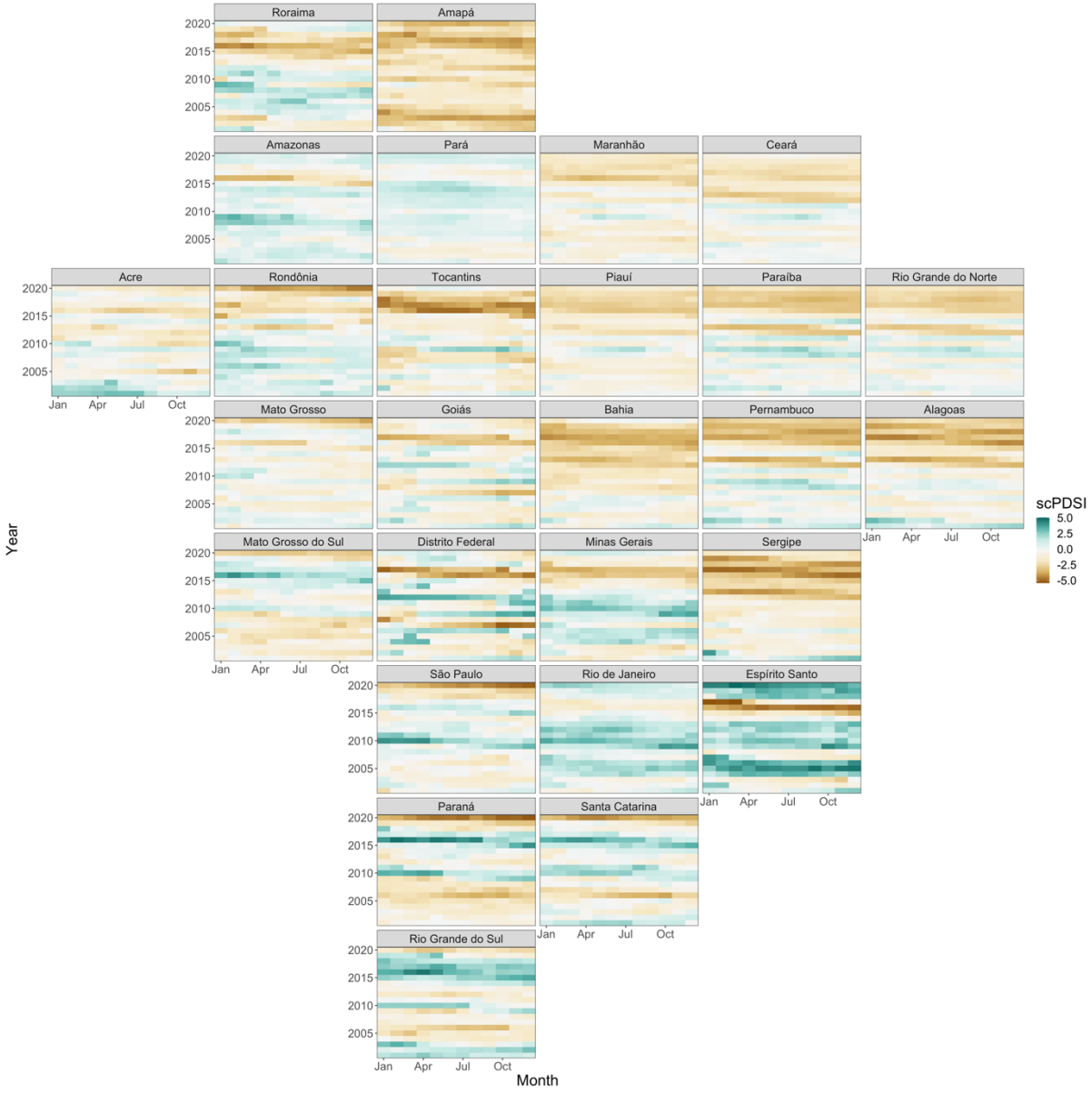


Fig B2: The average monthly self-calibrated Palmer Drought Severity Index (scPDSI) per state from January 2001 - December 2020. Values below -4 (shown in brown) are considered extremely dry compared to normal conditions, whereas values above 4 (shown in blue) are considered extremely wet. The north and east of Brazil has experienced increasingly severe droughts in recently years, in contrast states in the Southeast have experienced a number of extremely wet conditions.

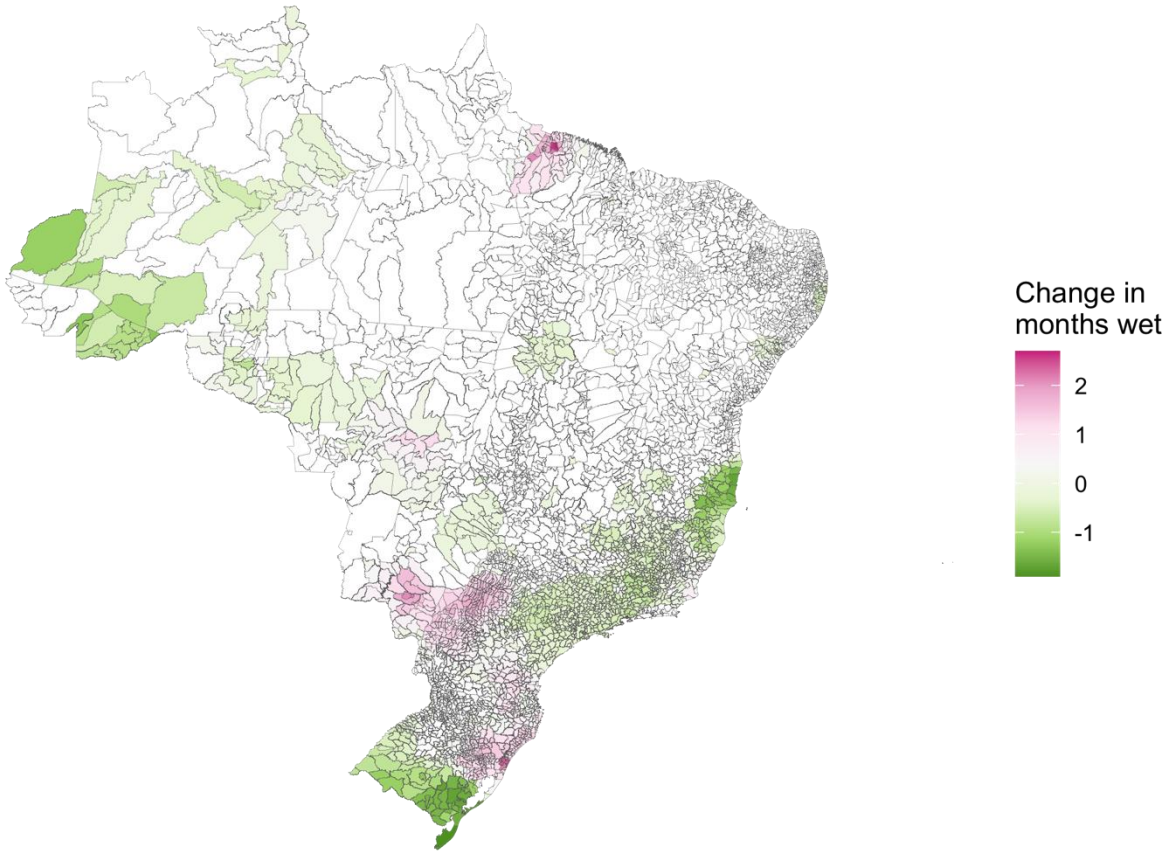


Fig B3: Map showing the difference in the average number of months per year considered extremely wet (scPDSI > 4) between 2001 - 2010 and 2011 - 2020. The number of months considered extremely wet has increased on average in parts of South and North Brazil (shown in pink). In comparison, the number of extremely wet months per year in the western Amazon and parts of South and Southeast Brazil have reduced. Maps were produced in R using the geobr package [5,6] (https://ipeagit.github.io/geobr/).

B.1.1.3 Socioeconomic factors

We obtained information about the percentage of residents in each municipality living in urban areas, the percentage with access to the piped water system, and the percentage that had refuse collected (either privately or using the municipal service) from the 2000 and 2010 censuses via DATASUS. Despite Brazil having the largest economy in South America, it has been the most unequal since 2015 [15]. Access to basic services differs greatly across the country and the

traditionally wealthier regions in the South and Southeast have almost universal coverage at the municipality level in contrast to rural parts of the North and Northeast which had little or no access, even in 2010. We found that the level of urbanisation was highly correlated to access to piped water (Fig B4, $r = 0.656$, 95% confidence interval: [0.641, 0.671], $p < 0.001$) and refuse collection (Fig B4, $r = 0.794$, 95% confidence interval: [0.784, 0.804] $p < 0.001$) when aggregated to the municipality level. Therefore, access to piped water and refuse collection were not included in the models as they were not useful at explaining the differences within cities at this level of aggregation and would likely introduce multicollinearity into the model.

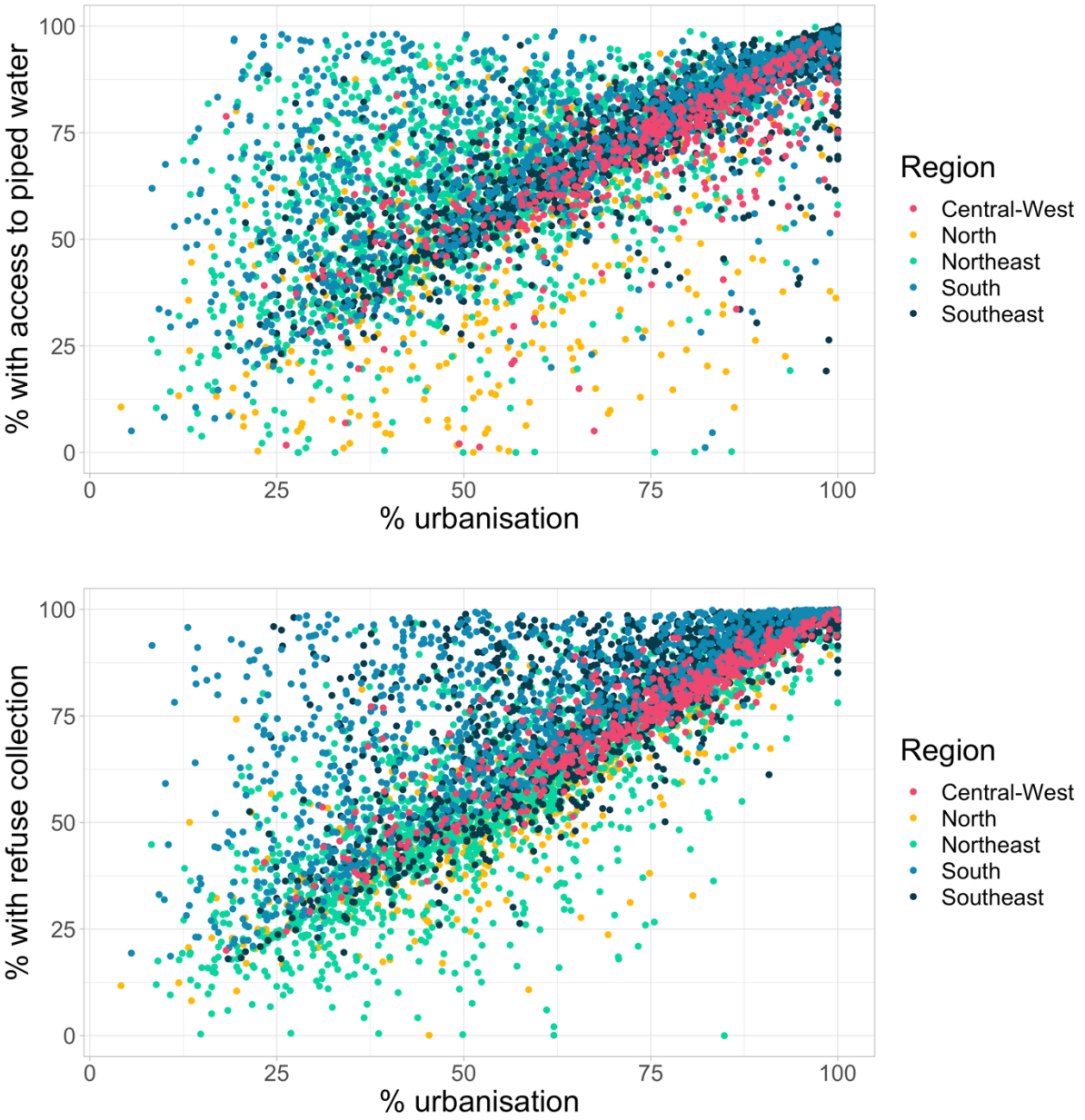


Fig B4: Scatterplot comparing the percentage of residents with access to piped water (top) and refuse collection (bottom) to the percentage living in urban areas from the 2010 census. Access to basic services was highly correlated to the level of urbanisation: highly urban areas had highest access to piped water and refuse collection.

B.1.1.4 Hierarchical levels of influence of cities

We extracted the level of influence of cities from the Regions of Influence of Cities ("Regiões de Influência das Cidades", REGIC) studies carried out by IBGE in 2007 and 2018 [16,17] to use as a proxy for human movement within our models. REGIC aims to recreate the complex urban network of Brazil using information from surveys about the frequency and reasons for the movement of people and goods around the country. The level of influence assigned to each city was based on the number of people travelling to the city but also the number of important institutions that attracts the movement of people from outside the city, such as hospitals, universities, business centres, government agencies, and cultural centres (such as theatres and shopping centres). Cities were classified into five levels:

1. Metropolis: the largest cities in Brazil, with strong connections throughout the entire country. This includes São Paulo, the capital Brasilia, and Rio de Janeiro.
2. Regional capital: large cities which are connected throughout the region in which they are located and to metropoles. This includes state capitals that were not classified as metropoles, such as Rio Branco, Campo Grande and Porto Velho.
3. Sub-regional capital: cities with a lower level of connectivity, mostly connected locally and to the three largest metropoles.
4. Zone centre: smaller cities with influences restricted to their immediate area, often neighbours.
5. Local centre: the smallest cities in the network which typically only serve residents of the municipality and are not connected elsewhere.

There were 12 metropoles, consisting of 203 municipalities, according to the 2007 REGIC study: São Paulo, Rio de Janeiro, Brasilia, Manaus, Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Curitiba, Goiânia and Porto Alegre. In 2018, this increased to 15 metropoles, consisting of 214 municipalities, as Campinas, Florianópolis and Vitória were re-classified from regional capitals to metropoles. The number of regional capitals and sub-regional centres also increased between 2007 and 2018 from 70 to 97 and from 169 to 352 respectively. The number of lower-level cities, zone centre and local centres, both decreased from 556 to 398, and from 4473 to 4037 (Appendix B.3, S1 Table). The distribution of highly connected urban centres is uneven across the country; the South and Southeast regions are particularly well connected, while the North and Northeast contain fewer high-level centres (Chapter 2, Fig 3 and Appendix B.3, S1 Table). The proportion of higher-level centres has increased in each

region of Brazil, although the Amazon rainforest remains less connected than other areas (Fig B5). Metropoles, regional capitals and sub-regional centres had higher levels of urbanisation, access piped water and refuse collection on average than less connected centres (Appendix B.2 Fig S6).

a) **b)**

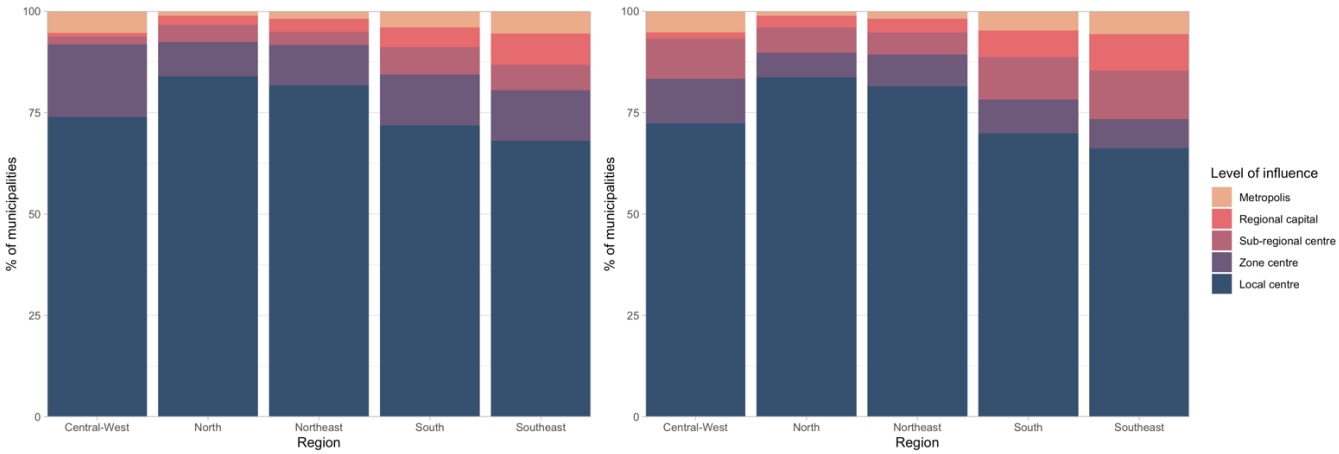


Fig B5: The proportion of cities in each region at each level of influence in the a) 2007 and b) 2018 REGIC study. The proportion of high-level cities has increased across the country but the North and Northeast still have noticeably less well-connected cities than other regions. The Southeast and South are by far the most connected regions.

B.1.1.5 Modelling approach

We formulated a spatio-temporal generalised additive model (GAM) to quantify the relationship between temperature suitability, level of connectivity and socioeconomic conditions on the odds of a municipality experiencing an outbreak. The response variable was a binary outbreak indicator defined as an annual dengue incidence rate of more than 300 cases per 100,000 residents. To account for spatial and temporal patterns in the data, smooth functions of the year and the coordinates of the centroids of municipalities were included in the model. We used thin plate regression splines to represent the smooth (2D) function of the coordinates. Thin plate splines are data-driven and estimate the best fitting function for the data [18]. To account for changes in spatial patterns over the period, we also included a space-time interaction term created by applying a tensor product smooth to the coordinates and the year.

Tensor product smooths allow interactions between variables that are measured on different scales (in this case, space and time). The final model equation was as follows:

$$Y_{it} \sim Bernoulli(p_{it})$$

$$logit(p_{it}) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{jit} + f_{spat}(lon_i, lat_i) + f_{time}(t) + f_{int}(lon_i, lat_i, t)$$

Where Y_{it}, binary outbreak indicator for municipality *i* (*i = 1, ..., 5,560*) in year *t* (*t = 2001, ..., 2020*), is expected to follow a Bernoulli distribution defined by p_{it}, the probability of an outbreak. The Bernoulli distribution is a special case of the binomial distribution where the number of trials is equal to 1. β_j are coefficient estimates associated with covariates X_{jit}. $f_{spat}(lon_i, lat_i)$ is the spatial smooth field based on the coordinates (lon_i, lat_i) of the centroid of municipality *i*, $f_{time}(t)$ is the temporal smooth function applied to year *t* and $f_{int}(lon_i, lat_i, t)$ is the spatio-temporal interaction term. This model is a type of structured additive regression (STAR) model which allows for Bayesian interpretations of additive models by specifying prior beliefs on the smooth functions [18,19].

We chose to include the number of months with temperature suitable for dengue transmission, the proportion of residents living in urban areas, the level of influence from the REGIC study, and the prior outbreak indicator as covariates in our final model to address our initial research questions. We also tested the number of months considered extremely wet according to the scPDSI as hydrometeorological factors are also recognised as important drivers of dengue transmission [7,8]. Although this coefficient was statistically significant (adjusted odds ratio: 1.11, 95% credible interval: 1.09, 1.14), the model fit was not improved with the addition of this covariate (S3 Table) and the covariate was excluded from further analysis.

Inference was performed using an empirical Bayesian approach with estimates calculated using restricted maximum likelihood (REML), an approach that has been shown to give more stable estimates than generalised cross validation [20], and more accurate estimates than a full Bayesian approach for binomial models (a simulation study showed increased coverage probability and reduced bias for the empirical Bayesian approach compared to fully Bayesian) [19]. We used the mgcv package in R [18] to fit the spatio-temporal models and to simulate

from the posterior distributions of the coefficients to produce mean estimates and 95% credible intervals.

B.2 Results

B.2.1 Outbreak threshold comparisons

To check whether our model results were robust to the definition of an outbreak, we compared our primary results to alternative outbreak indicators: over 100 cases per 100,000 residents (considered medium risk by the Brazilian Ministry of Health [21]), and above the 75th percentile of the yearly DIR between 2001 - 2020 for each municipality with a minimum threshold set as over 5 cases per year. Although the models agreed that the odds of an outbreak were significantly increased in highly connected, highly urbanised cities that had previously experienced an outbreak and had a suitable temperature, the coefficient estimates differed (Appendix B.4, S2 Table). The 75th percentile model had noticeably lower coefficient estimates for each parameter compared to the fixed threshold models (Appendix B.2, Fig S8). Most credible intervals for the coefficient estimates of the fixed threshold models overlapped, however the odds of experiencing an outbreak in municipalities that had previously experienced one was higher in the model using the DIR = 100 threshold (aOR: 2.42, 95% CI: 2.31, 2.56) compared to the DIR = 300 threshold (aOR: 2.03, 95% CI: 1.93, 2.15).

We assessed the model fit of these alternative outbreak threshold models using a receiver operating characteristic (ROC) curve which plots the true positive rate against the true negative rate at different thresholds to test the predictive ability of the model. The area under the ROC curve was calculated as this gives a measure of predictive ability compared to chance, which would return a value of 0.5. The closer the area under the ROC curve is to 1, the better the model fits the data. Care should be taken when data are imbalanced as the ROC curve can overestimate the accuracy of a model. The predictive ability of models were also compared using the Brier score [22]. The Brier score is the mean squared difference between the observed and expected outcomes; a lower Brier score represents a better fitting model. We found that the fixed threshold models fit the data better according to the ROC curve (Appendix B.2, Fig S7 and Appendix B.5, S3 Table), the Brier score also showed that these models had a better predictive ability than the 75th percentile model (Appendix B.5, S3 Table).

References

1. Silva MMO, Rodrigues MS, Paploski IAD, Kikuti M, Kasper AM, Cruz JS, et al. Accuracy of Dengue Reporting by National Surveillance System, Brazil. Emerg Infect Dis. 2016;22: 336–339. doi:10.3201/eid2202.150495
2. Badurdeen S, Valladares DB, Farrar J, Gozzer E, Kroeger A, Kuswara N, et al. Sharing experiences: towards an evidence based model of dengue surveillance and outbreak response in Latin America and Asia. BMC Public Health. 2013;13: 607. doi:10.1186/1471-2458-13-607
3. Codeco C, Coelho F, Cruz O, Oliveira S, Castro T, Bastos L. Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. Rev DÉpidémiologie Santé Publique. 2018;66: S386. doi:10.1016/j.respe.2018.05.408
4. Lowe R, Barcellos C, Coelho CAS, Bailey TC, Coelho GE, Graham R, et al. Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. Lancet Infect Dis. 2014;14: 619–626. doi:10.1016/S1473-3099(14)70781-9
5. Lee S.A, Economou T., Catão R., Barcellos C., Lowe R. Data and R code to accompany "The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil" (version 1.0.0). 2021. Available: https://github.com/sophie-a-lee/Dengue_expansion
6. Pereira, R.H.M., Gonçalves, C.N. geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. 2019. Available: https://github.com/ipeaGIT/geobr
7. Lowe R, Lee SA, O'Reilly KM, Brady OJ, Bastos L, Carrasco-Escobar G, et al. Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. Lancet Planet Health. 2021;5: e209–e219. doi:10.1016/S2542-5196(20)30292-8
8. Lowe R, Gasparrini A, Meerbeeck CJV, Lippi CA, Mahon R, Trotman AR, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. PLOS Med. 2018;15: e1002613. doi:10.1371/journal.pmed.1002613
9. Harris I, Osborn TJ, Jones P, Lister D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data. 2020;7: 109. doi:10.1038/s41597-020-0453-3
10. van der Schrier G, Barichivich J, Briffa KR, Jones PD. A scPDSI-based global data set of dry and wet spells for 1901–2009. J Geophys Res Atmospheres. 2013;118: 4025–4048. doi:10.1002/jgrd.50355

11. Palmer WC. Meteorological drought. US Department of Commerce, Weather Bureau; 1965.

12. Alley WM. The Palmer drought severity index: limitations and assumptions. J Appl Meteorol Climatol. 1984;23: 1100–1109.

13. Wells N, Goddard S, Hayes MJ. A self-calibrating Palmer drought severity index. J Clim. 2004;17: 2335–2351.

14. Baston, Daniel. exactextractr: Fast Extraction from Raster Datasets using Polygons. 2020. Available: https://CRAN.R-project.org/package=exactextractr

15. Gini index (World Bank estimate) - Latin America & Caribbean, Argentina, Brazil, Chile, Uruguay, Paraguay, Peru, Bolivia, Colombia, Guyana | Data. [cited 17 Jun 2021]. Available: https://data.worldbank.org/indicator/SI.POV.GINI?end=2019&locations=ZJ-AR-BR-CL-UY-PY-PE-BO-CO-GY&start=1979&view=chart

16. Estatística IB de G e. Regiões de influência das cidades 2007. IBGE Rio de Janeiro; 2008.

17. Estatística IB de G e. Regiões de influência das cidades 2018. IBGE Rio de Janeiro; 2020.

18. Wood SN. Generalized additive models: an introduction with R. CRC press; 2017.

19. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. Stat Sin. 2004; 731–761.

20. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Ser B Stat Methodol. 2011;73: 3–36. doi:10.1111/j.1467-9868.2010.00749.x

21. Barcellos C, Lowe R. Expansion of the dengue transmission area in Brazil: the role of climate and cities. Trop Med Int Health. 2014;19: 159–168. doi:10.1111/tmi.12227

22. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78: 1–3.

B.2 Supplementary Figures

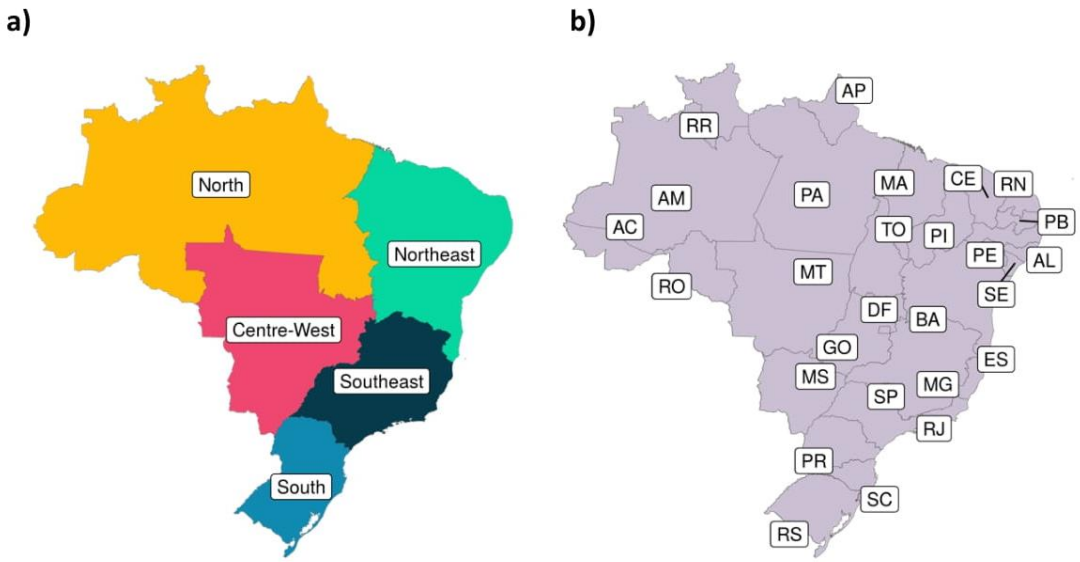


Fig S1: The organisation of Brazil into a) 5 geo-political regions, and b) 27 federal units. Abbreviations: AC = Acre, AL = Alagoas, AP = Amapá, AM = Amazonas, BA = Bahia, CE = Ceará, DF = Distrito Federal, ES = Espírito Santo, GO = Goiás, MA = Maranhão, MT = Mato Grosso, MS = Mato Grosso do Sul, MG = Minas Gerais, PA = Pará, PB = Paraíba, PR = Paraná, PR = Pernambuco, PI = Piauí, RJ = Rio de Janeiro, RN = Rio Grande do Norte, RS = Rio Grande do Sul, RO = Rondônia, RR = Roraima, SC = Santa Catarina, SP = São Paulo, SE = Sergipe, TO = Tocantins.

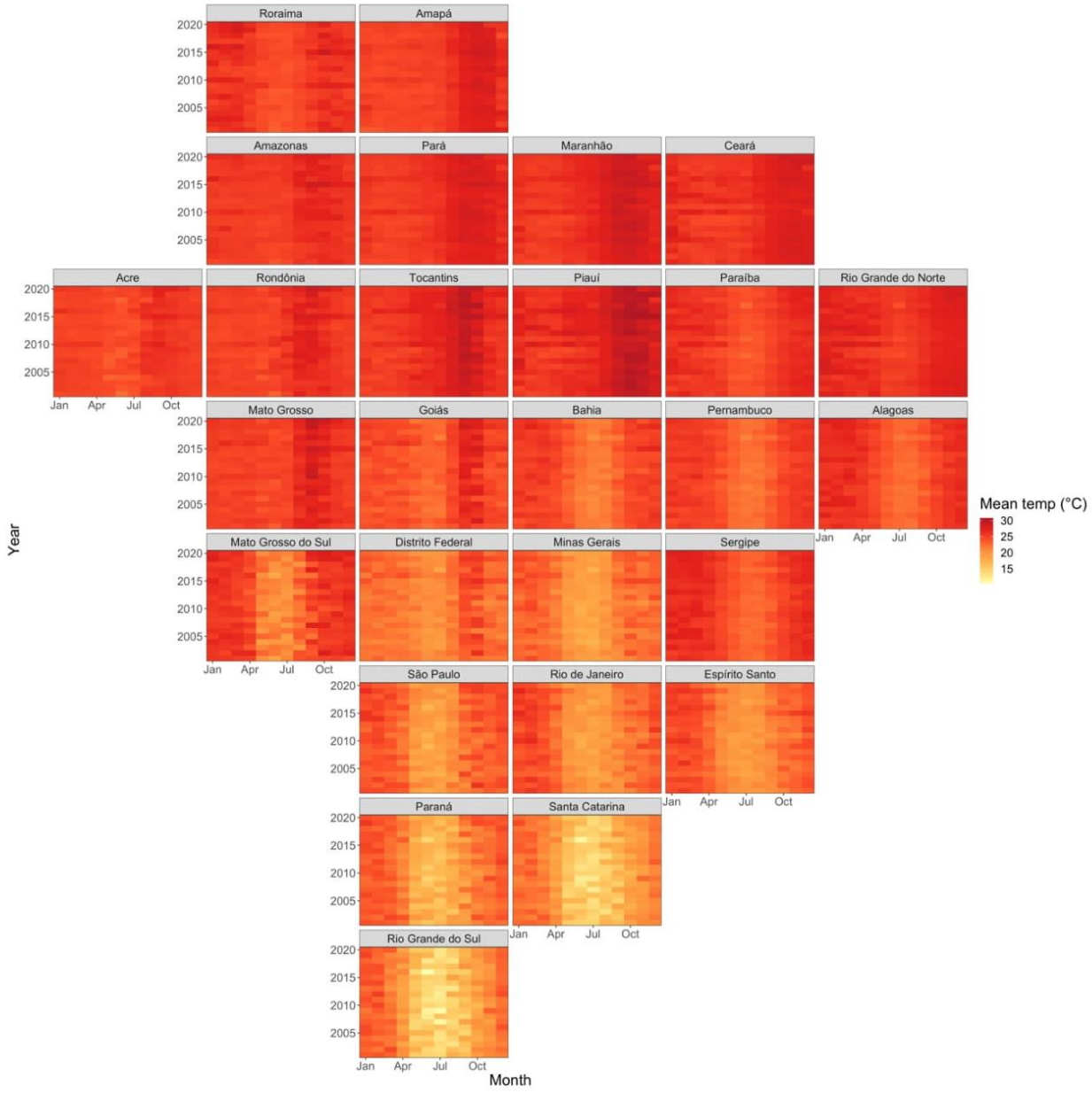


Fig S2: Average monthly mean temperature (°C) in each Brazilian state January 2001 - December 2020.

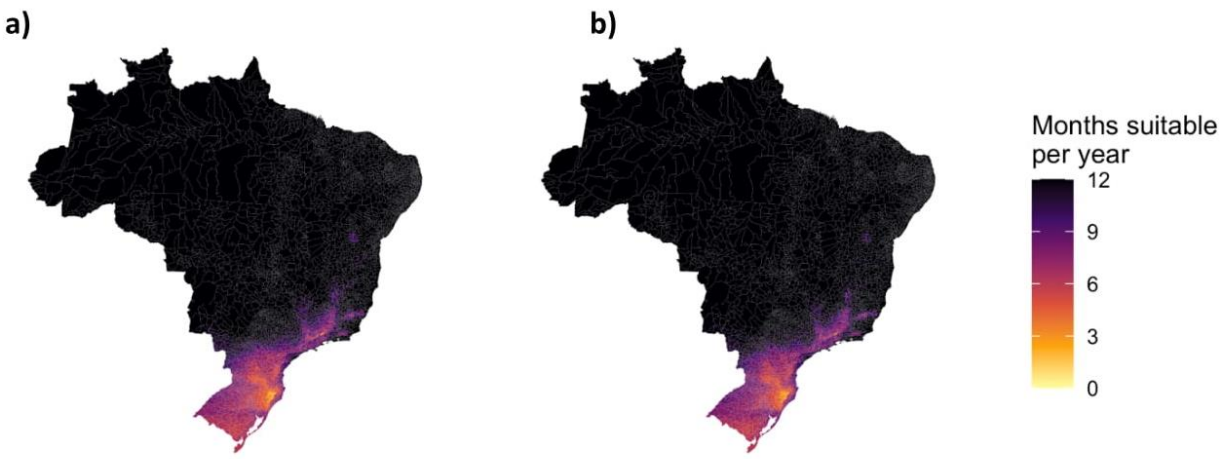


Fig S3: The average number of months suitable for dengue transmission per year a) 2001 - 2010, and b) 2011 - 2020. The average number of months with mean temperature between 16.2 and 34.5°C aggregated to the two decades of data. Most of Brazil experiences suitable temperatures year-round apart from areas of South Brazil and areas of high altitude in the Southeast which experience cool winters.

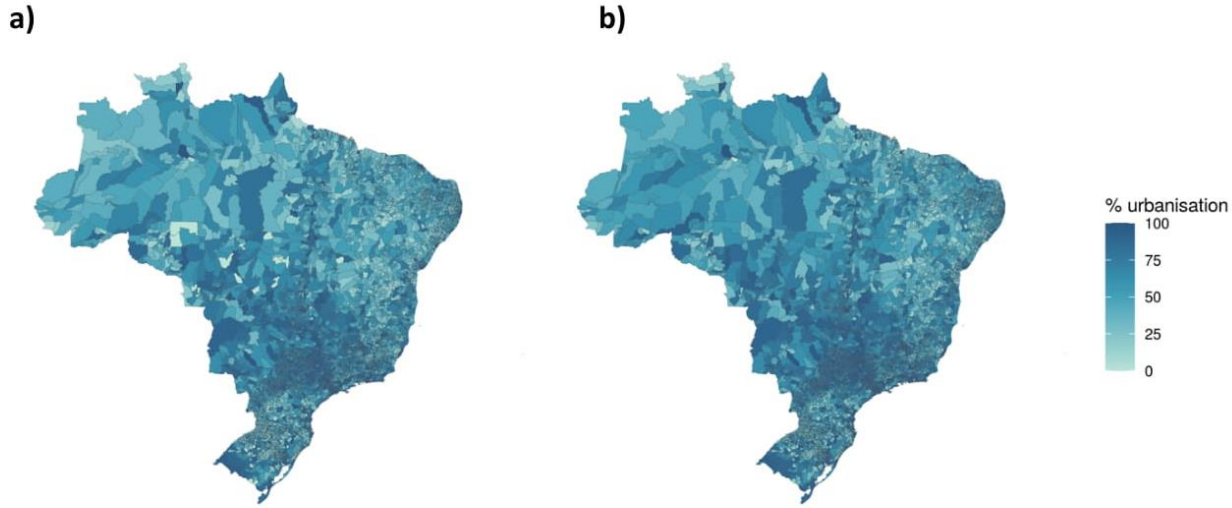


Fig S4: The percentage of residents living in urban areas of each municipality from the 2000 (a) and 2010 (b) censuses. Levels of urbanisation differ greatly across Brazil, with the majority of Southeast and South Brazil living in urban areas in comparison to the North and Northeast which has a larger rural population.

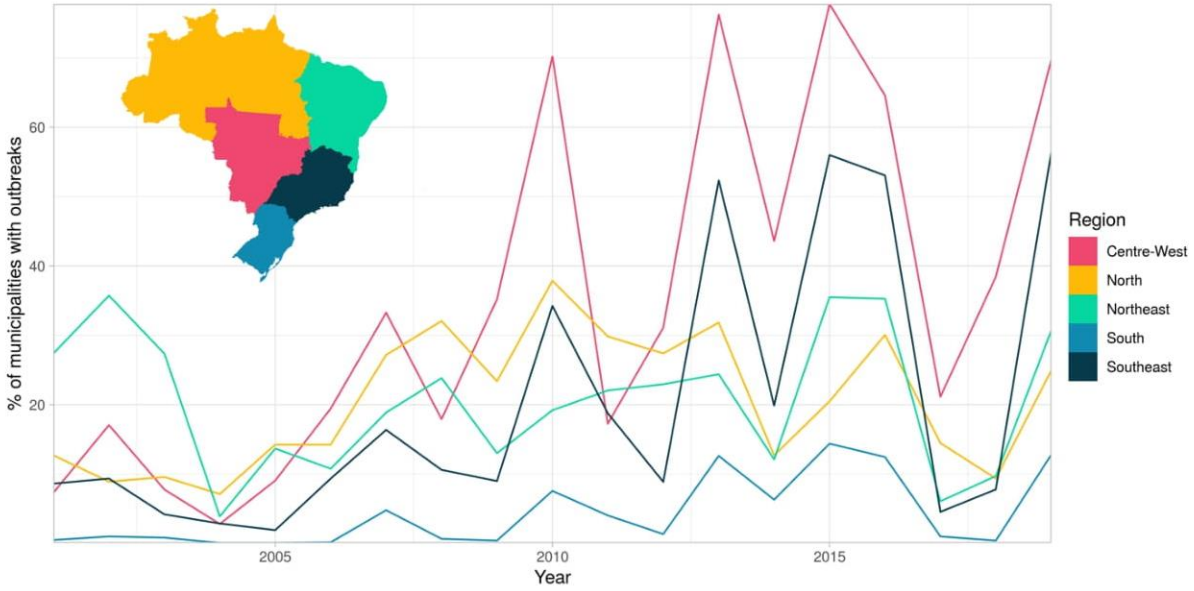
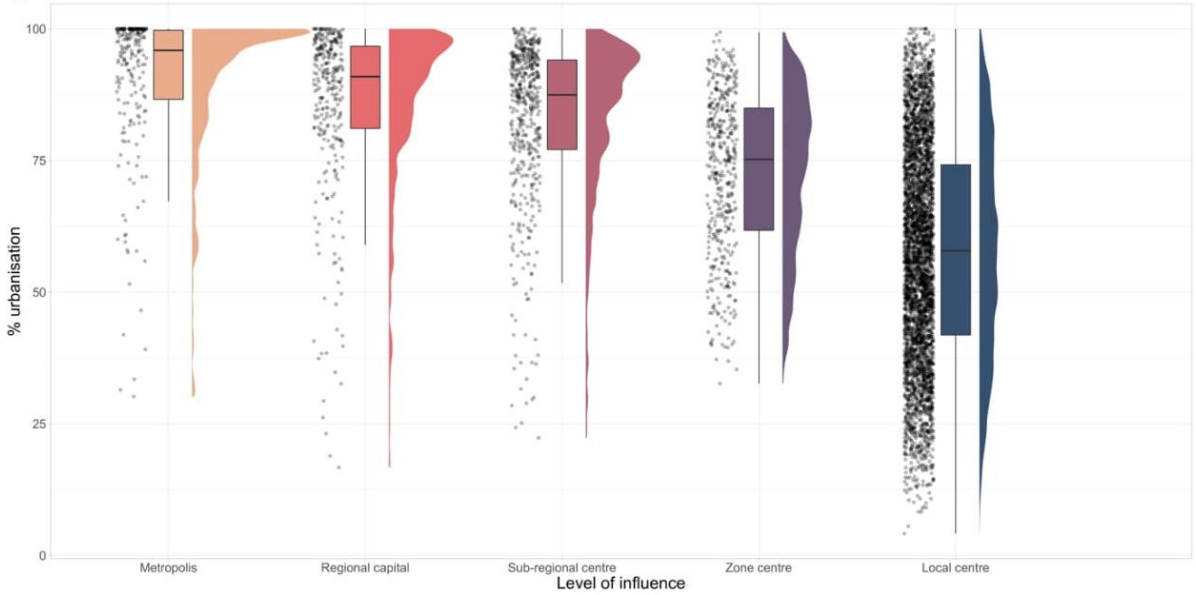
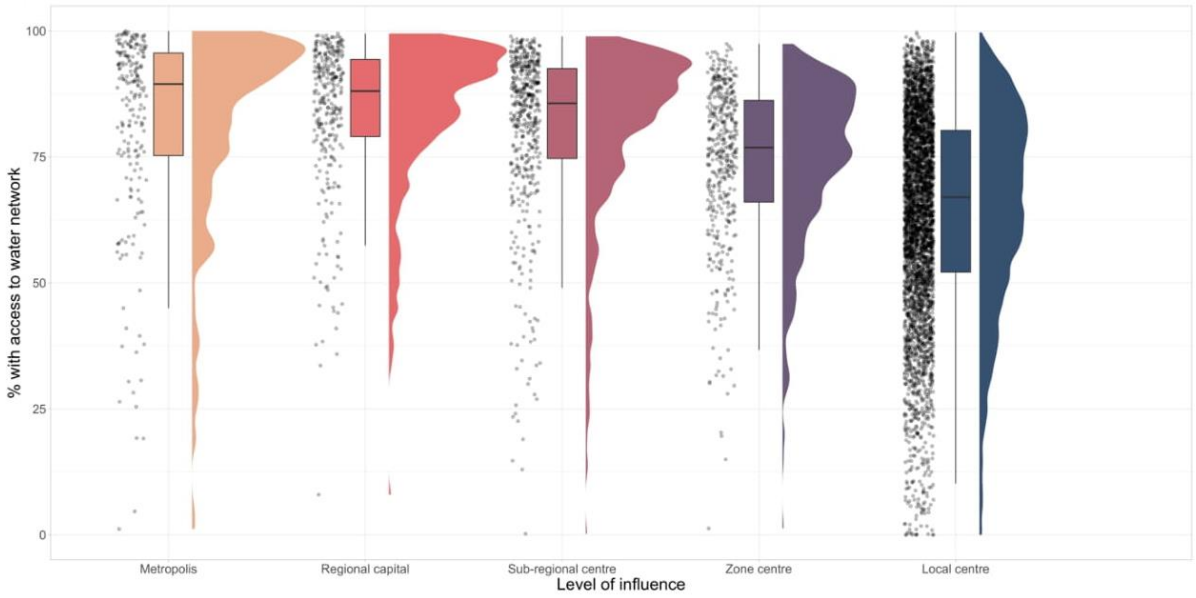


Fig S5: The proportion of municipalities in each region of Brazil experiencing an outbreak per year 2001 - 2020. The proportion of municipalities affected by outbreak has increased since 2010 in every region of the country, although outbreaks in South Brazil are still focused on a small part of the region.

a)


100
75
50
25
0
% urbanisation
Metropolis
Regional capital
Sub-regional centre
Zone centre
Local centre
Level of influence

b)


100
75
50
25
0
% with access to water network
Metropolis
Regional capital
Sub-regional centre
Zone centre
Local centre
Level of influence

c)

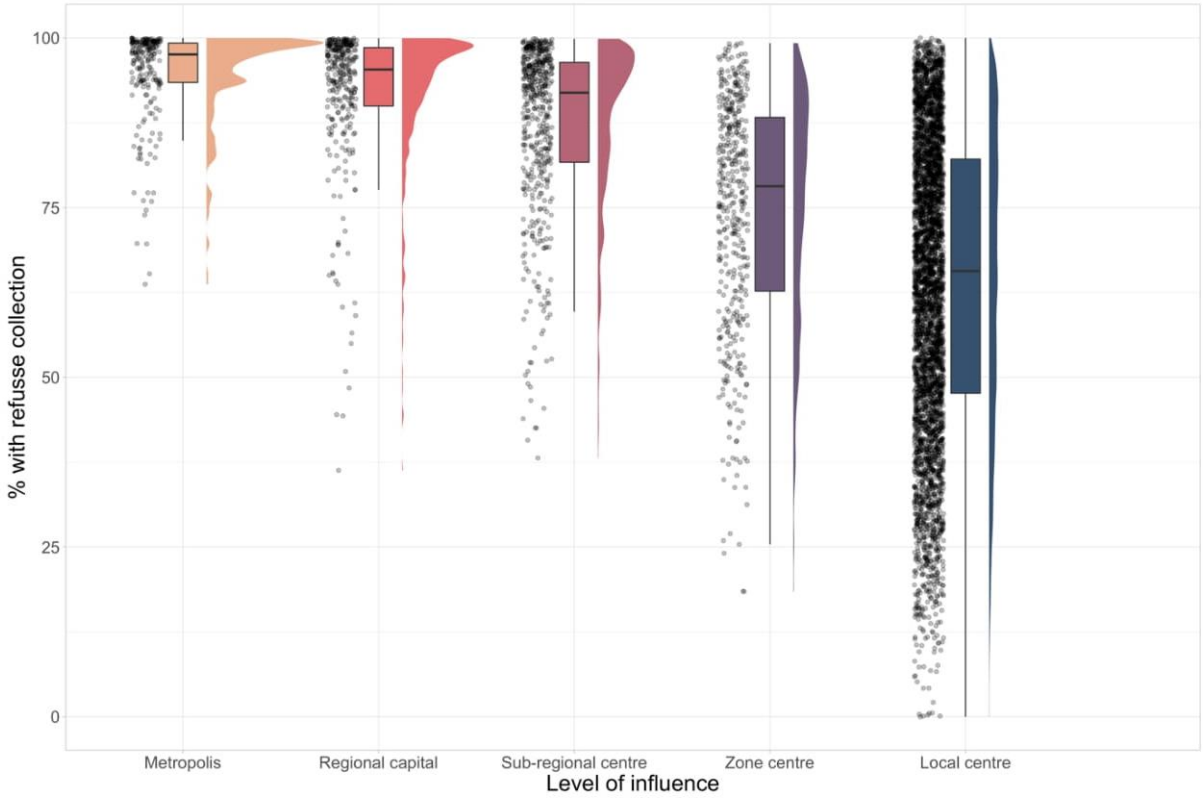


Fig S6: Raincloud plots exploring the relationship between REGIC level of influence and a) urbanisation, b) access to piped water, and c) refuse collection. Metropoles and regional capitals have higher levels of urbanisation and access to basic services than municipalities that had lower levels of connectivity within the urban network. Local centres were more varied in terms of basic services and urban levels than the other levels and covered a wide range of city types.

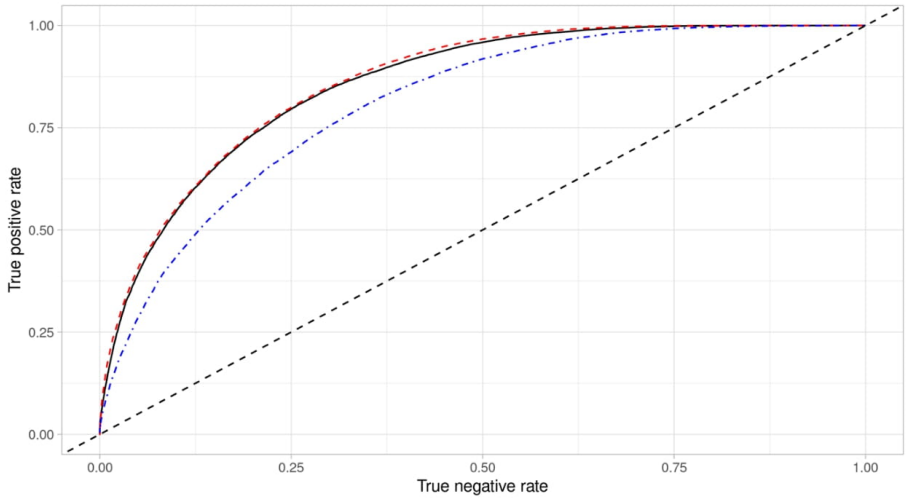


Fig S7: Receiver operating characteristic (ROC) curve for the final model (solid black line), the model using an outbreak threshold of over 100 cases per 100,000 residents (red dashed line), and the model using an outbreak threshold of over the 75th percentile (blue dashed line), compared to chance (black dashed line). The closer to the top-left corner, the better the predictive ability of a model. As the ROC curve lies above the dashed reference line, this model performs better than chance.

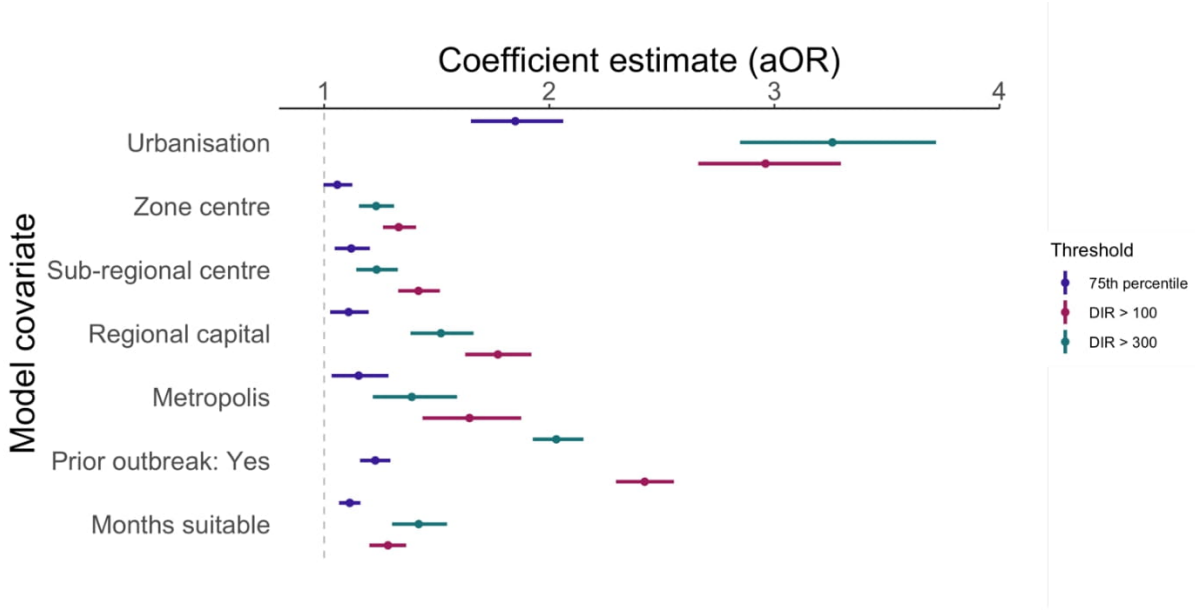


Fig S8: The mean and 95% credible interval of the posterior distribution for each model covariate under different outbreak threshold definitions. Coefficient estimates using the outbreak indicator based on the 75th percentile were noticeably smaller than the fixed threshold alternatives. The fixed threshold models (where outbreaks were defined as a dengue incidence rate of over 100 or 300) produced similar estimates, however the odds of an outbreak in municipalities after a previous outbreak was higher for the DIR = 100 model.

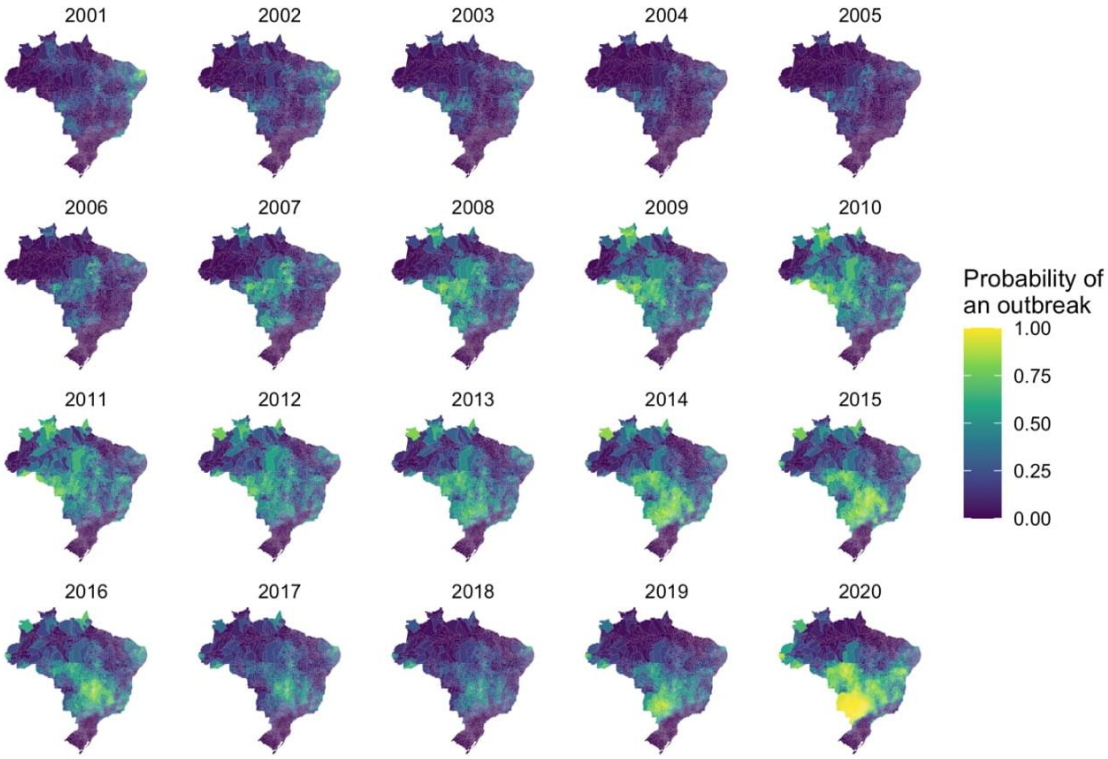
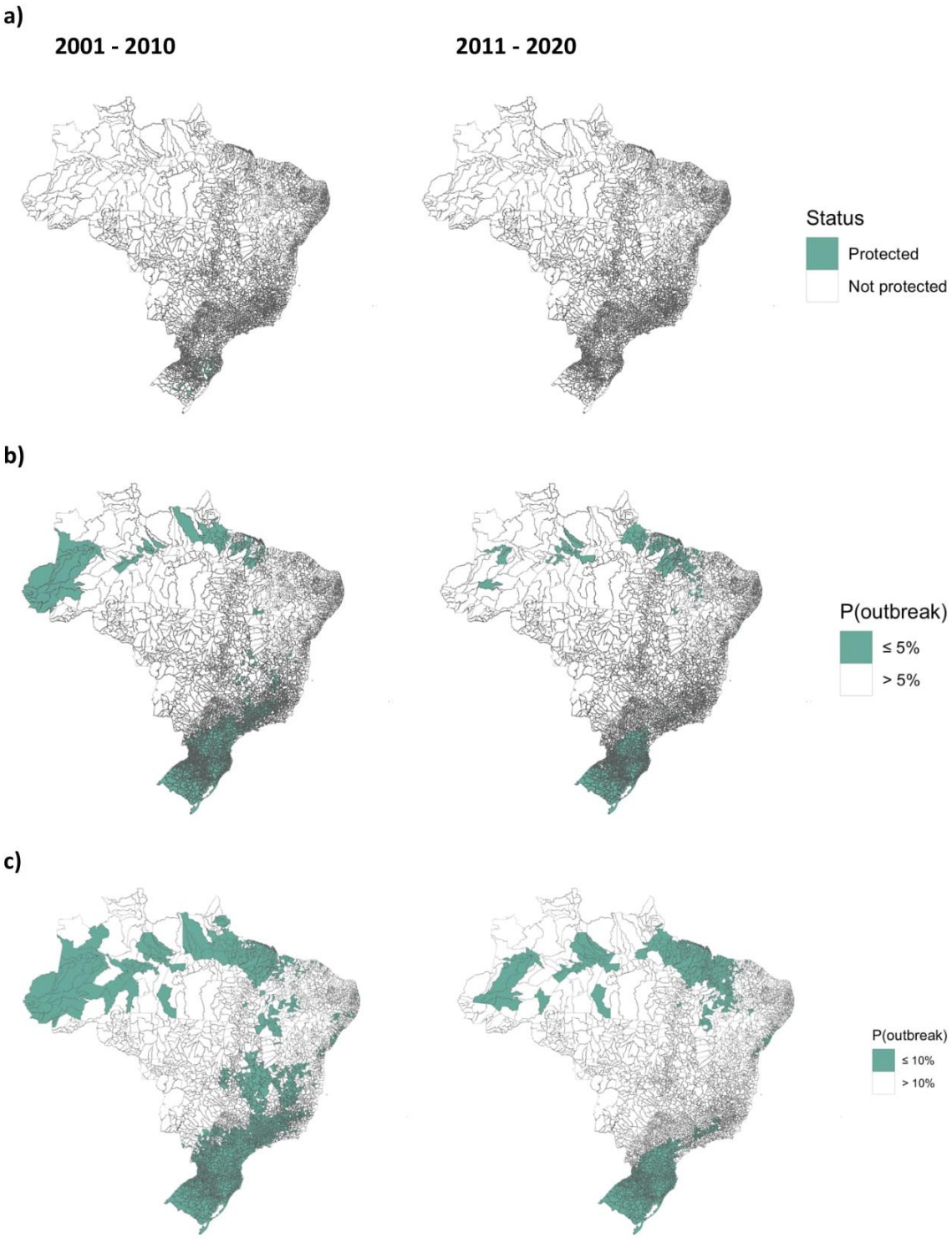


Fig S9: The probability of an outbreak estimated from the model for each year 2001 - 2020. The mean probability of an outbreak estimated by taking 1000 simulations from the posterior distribution of the response and transforming the outcome using a probit function


a)
2001 - 2010
2011 - 2020
Status
Protected
Not protected
b)
P(outbreak)
≤ 5%
> 5%
c)
P(outbreak)
≤ 10%
> 10%

d)

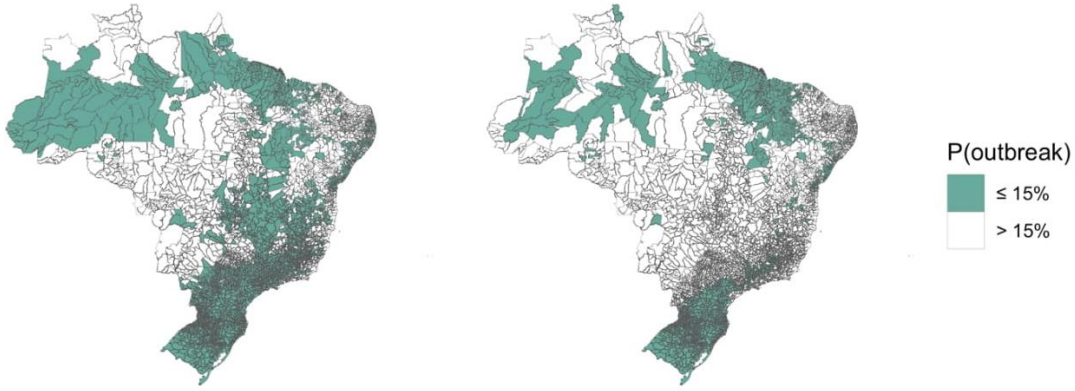


Fig S10: Comparison of different risk thresholds to define current geographical barriers to dengue outbreaks. Municipalities were considered 'protected' if the probability of an outbreak was less than or equal to the threshold a) 0%, b) 5%, c) 10% or d) 15%. The threshold of 10% was chosen as it was the most comparable with previous studies.

B.3 Table S1

Table S1: Distribution of municipalities at each level of influence in the urban network, 2007 (1) and 2018 (2). The number of municipalities classified as metropoles (largest cities in Brazil, connected throughout the entire country), regional capitals (large cities connected regionally and to metropoles), sub-regional capitals (cities connected locally and to the three largest metropoles), zone centres (smaller cities generally connected only to their neighbours), and local centres (smallest cities typically disconnected from the urban network).

Region	Metropolis		Regional capital		Sub-regional capital		Zone Centre		Local Centre	
	2007	2018	2007	2018	2007	2018	2007	2018	2007	2018
Brazil	**203**	**214**	**259**	**308**	**270**	**496**	**658**	**437**	**4170**	**4105**
North	5	5	10	13	19	28	38	27	377	376
Northeast	33	33	59	62	57	97	179	140	1464	1460
Central-West	25	24	4	7	9	46	83	51	342	335
Southeast	92	95	129	149	104	200	209	120	1134	1104
South	48	57	57	77	81	125	149	99	853	830

Table S2: Posterior mean and 95% credible interval (CI) estimates for linear effect parameters, calculated using an outbreak threshold of 100 cases per 100,000 residents, calculated using an outbreak threshold based on the 75th percentile of dengue incidence rate, and using temperature suitability related to *Aedes aegypti* only shown on the adjusted odds ratio (aOR) scale.

Coefficient	aOR (95% CI)		
	Medium risk model[1]	75th percentile model[2]	Aedes aegypti model[3]
Urbanisation	2.96 (2.66, 3.30)	1.86 (1.66, 2.08)	3.21 (2.80, 3.65)
REGIC level: metropolis	1.65 (1.44, 1.88)	1.15 (1.03, 1.28)	1.38 (1.20, 1.58)
REGIC level: regional capital	1.77 (1.63 1.92)	1.11 (1.03, 1.21)	1.51 (1.38, 1.66)
REGIC level: sub-regional centre	1.42 (1.33, 1.51)	1.12 (1.05, 1.19)	1.23 (1.14, 1.34)
REGIC level: zone centre	1.33 (1.26, 1.41)	1.06 (0.99, 1.12)	1.24 (1.16, 1.31)
Prior outbreak: yes	2.42 (2.30, 2.55)	1.23 (1.16, 1.29)	2.01 (1.91, 2.12)
Months with suitable temperature	1.29 (1.20, 1.37)	1.11 (1.07, 1.16)	1.34 (1.28, 1.40)

[1] Response variable is dengue outbreak defined as over 100 cases per 100,000 inhabitants

[2] Response variable is dengue outbreak defined as above the 75th percentile of the annual dengue incidence rate between 2001 - 2020, with a minimum incidence equivalent to 5 cases per year.

[3] Temperature suitability set to *Aedes aegypti* limits, between 17.8° and 34.5°C

B.5 S3 Table

Table S3: Model comparison statistics. Area under the receiver operator curve (AUROC) and Brier scores for models assuming an outbreak threshold of over 300 cases per 100,000 residents (high risk model), over 100 cases per 100,000 (medium risk model), over the 75th percentile of incidence rates, and a model including the number of months considered extremely wet.

Model formula	AUROC (95% confidence interval)	Brier score
High risk model	0.858 (0.856, 0.861)	0.109
Medium risk model	0.864 (0.861, 0.866)	0.138
75th percentile model	0.809 (0.807, 0.812)	0.125
Extremely wet model	0.859 (0.856, 0.861)	0.109

Appendix C: Supplementary Material Chapter 3

C.1 Technical appendix

Technical appendix to support Chapter 3: *Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions*. Contains detailed descriptions about the models identified within the review. Taken from https://doi.org/10.1098/rsif.2021.0096.

Technical appendix: Detailed descriptions of spatial models identified in the systematic review

We identified 17 distinct spatial models in this review: 9 statistical, 4 machine learning and 4 mechanistic models. This document provides detailed information about how these models are structured and the ways in which spatial connectivity was accounted for in these models.

Table of Contents

1. Statistical models

1.1 Generalised linear model (GLM)

A GLM aims to represent the relationship between a dependent variable (in this case, some outcome related to mosquito-borne disease transmission) and one or more explanatory variable(s) using a linear model. GLMs assume that the relationship between the outcome and explanatory variables is fixed and constant over time and space. One limitation of these models is that all observations are assumed to be independent. To overcome this, researchers included spatial variables, which aim to account for connectivity within the data, as covariates in the GLM. Spatial GLMs took the form:

$$\eta(\mu(Y_i)) = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j + \sum_{k=1}^{p} \zeta_k S_{ik},$$

where $\mu(Y_i)$ is the mean of Y_i, the dependent variable, and X_{ij} are *m* explanatory variable(s) in location *i*. η is a link function determined by the model choice, β_j are regression coefficients, S_{ik} are the *p* spatial covariates and ζ_k are the corresponding regression coefficient(s).

There were 17 studies included in this review that used a GLM. Fifteen of these studies included spatial covariates informed by other, connected locations. Spatial covariates included the number of cases observed in connected regions (1–7) or the number of connected regions that had a 'high' incidence (8), the coordinates of the centroid of a region (9,10), the number of people moving between regions (11–13) or the amount of time they spent there (14), the distance between regions (11,12), time taken to travel between regions (12), and eigenvectors created using spatial filtering (15). Spatial filtering aims to account for shared, unobservable characteristic between neighbours by decomposing Moran's I (a measure of spatial autocorrelation) into eigenvectors which were included in the model as covariates (16). One GLM included connectivity within the dependent variable, modelling the time taken between imported cases and indigenous cases in connected regions (17). The final GLM did not include spatial covariates but used generalised estimating equations to fit the model which does not assume independence between observations and allows for a more flexible covariance structure when this structure is not known (18).

Spatial GLMs are easier to fit and interpret if it is possible to determine which regions or observations are connected. Otherwise, spatial GLMs are not appropriate as any remaining spatial autocorrelation not accounted for can lead to over-precise estimates. Including information from connected regions allows the model to 'borrow strength' from other parts of the data to obtain more precise estimates; this is particularly useful when a study area is broken into small regions or where observed cases are low. However, the inclusion of a large number of spatial covariates risks overfitting the data or introducing multicollinearity. The inclusion of spatial covariates as fixed effects assumes that the relationship between them and the outcome is stationary (the same across the whole spatial area) which may not be appropriate, particularly in large-scale studies. Spatial GLMs may be a useful exploratory tool but are not recommended for making inferences or predictions within a spatial study.

1.2 Geographically weighted regression (GWR)

GWR fits a local GLM to each region using data from connected locations. Each local model has different regression coefficients, overcoming the stationarity assumption of spatial GLMs:

$$\eta(\mu(Y_i)) = \beta_{io} + \sum_{j=1}^{m} X_{ij}\beta_{ij},$$

where $\mu(Y_i)$ is the mean of Y_i, the dependent variable, and η is a link function determined by the model choice. X_{ij} are m explanatory variable(s) in location i, and $\beta_{i,j}$ are the corresponding local regression coefficients for region i.

Local GLMs are estimated using information from connected regions, weighted using a function based on distance that assumes regions are more connected the closer they are. The functions used to estimate these weights are known as kernel density functions; they are based on the distance between observations and some bandwidth, estimated as part of model fitting. Kernels can be fixed (where a set number of connected regions are used to fit each model) or adaptive (where the distance within which areas are considered connected differs) (19).

We identified 20 studies that used GWR to model mosquito-borne disease transmission (20–39). GWR is a useful exploratory tool to investigate how the relationship between variables and the outcome differs across space. The model fitting process is fairly efficient, and interpretation of the models is no different from standard GLMs. As GWR does not produce a global model, it is inappropriate to use this method to make inferences about an entire region. GWR can be used as an exploratory tool alongside an alternative spatial modelling method when making inferences or predictions.

1.3 Generalised additive model (GAM)

GAMs extend the GLM framework to allow nonlinear relationships between covariates and the dependent variable using smoothing functions. These smoothing functions can be applied to spatial variables to account for spatial dependency, such as distance between observations, or can be applied to coordinates to create a spatially structured random field, which can be included in the model. Spatial GAMs take the form:

$$\eta(\mu(Y_i)) = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j + \sum_{k=1}^{q} f_k(Z_{ik}) + f_{spat},$$

where $\mu(Y_i)$ is the mean of Y_i, the dependent variable, and η is a link function determined by the model choice. X_{ij} are m explanatory variable(s) in location i, β_j are the corresponding regression coefficients. f_k are smooth functions applied to each of the q covariates, Z_{ik}, in location i, and f_{spat} is a spatially structured surface defined as either a bi-dimensional smooth function in space or a random field based on the location of observation i.

We identified 18 studies that used a spatial GAM; 14 used a bi-dimensional smoothing function which is applied to coordinates to define f_{spat}. This function assumes that connectivity between observations decays as the distance between them increases. The smoothing function was either based on coordinates of cases (40–44), households (45–48), or the centroid of a region (9,49–52). One study included f_{spat} defined by a Markov random

field (53); this assumes that regions are connected if and only if they share a border (54). The remaining 3 studies included spatial covariates in their GAM to account for connectivity rather than a spatially smoothed field. Two transformed these spatial covariates using a smoothing function (the incidence of surrounding districts (55), and the distance from an index location (56)), the other included case coordinates but did not specify whether a smoothing function was applied (57).

Generalised additive models are useful to account for nonlinear, complex relationships between the outcome and covariates. As with GLM, the inclusion of many spatial covariates risks introducing multicollinearity and overfitting the data. Even if the spatial covariate has a smoothing function applied to it, the model still assumes this relationship is stationary which may not be appropriate for large-scale studies. The inclusion of a spatially structured random field overcomes some of the issues associated with multicollinearity and overfitting as it adjusts the covariance structure of the model rather than the estimated outcome. Spatially structured random fields can be plotted to generate hypotheses related to connectivity and identify factors that may improve the model.

1.4 Autoregressive distributed lag model (ADLM):

A distributed lag model is a GLM that assumes covariates have an influence on the outcome over time rather than at a single time point. The model contains covariates from previous time points to understand how this relationship changes through time, each with their own regression coefficient. Researchers must choose a sensible lag period, in this equation *j,* which is the length of time covariates are expected to be related to the outcome:

$$\eta\big(\mu(Y_{i,t})\big) = \beta_0 + \sum_{j=1}^{m} X_{i,t-j}\beta_{t-j},$$

where $\mu(Y_{i,t})$ is the mean of the dependent variable in region *i* at time t, and η is a link function determined by the model choice. $X_{i,t-j}$ are the explanatory variable(s) at time *t - j*, and β_{t-j} are their corresponding regression coefficients.

The ADLM extends this framework to take account of spatial autocorrelation by fitting a separate regression model to error terms (the difference between the observed and fitted outcomes) of the model, using lagged outcome values from connected areas as covariates. The model is fitted using an iterative procedure:

Let $w_{i,t}$ be the difference between the fitted values and observed values. These error terms are regressed against previous outcomes in surrounding districts:

$$w_{i,t} = \sum_{p=1}^{q} \phi_{i,p} \sum_{l \in L_i} \lambda_{i,l} \eta\big(Y_{l,t-p}\big),$$

where $\lambda_{i,l}$ are coefficients measuring the influence of the i^{th} region belonging to the neighbourhood L_i (containing all regions connected to region *i*), and $\phi_{i,p}$ measures the relationship between connected regions at a lag of *p* time steps. We identified one ADLM model (58).

Although the ADLM allows researchers to investigate the impact covariates have on transmission over time, it has several limitations. The inclusion of multiple spatial covariates from neighbouring regions is likely to introduce multicollinearity into the model, making results unstable. The high number of covariates risks overfitting the data, limiting broader inferences. The iterative procedure makes model fitting more computationally intensive than other statistical models and difficult to interpret, as well as making uncertainty quantification not straightforward.

1.5 Generalised linear mixed model (GLMM) and Generalised additive mixed model (GAMM)

GLMM and GAMM are extensions GLM and GAM respectively that include one or more random effect(s). Random effects are used to overcome the independence assumption that underpins GLM/GAM by accounting for residual autocorrelation and to allow for pooling of the data in regions with little information. They can either be unstructured, where the source of autocorrelation is unknown, or structured, for example in spatial or temporal data. Structured random effects require a covariance structure to be specified that defines this connectivity. GLMM/GAMMs are also known as hierarchical or multilevel models as they take account of data that are clustered or grouped at different levels (for example within households or administrative regions). GAMMs take the form:

$$\eta(\mu(Y_i)) = \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j + \sum_{k=1}^{q} f_k(Z_{ik}) + \boldsymbol{b_i},$$

where $\mu(Y_i)$ is the mean of Y_i, the dependent variable, and η is a link function determined by the model choice. X_{ij} are m explanatory variable(s) in location i, β_j are the corresponding regression coefficients. f_k are smooth functions applied to each of the q covariates, Z_{ik}, in location i, and $\boldsymbol{b_i}$ are one or more random effects. Note that GLMMs take the same form but do not contain nonlinear $f_k(Z_{ik})$ terms.

We identified 139 studies that used a GLMM, 14 that used a GAMM, and one that used both. Most (147/152) of these studies used spatially structured random effects to account for connectivity between regions, the remaining 5 used spatial covariates (2 used spatial filtering to create eigenvectors which were included as covariates (59,60), one used a covariate created to measure the 'risk of importation' based on incidence and number of travellers arriving from other countries (61), and one used geo-located Twitter data to create a covariate that reflected commuting behaviour (62)), or a spatially smoothed random field (based on the coordinates of villages (63)), combined with an unstructured random effect to take account of other, unobserved variation between regions. Spatial structures used to calculate random effects were either distance- or neighbourhood-based. There were 91 studies that used a neighbourhood-based approach (9,15,37,64–150): random effects were structured using a Gaussian Markov random field which assumes regions are connected if and only if they are neighbours (151) (either through a shared border or if they lay within a set distance (64)), this is also known as a conditional autoregressive (CAR) model. 57 studies used a distance-based structure to formulate their random effects, most often based on the Matérn covariance function (152), which assumes that connectivity decays exponentially as the distance between regions or observations increases (134,153–205).

The inclusion of spatial connectivity within a random, rather than fixed, effect overcomes issues of stationarity, multicollinearity and overfitting that were present when using spatial covariates. These methods are more computationally intensive but the development of methods designed to incorporate these spatial structures, such as integrated Laplace approximations (INLA) (206), has made these more accessible in recent years. At present, only two approaches to structuring random effects were identified which were both compatible with distance-based assumptions of connectivity. No random effect structures were identified that allowed for connectivity between distant regions or observations, for example through international travel.

1.6 Distributed lag nonlinear models (DLNM)

DLNMs are an extension of the GAM framework that allow the model to explore nonlinear, delayed associations between exposures, the response, and temporal lags. DLNMs do this by including cross-basis functions, bi-dimensional functions that specify the dependency along the exposure and along time lags. The cross-basis function combines the exposure-response smoothing function, f(x), used to model nonlinear associations in GAMs, with a lag-response function, w(ℓ), to create a bi-dimensional exposure-lag-response function f.w(x, ℓ) (207). The general form of a DLNM is:

$$\eta(\mu(Y_i)) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{m=p+1}^{r} f.w(X_{im}, \ell),$$

where $\mu(Y_i)$ is the mean of Y_i, the dependent variable, and η is a link function determined by the model choice. X_{ij} are m explanatory variable(s) in location i, β_j are the corresponding regression coefficients, and $f.w(X_{im}, \ell)$ are the exposure-lag-response functions with a maximum lag of ℓ.

These models can be extended to include structured or unstructured random effects (as seen in mixed models), to include a spatially smoothed random field (as seen in GAMs), or to include spatial covariates. We identified 4 papers that used a DLNM: 3 included spatially structured random effects using a Gaussian Markov random field, assuming connectivity exists between adjacent regions (208–210), the other included the number of cases observed in surrounding regions, firstly as a smoothed spatial covariate, then within a cross-basis function to account for lagged effects (55).

DLNM models allow for the relationship between covariates and the outcome to be explored over time, rather than at a single point. This is particularly useful when studying the impact of climate or vector control on mosquito-borne disease which is often complex, nonlinear and delayed by a period of time (211). Unlike the ADLM, lagged variables are included within the exposure-lag-response function rather than as separate fixed effects which overcomes some issues related to overfitting and multicollinearity. The model structure is flexible and can be extended to include different spatial methods (such as random effects or spatial covariates). Due to the complex nature of the exposure-lag-response function, these models are more difficult to interpret and visualise than other statistical models.

1.7 Bespoke statistical models

Two statistical modelling studies carried out novel statistical methods that did not fall into the above categories.

One study aimed to estimate the probability of a case, *i*, being the 'source' of another case, *j*, based on the distance and time between them. The hazard of infection was estimated as the product of a temporal kernel (an exponential function that reduces as time lag increases) and a spatial kernel (an exponential function that decays as distance increases). The spatial kernel used the distance between the original case and the potential connected case's work and home addresses. These addresses were estimated using data from the public transportation network in Singapore. The probability of case *j* arising from case *i* is estimated using the hazard as defined above, divided by the total hazard of all other cases that occurred prior to case *j* (212). This study provided a novel approach to investigating the role of work places and commuting in the risk of Zika transmission within an urban environment by treating humans as vectors as well as hosts (as mosquitoes were largely naïve, humans infecting mosquitoes were expected to drive expansion rather than mosquitoes infecting humans). Although this may be appropriate for investigating a newly-emerging or re-emerging disease in a largely naïve mosquito population, the impact of people infecting mosquitoes becomes less important in endemic settings. The method may be less suitable for diseases transmitted by night-biting mosquitoes (such as malaria) or larger-scale studies where other forms of travel drive importation.

Another study estimated the instantaneous risk of chikungunya transmission between areas as a product of connectivity measures (including distance, number of air passengers and mobility estimated using the gravity model) and a coefficient. This was then used to estimate the likelihood of importation of dengue, although the underlying probability assumptions and the method used to estimate coefficients were not given in the paper (213).

2. Machine learning methods

2.1 Neural network models

Neural networks combine multiple models to approximate the relationship between an outcome (or output) and a number of covariates (or inputs). A neural network involves weighting the inputs together to create a temporary output called a node. This process is repeated to create multiple nodes, which are referred to as a layer. This process is repeated by weighting the nodes together to create a new layer of nodes. The process is repeated multiple times and the layers are referred to as hidden layers. The nodes in the final hidden layer are then weighted together to create the output. The use of multiple layers and the flexibility of functions that can be used for the weighting allows neural networks to be used for relationships that are nonlinear and complex (Figure 1). Without the hidden layers, a neural network is equivalent to a nonlinear regression model (214). As the relationships are hidden and complex, it is difficult to use a neural network to understand the relationship between covariates and an outcome, but they are capable of providing accurate predictions if carried out correctly.

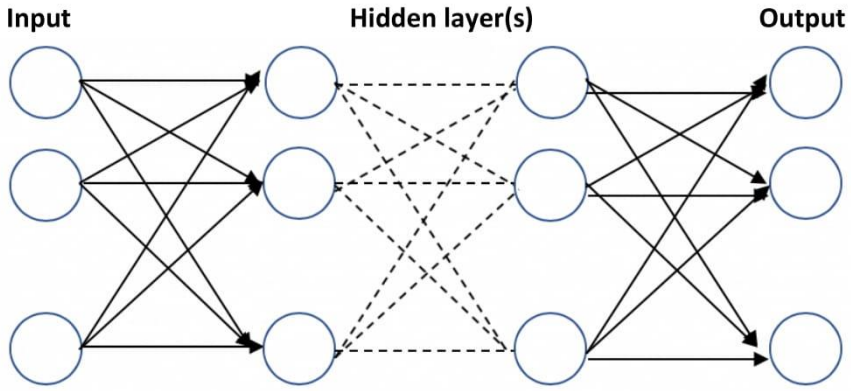


Figure 1. A visual representation of the neural network framework

There were 4 studies that included a neural network. Two studies included spatial variables within their input: one included the number of cases observed in adjacent districts (35), the other included connectivity-risk variables calculated the volume of passengers moving between countries and the incidence rate at their origin and destination (214). The other 2 studies created movement matrices based public transportation data; one used this matrix to weight links between hidden layers (216), the other used this matrix as an input for their network (217).

Neural networks typically require a large amount of data, however they are able to include complex dependencies between variables without being limited by a priori model assumptions as is the case with statistical models. Neural networks are useful for studies that require accurate predictions where a large amount of data is available but are less appropriate to understand the relationships between covariates and the outcome or the underlying process of transmission.

2.2 Regression trees

There were two papers that used regression tree methods: one used random forest regression (13) and the other used boosted regression trees (218). Regression trees are a machine learning approach that partitions the outcome of interest using a series of binary rules based on covariates and computes a summary statistic of the outcome in each partition (214). Both random forest and boosted regression methods fit many regression trees and combine the results to overcome issues of overfitting and bias; random forest regression combines the results of these trees at the end of the process, in this case using bootstrapping, to obtain an overall result, whereas boosted regression trees refit regression trees iteratively, weighting observations that were poorly modelled to produce a final, parsimonious additive model (219). Unlike neural networks, regression trees may be used to explore the association between covariates and the outcome by calculating their relative importance in the model. Both papers included spatial covariates as inputs, such as distance and number of people moving between countries or cities, to account for connectivity.

Regression tree methods are more flexible than standard regression; as with other machine learning methods, the model structure does not need to be defined prior to model fitting, the method 'learns' from the data. Models allow multiple interactions and complex models to be explored and can provide more robust predictions than statistical regression models (214). Although the relative importance of a covariate can be calculated from this method, the binary separation of covariates means these models cannot explore continuous relationships. This method should only be used in studies that require accurate predictions from a large amount of data but are not suitable for causal inference or to understand the process of transmission.

2.3 Bayesian networks

Bayesian networks use directed acyclic graphs to illustrate the causal relationship between an outcome and covariates. Each variable is represented by a node and these nodes are connected by arrows representing causal pathways. Bayesian networks assign conditional probabilities to each of these arrows, describing the relationship between nodes. One study was identified that used a spatial Bayesian network to understand the relationship between environmental covariates and malaria (220). Researchers included the total number of cases in connected regions within the network. Bayesian networks require prior understanding of the research area to determine where causal pathways exist. They also require numerical variables to be categorised before the conditional probabilities could be obtained, losing information about continuous relationships.

3. Mechanistic models

3.1 Compartmental model

Compartmental models describe the movement of hosts (in this case, humans) through infectious stages (e.g. susceptible, infected and recovered) to replicate the process of disease transmission in the real world. The movement between infectious stages is described using differential equations and may include biological, environmental, and entomological factors as parameters that explain this process. Some studies also include a compartmental model for vectors that interacts with the human model. The choice of compartments depends on how the disease progresses. A graphical representation of a compartmental model used to explain mosquito-borne disease is given in Figure 2.

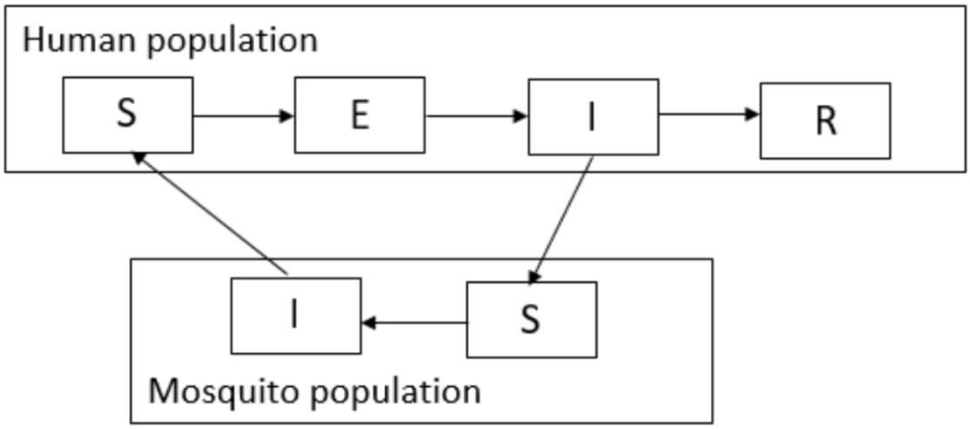


Figure 2. A visual representation of an SEIR-SEI compartmental model. At any given time, humans are susceptible (S), exposed but not infectious (E), infectious (I) or recovered (R) from a given mosquito-borne disease, and mosquitoes are either susceptible (S) or infectious (I). Arrows represent potential movement or interaction between compartments: susceptible humans can become exposed after interaction with an infectious mosquito, and a susceptible mosquito can become infectious after an interaction with an infectious human.

We identified 5 studies that used compartmental models. Three included a spatial parameters within the model equations to allow for connectivity between regions within the transmission process: two used an exponential distance-decay function to estimate the force of infection between individuals (221,222), the other included a 'source term' to account for infected individuals travelling into the area from other cities (223). The remaining two studies used statistical models to estimate parameters that were included in the compartmental model equations: one estimated mosquito abundance as a proxy for biting patterns using a spatial GAM containing a spatially structured random field, accounting for distance between mosquito traps (224), the other used a GLMM with CAR random effects to estimate a coefficient accounting for the movement of infectious individuals from neighbouring regions.

Compartmental models are useful to investigate the process of transmission. These models require less data as parameters can be taken from previous experiments when data is not available, making them useful when studying rare or (re-)emerging diseases. Mechanistic

models require knowledge of the biological and entomological processes that drive transmission as incorrect parameter specification may lead to invalid results. The models can extrapolate beyond the data, making them useful to investigate the impact of vector control and eradication programmes, but only within the specific settings they have been parameterised for.

3.2 Metapopulation model

Metapopulations are spatially separate groups of the same population that are connected in some way. In metapopulation models, these are generally towns, villages, or clusters within a region of interest. Metapopulation models identified in this review modelled mosquito-borne transmission within each metapopulation whilst allowing for migration between these regions. We identified 24 metapopulation modelling studies; 20 fitted separate compartmental models to each metapopulation (217,225–243) and 4 estimated disease transmission but did not explicitly model movement between compartments (244–247). There were 17 metapopulation models that used a movement matrix to represent the number of people (and in one case, vectors and livestock (229)) moving between metapopulations. These were created using movement models, such as gravity or radiation models, that aimed to mimic human commuting behaviour (226,228–232,243,245,246), movement data collected from geo-located tweets (232,239,247), air travel (234,235) or public transportation (217), distance (227), or using a fixed value based on the type of neighbourhood (e.g. city or suburb) (237,238). There were 7 metapopulation models that used a spatial parameter to account for connectivity, including it within the model process rather than explicitly modelling movement between metapopulations (225,233,236,240–242,244). O'Reilly et al. (233) fitted separate compartmental models for each location for the population that remained and the population that left.

Due to their design, metapopulations are ideal for accounting for complex mobility patterns by explicitly allowing movement around a network, however this added complexity means the method is computationally intensive and often difficult to implement. As with other mechanistic models, metapopulation models can 'borrow' parameters from other experiments and so require less data than statistical or machine learning, but prior knowledge of the transmission process is required. Although human movement can be included into a metapopulation model, obtaining unbiased, adequate data can be difficult. It is important to consider the most appropriate data or proxy to represent the type of movement expected to drive transmission (short- vs. long-term, domestic vs. international, etc.). Mathematical models of human movement, such as gravity and radiation models, provide an appealing alternative when the process is thought to be driven by commuting but may not be useful in large, international studies.

3.3 Agent-based model (ABM)

ABMs simulate the movement of individuals (or agents) around a system whilst measuring their impact on the outcome. Agents may interact with one another (for example, infected mosquitoes may interact and infect susceptible humans at a given time step) and are tracked through infectious stages, either within a compartmental system or alternative methods. We

identified 4 studies using an ABM: three used a compartmental model for each agent (248–250), simulating movement between infectious stages as well as around the spatial system based on a movement matrix, the other model did not explicitly use a compartmental model but estimated the number of new infections based on where agents were in the system and the distance between them (251). The compartmental-based ABMs used a movement matrix to describe the movement of agents around the study area which were informed by commuting patterns for humans (based on mobile phone or survey data, estimating home and work/school addresses (249,250)) or based on flight patterns of mosquitoes (either weighted by proximity to humans and breeding grounds (248), or into adjacent cells (249)).

As with metapopulations models, the ABM structure is useful for settings where human and/or vector movement is thought to be an important driver of disease and allows this to be explicitly included within a model. Mixing between hosts and vectors is also explicitly measured, rather than implicitly through parameters as is the case in metapopulation models, allowing more complex mixing dynamics to be explored. Due to their extra complexity, ABMs are extremely computationally intensive and can be hard to fit. Inclusion of movement matrices requires appropriate data which may be difficult or, in the case of vector movement, impossible to obtain.

3.4 Generalised inverse infection model (GIIM)

The GIIM was developed to estimate parameters of the transmission process using outbreak data rather than simulating the outbreak itself. The model is based on a network, similar to the metapopulation model, where each node of the network represents a metapopulation which has experienced an outcome. Nodes may also be separated into disease compartments. Nodes are connected to one another by edges. Unlike the metapopulation model, it is the edge parameters that are estimated and the process of transmission rather than the outcome itself (which has already been observed). There was one study that used a GIIM to investigate the emergence of Zika in the Americas (252). Each node was a country or state classified using a compartmental model (susceptible or infected based on case data) and each edge represented the probability of importation from an infectious node to a susceptible one. The GIIM is useful when the process of transmission is not fully understood and may be used to inform other mechanistic models. The model requires accurate outcome data (in this case, information regarding Zika outbreaks) so cannot be used to investigate areas with poor surveillance or places currently experiencing an outbreak.

References

1. Wu PC, Lay JG, Guo HR, Lin CY, Lung SC, Su HJ. Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. Sci Total Environ. 2009;407(7):2224–33.

2. DeGroote JP, Sugumaran R. National and regional associations between human West Nile virus incidence and demographic, landscape, and land use conditions in the coterminous United States. Vector-Borne Zoonotic Dis. 2012;12(8):657–65.

3. Chuang TW, Ng KC, Nguyen TL, Chaves LF. Epidemiological characteristics and space-time analysis of the 2015 dengue outbreak in the metropolitan region of Tainan city, Taiwan. Int J Environ Res Public Health. 2018;15(3):396.

4. Wen T, Hsu C, Hu M. Evaluating neighborhood structures for modeling intercity diffusion of large-scale dengue epidemics. Int J Health Geogr. 2018;17(9).

5. Ashmore P, Lindahl JF, Colón-González FJ, Sinh Nam V, Quang Tan D, Medley GF. Spatiotemporal and socioeconomic risk factors for Dengue at the Province Level in Vietnam, 2013–2015: clustering analysis and regression model. Trop Med Infect Dis. 2020;5(2):81.

6. Xu Z, Bambrick H, Pongsumpun P, Tang IM, Yakob L, Devine G, et al. Does Bangkok have a central role in the dengue dynamics of Thailand? Parasit Vectors. 2020;13(1):1–9.

7. Impoinvil DE, Solomon T, Schluter WW, Rayamajhi A, Bichha RP, Shakya G, et al. The spatial heterogeneity between Japanese encephalitis incidence distribution and environmental variables in Nepal. PloS One. 2011;6(7):e22192.

8. Astutik S, Rahayudi B, Iskandar A, Fitriani R. Bayesian spatial-temporal autologistic regression model on dengue hemorrhagic fever in East Java, Indonesia. Appl Math Sci. 2013;7(9):435–43.

9. Kazembe LN. Spatial modelling and risk factors of malaria incidence in northern Malawi. Acta Trop. 2007;102(2):126–37.

10. Wang X, Su L, Zhu H, Hu W, An J, Wang C, et al. Long-Term Epidemiological Dynamics of Japanese Encephalitis Infection in Gansu Province, China: A Spatial and Temporal Analysis. Am J Trop Med Hyg. 2020 Nov 4;103(5):2065–76.

11. Tao H, Wang K, Zhuo L, Li X, Li Q, Liu Y, et al. A comprehensive framework for studying diffusion patterns of imported dengue with individual-based movement data. Int J Geogr Inf Sci. 2019;34(3):604–24.

12. Kraemer MU, Faria NR, Reiner Jr RC, Golding N, Nikolay B, Stasse S, et al. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. Lancet Infect Dis. 2017;17(3):330–8.

13. Chen Y, Yang Z, Jing Q, Huang J, Guo C, Yang K, et al. Effects of natural and socioeconomic factors on dengue transmission in two cities of China from 2006 to 2017. Sci Total Environ. 2020 Jul 1;724:138200.

14. Gomes MFC, Codeço CT, Bastos LS, Lana RM. Measuring the contribution of human mobility to malaria persistence. Malar J. 2020 Nov 11;19(1):404.

15. Griffith DA. A comparison of six analytical disease mapping techniques as applied to West Nile Virus in the coterminous United States. Int J Health Geogr. 2005;4(1):18.

16. Griffith DA. A linear regression solution to the spatial autocorrelation problem. J Geogr Syst. 2000 Jul 1;2(2):141–56.

17. Wen T-H, Tsai C-T. Evaluating the role of disease importation in the spatiotemporal transmission of indigenous dengue outbreak. Appl Geogr. 2016;76:137–46.

18. Thomson MC, Connor SJ, D'Alessandro U, Rowlingson B, Diggle P, Cresswell M, et al. Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results. Am J Trop Med Hyg. 1999/08/04 ed. 1999 Jul;61(1):2–8.

19. Brunsdon C, Fotheringham S, Charlton M. Geographically Weighted Regression. J R Stat Soc Ser Stat. 1998;47(3):431–43.

20. Delmelle E, Hagenlocher M, Kienberger S, Casas I. A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. Acta Trop. 2016;164:169–76.

21. Robertson C, Pant DK, Joshi DD, Sharma M, Dahal M, Stephen C. Comparative spatial dynamics of Japanese encephalitis and acute encephalitis syndrome in Nepal. PloS One. 2013;8(7):e66168.

22. Ehlkes L, Krefis AC, Kreuels B, Krumkamp R, Adjei O, Ayim-Akonor M, et al. Geographically weighted regression of land cover determinants of Plasmodium falciparum transmission in the Ashanti Region of Ghana. Int J Health Geogr. 2014;13(1):1–11.

23. Atique S, Chan TC, Chen CC, Hsu CY, Iqtidar S, Louis VR, et al. Investigating spatio-temporal distribution and diffusion patterns of the dengue outbreak in Swat, Pakistan. J Infect Public Health. 2018;11(4):550–7.

24. Manyangadze T, Chimbari MJ, Macherera M, Mukaratirwa S. Micro-spatial distribution of malaria cases and control strategies at ward level in Gwanda district, Matabeleland South, Zimbabwe. Malar J. 2017;16(1):1–11.

25. Khormi HM, Kumar L. Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study. Sci Total Environ. 2011;409(22):4713–9.

26. Acharya BK, Cao CX, Lakes T, Chen W, Naeem S, Pandit S. Modeling the spatially varying risk factors of dengue fever in Jhapa district, Nepal, using the semi-parametric geographically weighted regression model. Int J Biometeorol. 2018;62(11):1973–86.

27. Hasyim H, Nursafingi A, Haque U, Montag D, Groneberg DA, Dhimal M, et al. Spatial modelling of malaria cases associated with environmental factors in South Sumatra, Indonesia. Malar J. 2018;17(1):1–15.

28. Homan T, Maire N, Hiscox A, Di Pasquale A, Kiche I, Onoka K, et al. Spatially variable risk factors for malaria in a geographically heterogeneous landscape, western Kenya: An explorative study. Malar J. 2016;15(1):1.

29. Hsueh YH, Lee J, Beltz L. Spatio-temporal patterns of dengue fever cases in Kaoshiung City, Taiwan, 2003–2008. Appl Geogr. 2012;34:587–94.

30. Ren H, Wu W, Li T, Yang Z. Urban villages as transfer stations for dengue fever epidemic: A case study in the Guangzhou, China. PLoS Negl Trop Dis. 2019;13(4):e0007350.

31. Lin C-H, Wen T-H. Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. Int J Environ Res Public Health. 2011;8(7):2798–815.

32. Grillet ME, Barrera R, Martínez JE, Berti J, Fortin MJ. Disentangling the effect of local and global spatial variation on a mosquito-borne infection in a neotropical heterogeneous environment. Am J Trop Med Hyg. 2010;82(2):194–201.

33. Yang D, Xu C, Wang J, Zhao Y. Spatiotemporal epidemic characteristics and risk factor analysis of malaria in Yunnan Province, China. BMC Public Health. 2017;17(1):1–10.

34. Gopal S, Ma Y, Xin C, Pitts J, Were L. Characterizing the spatial determinants and prevention of malaria in Kenya. Int J Environ Res Public Health. 2019;16(24):5078.

35. Ren H, Zheng L, Li Q, Yuan W, Lu L. Exploring determinants of spatial variations in the dengue fever epidemic using geographically weighted regression model: A case study in the joint Guangzhou-Foshan area, China, 2014. Int J Environ Res Public Health. 2017;14(12):1518.

36. Halimi M, Farajzadeh M, Delavari M, Takhtardeshir A, Moradi A. Modelling spatial relationship between climatic conditions and annual parasite incidence of malaria in southern part of Sistan&Balouchistan Province of Iran using spatial statistic models. Asian Pac J Trop Dis. 2014;4(Supplement1):S167–72.

37. Ge Y, Song Y, Wang J, Liu W, Ren Z, Peng J, et al. Geographically weighted regression-based determinants of malaria incidences in northern China. Trans GIS. 2017;21(5):934–53.

38. de Oliveira Padilha MA, de Oliveira Melo J, Romano G, de Lima MVM, Alonso WJ, Sallum MAM, et al. Comparison of malaria incidence rates and socioeconomic-

environmental factors between the states of Acre and Rondonia: a spatio-temporal modelling study. Malar J. 2019;18(1):04.

39. Anjos RS dos, Nóbrega RS, Ferreira H dos S, Lacerda AP de, Sousa-Neves N de. Exploring local and global regression models to estimate the spatial variability of Zika and Chikungunya cases in Recife, Brazil. Rev Soc Bras Med Trop. 2020;53.

40. Chiaravalloti-Neto F, Pereira M, Fávaro EA, Dibo MR, Mondini A, Rodrigues-Junior AL, et al. Assessment of the relationship between entomologic indicators of Aedes aegypti and the epidemic occurrence of dengue virus 3 in a susceptible population, São José do Rio Preto, São Paulo, Brazil. Acta Trop. 2015;142:167–77.

41. Farinelli EC, Baquero OS, Stephan C, Chiaravalloti-Neto F. Low socioeconomic condition and the risk of dengue fever: A direct relationship. Acta Trop. 2018;180:47–57.

42. Charlwood JD, Tomas EVE, Braganca M, Cuamba N, Alifrangis M, Stanton M. Malaria prevalence and incidence in an isolated, meso-endemic area of Mozambique. Peerj. 2015 Nov;3:22.

43. Ugwu CLJ, Zewotir T. Evaluating the Effects of Climate and Environmental Factors on Under-5 Children Malaria Spatial Distribution Using Generalized Additive Models (GAMs). J Epidemiol Glob Health. 2020 Aug;10(4):304–14.

44. Mutucumarana CP, Bodinayake CK, Nagahawatte A, Devasiri V, Kurukulasooriya R, Anuradha T, et al. Geospatial analysis of dengue emergence in rural areas in the Southern Province of Sri Lanka. Trans R Soc Trop Med Hyg. 2020 Jun 1;114(6):408–14.

45. Vazquez-Prokopec GM, Kitron U, Montgomery B, Horne P, Ritchie SA. Quantifying the spatial dimension of dengue virus epidemic spread within a tropical urban environment. PLoS Negl Trop Dis. 2010;4(12):e920.

46. Braga C, Luna CF, Martelli CM, de Souza WV, Cordeiro MT, Alexander N, et al. Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. Acta Trop. 2010 Mar;113(3):234–40.

47. Honório NA, Nogueira RMR, Codeço CT, Carvalho MS, Cruz OG, Magalhães M de AFM, et al. Spatial Evaluation and Modeling of Dengue Seroprevalence and Vector Density in Rio de Janeiro, Brazil. PLoS Negl Trop Dis. 2009 Nov 10;3(11):e545.

48. Siqueira-Junior JB, Maciel IJ, Barcellos C, Souza WV, Carvalho MS, Nascimento NE, et al. Spatial point analysis based on dengue surveys at household level in central Brazil. BMC Public Health. 2008 Oct 20;8(1):361.

49. Kazembe LN, Mathanga DP. Estimating risk factors of urban malaria in Blantyre, Malawi: A spatial regression analysis. Asian Pac J Trop Biomed. 2016;6(5):376–81.

50. Chien L-C, Sy F, Pérez A. Identifying high risk areas of Zika virus infection by meteorological factors in Colombia. BMC Infect Dis. 2019 Oct 24;19(1):888.

51. Hundessa S, Williams G, Li S, Liu DL, Cao W, Ren H, et al. Projecting potential spatial and temporal changes in the distribution of Plasmodium vivax and Plasmodium falciparum malaria in China with climate change. Sci Total Environ. 2018;627:1285–93.

52. Cissoko M, Sagara I, Sankaré MH, Dieng S, Guindo A, Doumbia Z, et al. Geo-Epidemiology of Malaria at the Health Area Level, Dire Health District, Mali, 2013–2017. Int J Environ Res Public Health. 2020 Jan;17(11):3982.

53. Watts MJ, Kotsila P, Mortyn PG, Sarto i Monteys V, Urzi Brancati C. Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico. Int J Health Geogr. 2020 Nov 2;19(1):44.

54. Wood SN. Generalized additive models: an introduction with R. CRC press; 2017.

55. Jain R, Sontisirikit S, Iamsirithaworn S, Prendinger H. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect Dis. 2019;19(1):272.

56. Da Silva-Nunes M, Codeço CT, Malafronte RS, Da Silva NS, Juncansen C, Muniz PT, et al. Malaria on the amazonian frontier: Transmission dynamics, risk factors, spatial distribution, and prospects for control. Am J Trop Med Hyg. 2008;79(4):624–35.

57. Parra MCP, Fávaro EA, Dibo MR, Mondini A, Eiras ÁE, Kroon EG, et al. Using adult Aedes aegypti females to predict areas at risk for dengue transmission: A spatial case-control study. Acta Trop. 2018;182:43–53.

58. Laguna F, Grillet ME, León JR, Ludeña C. Modelling malaria incidence by an autoregressive distributed lag model with spatial component. Spat Spatio-Temporal Epidemiol. 2017;22:27–37.

59. Tevie J, Bohara A, Valdez RB. Examination of the geographical variation in human West Nile virus: A spatial filtering approach. Epidemiol Infect. 2014;142(12):2522–9.

60. Kim S, Kim Y. Spatially Filtered Multilevel Analysis on Spatial Determinants for Malaria Occurrence in Korea. Int J Environ Res Public Health. 2019 Apr;16(7):11.

61. Salami D, Capinha C, Martins M do RO, Sousa CA. Dengue importation into Europe: a network connectivity-based approach. PloS One. 2020;15(3):e0230274.

62. Ramadona AL, Tozan Y, Lazuardi L, Rocklöv J. A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. PLoS Negl Trop Dis. 2019;13(4):e0007298.

63. Dieng S, Ba EH, Cissé B, Sallah K, Guindo A, Ouedraogo B, et al. Spatio-temporal variation of malaria hotspots in Central Senegal, 2008–2012. BMC Infect Dis. 2020 Jun 17;20(1):424.

64. Maheu-Giroux M, Castro MC. Impact of community-based larviciding on the prevalence of malaria infection in Dar es Salaam, Tanzania. PLoS ONE. 2013;8(8):e71638.

65. Flórez-Lozano K, Navarro-Lechuga E, Llinás-Solano H, Tuesca-Molina R, Sisa-Camargo A, Mercado-Reyes M, et al. Spatial distribution of the relative risk of Zika virus disease in Colombia during the 2015–2016 epidemic from a Bayesian approach. Int J Gynecol Obstet. 2020;148(S2):55–60.

66. Ferreira GS, Schmidt AM. Spatial modelling of the relative risk of dengue fever in Rio de Janeiro for the epidemic period between 2001 and 2002. Braz J Probab Stat. 2006;20:29–47.

67. Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, Alegana VA, et al. The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: A spatial and temporal analysis of transmission intensity. The Lancet. 2014;383(9930):1739–47.

68. Ssempiira J, Kissa J, Nambuusi B, Kyozira C, Rutazaana D, Mukooyo E, et al. The effect of case management and vector-control interventions on space-time patterns of malaria incidence in Uganda. Malar J. 2018/04/14 ed. 2018 Apr 12;17(1):162.

69. Villalta D, Guenni L, Rubio-Palis Y, Ramírez Arbeláez R. Bayesian space-time modeling of malaria incidence in Sucre state, Venezuela. AStA Adv Stat Anal. 2013;97(2):151–71.

70. Jaya I, Abdullah AS, Hermawan E, Ruchjana BN. Bayesian Spatial Modeling and Mapping of Dengue Fever: A Case Study of Dengue Feverin the City of Bandung, Indonesia. Int J Appl Math Stat. 2016;54(3):94–103.

71. Thway AM, Rotejanaprasert C, Sattabongkot J, Lawawirojwong S, Thi A, Hlaing TM, et al. Bayesian spatiotemporal analysis of malaria infection along an international border: Hlaingbwe Township in Myanmar and Tha-Song-Yang District in Thailand. Malar J. 2018;17(1):428.

72. Jaya I, Folmer H. Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia. J Geogr Syst. 2020 Jan;22(1):105–42.

73. Rouamba T, Samadoulougou S, Tinto H, Alegana VA, Kirakoya-Samadoulougou F. Bayesian Spatiotemporal Modeling of Routinely Collected Data to Assess the Effect of Health Programs in Malaria Incidence During Pregnancy in Burkina Faso. Sci Rep. 2020;10(1):14.

74. Rotejanaprasert C, Ekapirat N, Areechokchai D, Maude RJ. Bayesian spatiotemporal modeling with sliding windows to correct reporting delays for real-time dengue surveillance in Thailand. Int J Health Geogr. 2020;19(1):1–13.

75. Snow RW, Kibuchi E, Karuri SW, Sang G, Gitonga CW, Mwandawiro C, et al. Changing malaria prevalence on the Kenyan coast since 1974: Climate, drugs and vector control. PLoS ONE. 2015;10(6):e0128792.

76. Reid HL, Haque U, Roy S, Islam N, Clements ACA. Characterizing the spatial and temporal variation of malaria incidence in Bangladesh, 2007. Malar J. 2012;11:170.

77. Aswi A, Cramb S, Duncan E, Hu W, White G, Mengersen K. Climate variability and dengue fever in Makassar, Indonesia: Bayesian spatio-temporal modelling. Spat Spatio-Temporal Epidemiol. 2020;33:100335.

78. Lowe R, Barcellos C, Coelho CAS, Bailey TC, Coelho GE, Graham R, et al. Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. Lancet Infect Dis. 2014;14(7):619–26.

79. Wang G, Minnis RB, Belant JL, Wax CL. Dry weather induces outbreaks of human West Nile virus infections. BMC Infect Dis. 2010 Feb 24;10(1):38.

80. Wimberly MC, Hildreth MB, Boyte SP, Lindquist E, Kightlinger L. Ecological niche of the 2003 West Nile virus epidemic in the northern Great Plains of the United States. PLoS ONE. 2008;3(12):e3744.

81. Ouédraogo M, Rouamba T, Samadoulougou S, Kirakoya-Samadoulougou F. Effect of Free Healthcare Policy for Children under Five Years Old on the Incidence of Reported Malaria Cases in Burkina Faso by Bayesian Modelling: "Not only the Ears but also the Head of the Hippopotamus". Int J Environ Res Public Health. 2020 Jan;17(2):417.

82. Umer MF, Zofeen S, Majeed A, Hu W, Qi X, Zhuang G. Effects of Socio-Environmental Factors on Malaria Infection in Pakistan: A Bayesian Spatial Analysis. Int J Environ Res Public Health. 2019;16(8):1365.

83. Chen Z, Shi L, Zhou XN, Xia ZG, Bergquist R, Jiang QW. Elimination of malaria due to Plasmodium vivax in central part of the People's Republic of China: Analysis and prediction based on modelling. Geospatial Health. 2014;9(1):169–77.

84. Abd Naeeim NS, Rahman NA. Estimating relative risk for dengue disease in Peninsular Malaysia using INLA. Malays J Fundam Appl Sci. 2017 Oct;13(4):721–7.

85. Alegana VA, Atkinson PM, Wright JA, Kamwi R, Uusiku P, Katokele S, et al. Estimation of malaria incidence in northern Namibia in 2009 using Bayesian conditional-autoregressive spatial-temporal models. Spat Spatio-Temporal Epidemiol. 2013;7:25–36.

86. Mukhsar, Abapihi B, Sani A, Cahyono E, Adam P, Abdullah FA. Extended convolution model to bayesian spatio-temporal for diagnosing the DHF endemic locations. J Interdiscip Math. 2016;19(2):233–44.

87. Husnina Z, Clements ACA, Wangdi K. Forest cover and climate as potential drivers for dengue fever in Sumatra and Kalimantan 2006–2016: a spatiotemporal analysis. Trop Med Int Health. 2019;24(7):888–98.

88. Lekdee K, Ingsrisawang L. Generalized linear mixed models with spatial random effects for spatio-temporal data: An application to dengue fever mapping. J Math Stat. 2013;9(2):137–43.

89. Nkurunziza H, Gebhardt A, Pilz J. Geo-additive modelling of malaria in Burundi. Malar J. 2011;10:234.

90. Zayeri F, Salehi M, Pirhosseini H. Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran. Asian Pac J Trop Med. 2011;4(12):985–92.

91. Ssempiira J, Kissa J, Nambuusi B, Mukooyo E, Opigo J, Makumbi F, et al. Interactions between climatic changes and intervention effects on malaria spatio-temporal dynamics in Uganda. Parasite Epidemiol Control. 2018;3(3):e00070.

92. Zhao X, Cao M, Feng HH, Fan H, Chen F, Feng Z, et al. Japanese encephalitis risk and contextual risk factors in Southwest China: A Bayesian hierarchical spatial and spatiotemporal analysis. Int J Environ Res Public Health. 2014;11(4):4201–17.

93. Martínez-Bello DA, López-Quílez A, Prieto AT. Joint estimation of relative risk for dengue and zika infections, Colombia, 2015-2016. Emerg Infect Dis. 2019;25(6):1118–26.

94. Zacarias OP, Andersson M. Mapping malaria incidence distribution that accounts for environmental factors in Maputo Province - Mozambique. Malar J. 2010;9(1):79.

95. Huang F, Zhou S, Zhang S, Zhang H, Li W. Meteorological factors - Based spatio-temporal mapping and predicting malaria in central China. Am J Trop Med Hyg. 2011;85(3):560–7.

96. Alegana VA, Wright JA, Nahzat SM, Butt W, Sediqi AW, Habib N, et al. Modelling the incidence of Plasmodium vivax and Plasmodium falciparum malaria in Afghanistan 2006-2009. PLoS ONE. 2014;9(7):e102304.

97. Restrepo AC, Baker P, Clements ACA. National spatial and temporal patterns of notified dengue cases, Colombia 2007-2010. Trop Med Int Health. 2014;19(7):863–71.

98. Santos-Vega M, Bouma MJ, Kohli V, Pascual M. Population Density, Climate Variables and Poverty Synergistically Structure Spatial Risk in Urban Malaria in India. PLoS Negl Trop Dis. 2016;10(12):e0005155.

99. Lowe R, Cazelles B, Paul R, Rodó X. Quantifying the added value of climate information in a spatio-temporal dengue model. Stoch Environ Res Risk Assess. 2016;30(8):2067–78.

100. Lowe R, Chirombo J, Tompkins AM. Relative importance of climatic, geographic and socio-economic determinants of malaria in Malawi. Malar J. 2013;12(1):416.

101. Sani A, Abapihi B, Mukhsar M, Kadir K. Relative risk analysis of dengue cases using convolution extended into spatio-temporal model. J Appl Stat. 2015;42(11):2509–19.

102. Samat NA, Pei Zhen W. Relative risk estimation for dengue disease mapping in Malaysia based on Besag, York and Mollié model. Pertanika J Sci Technol. 2017;25(3):759–66.

103. Martínez-Bello DA, López-Quílez A, Torres Prieto A. Relative risk estimation of dengue disease at small spatial scale. Int J Health Geogr. 2017 Aug 15;16(1):31.

104. Kristiani F, Yong B, Irawan R. Relative risk estimation of dengue disease in Bandung, Indonesia, using poisson-gamma and bym models considering the severity level. J Teknol. 2016;78(11):57–64.

105. Rouamba T, Samadoulougou S, Tinto H, Alegana VA, Kirakoya-Samadoulougou F. Severe-malaria infection and its outcomes among pregnant women in Burkina Faso health-districts: Hierarchical Bayesian space-time models applied to routinely-collected data from 2013 to 2018. Spat Spatio-Temporal Epidemiol. 2020 Jun 1;33:100333.

106. Manh BH, Clements ACA, Thieu NQ, Hung NM, Hung LX, Hay SI, et al. Social and environmental determinants of malaria in space and time in Viet Nam. Int J Parasitol. 2011;41(1):109–16.

107. Clements AC, Barnett AG, Cheng ZW, Snow RW, Zhou HN. Space-time variation of malaria incidence in Yunnan province, China. Malar J. 2009/08/04 ed. 2009 Jul 31;8:180.

108. de Almeida AS, de Andrade Medronho R, Ortiz Valencia LI. Spatial analysis of dengue and the socioeconomic context of the city of Rio de Janeiro (Southeastern Brazil). Rev Saude Publica. 2009;43(4):666–73.

109. Honorato T, Lapa PPA, Sales CMM, Reis-Santos B, Tristão-Sá R, Bertolde AI, et al. Spatial analysis of distribution of dengue cases in Espírito Santo, Brazil, in 2010: Use of Bayesian model. Rev Bras Epidemiol. 2014;17:150–9.

110. Noh M, Lee Y, Oh S, Chu C, Gwack J, Youn SK, et al. Spatial and Temporal Distribution of Plasmodium vivax Malaria in Korea Estimated with a Hierarchical Generalized Linear Model. Osong Public Health Res Perspect. 2012;3(4):192–8.

111. Phanitchat T, Zhao B, Haque U, Pientong C, Ekalaksananan T, Aromseree S, et al. Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. BMC Infect Dis. 2019 Aug 23;19(1):743.

112. Wangdi K, Clements ACA, Du T, Nery SV. Spatial and temporal patterns of dengue infections in Timor-Leste, 2005-2013. Parasit Vectors. 2018;11(1):9.

113. Ouédraogo M, Samadoulougou S, Rouamba T, Hien H, Sawadogo JEM, Tinto H, et al. Spatial distribution and determinants of asymptomatic malaria risk among children under 5 years in 24 districts in Burkina Faso. Malar J. 2018 Dec 7;17(1):460.

114. Kikuti M, Cunha GM, Paploski IAD, Kasper AM, Silva MMO, Tavares AS, et al. Spatial Distribution of Dengue in a Brazilian Urban Slum Setting: Role of Socioeconomic Gradient in Disease Risk. PLoS Negl Trop Dis. 2015 Jul 21;9(7):e0003937.

115. Costa JV, Donalisio MR, Silveira LV de A. Spatial distribution of dengue incidence and socio-environmental conditions in Campinas, Sao Paulo State, Brazil, 2007. Cad Saude Publica. 2013;29(8):1522–32.

116. Teixeira TR de A, Cruz OG. Spatial modeling of dengue and socio-environmental indicators in the city of Rio de Janeiro, Brazil. Cad Saude Publica. 2011;27(3):591–602.

117. Hu WB, Clements A, Williams G, Tong SL, Mengersen K. Spatial Patterns and Socioecological Drivers of Dengue Fever Transmission in Queensland, Australia. Environ Health Perspect. 2012 Feb;120(2):260–6.

118. Abellana R, Ascaso C, Aponte J, Saute F, Nhalungo D, Nhacolo A, et al. Spatio-seasonal modeling of the incidence rate of malaria in Mozambique. Malar J. 2008 Oct 31;7(1):228.

119. Mabaso MLH, Vounatsou P, Midzi S, Da Silva J, Smith T. Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. Int J Health Geogr. 2006;5:20.

120. Martínez-Bello DA, López-Quílez A, Torres Prieto A. Spatio-Temporal Modeling of Zika and Dengue Infections within Colombia. Int J Environ Res Public Health. 2018 Jul;15(7):1376.

121. Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, et al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Comput Geosci. 2011;37(3):371–81.

122. Nobre AA, Schmidt AM, Lopes HF. Spatio-temporal models for mapping the incidence of malaria in Pará. Environmetrics. 2005;16(3):291–304.

123. Bett B, Grace D, Lee HS, Lindahl J, Nguyen-Viet H, Phuc P-D, et al. Spatiotemporal analysis of historical records (2001–2012) on dengue fever in Vietnam and development of a statistical model for forecasting risk. PLOS ONE. 2019 Nov 27;14(11):e0224353.

124. McHale TC, Romero-Vivas CM, Fronterre C, Arango-Padilla P, Waterlow NR, Nix CD, et al. Spatiotemporal Heterogeneity in the Distribution of Chikungunya and Zika Virus Case Incidences during their 2014 to 2016 Epidemics in Barranquilla, Colombia. Int J Environ Res Public Health. 2019 Jan;16(10):1759.

125. Martínez-Bello D, López-Quílez A, Prieto AT. Spatiotemporal modeling of relative risk of dengue disease in Colombia. Stoch Environ Res Risk Assess. 2018;32(6):1587–601.

126. Lowe R, Bailey TC, Stephenson DB, Jupp TE, Graham RJ, Barcellos C, et al. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. Stat Med. 2013;32(5):864–83.

127. Wijayanti SPM, Porphyre T, Chase-Topping M, Rainey SM, McFarlane M, Schnettler E, et al. The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia. PLoS Negl Trop Dis. 2016 Sep;10(9):15.

128. Mabaso MLH, Craig M, Vounatsou P, Smith T. Towards empirical description of malaria seasonality in southern Africa: The example of Zimbabwe. Trop Med Int Health. 2005;10(9):909–18.

129. Adin A, Martínez-Bello DA, López-Quílez A, Ugarte MD. Two-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia. PLOS ONE. 2018 Sep 11;13(9):e0203382.

130. Achcar JA, Martinez EZ, Souza AD, Tachibana VM, Flores EF. Use of Poisson spatiotemporal regression models for the Brazilian Amazon Forest: malaria count data. Rev Soc Bras Med Trop. 2012/01/11 ed. 2011 Nov;44(6):749–54.

131. Hanandita W, Tampubolon G. Geography and social distribution of malaria in Indonesian Papua: a cross-sectional study. Int J Health Geogr. 2016 Apr 12;15(1):13.

132. Zhang SB, Hu WB, Qi X, Zhuang GH. How Socio-Environmental Factors Are Associated with Japanese Encephalitis in Shaanxi, China-A Bayesian Spatial Analysis. Int J Environ Res Public Health. 2018 Apr;15(4):13.

133. Battle KE, Lucas TCD, Nguyen M, Howes RE, Nandi AK, Twohig KA, et al. Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. The Lancet. 2019;394(10195):332–43.

134. Abd Naeeim NS, Rahman NA, Fahimi FAM. A spatial-temporal study of dengue in Peninsular Malaysia for the year 2017 in two different space-time model. J Appl Stat. 2020 Mar;47(4):739–56.

135. Kleinschmidt I, Sharp B, Mueller I, Vounatsou P. Rise in malaria incidence rates in South Africa: A small-area spatial analysis of variation in time trends. Am J Epidemiol. 2002 Feb;155(3):257–64.

136. Bisanzio D, Mutuku F, LaBeaud AD, Mungai PL, Muinde J, Busaidy H, et al. Use of prospective hospital surveillance data to define spatiotemporal heterogeneity of malaria risk in coastal Kenya. Malar J. 2015;14(1):482.

137. Nakhapakorn K, Sancharoen W, Mutchimwong A, Jirakajohnkool S, Onchang R, Rotejanaprasert C, et al. Assessment of Urban Land Surface Temperature and Vertical City Associated with Dengue Incidences. Remote Sens. 2020 Jan;12(22):3802.

138. Semakula M, Niragire F, Faes C. Bayesian spatio-temporal modeling of malaria risk in Rwanda. PLOS ONE. 2020 Sep 10;15(9):e0238504.

139. Akter R, Hu W, Gatton M, Bambrick H, Cheng J, Tong S. Climate variability, socio-ecological factors and dengue transmission in tropical Queensland, Australia: A Bayesian spatial analysis. Environ Res. 2020 Oct 4;110285.

140. Kristiani F, Claudia Y, Yong B, Hilsdon A-M. A comparative analysis of frequentist and Bayesian approaches to estimate dengue disease transmission in Bandung-Indonesia. J Stat Manag Syst. 2020 Aug 5;0(0):1–17.

141. Ye J, Moreno-Madriñán MJ. Comparing different spatio-temporal modeling methods in dengue fever data analysis in Colombia during 2012–2015. Spat Spatio-Temporal Epidemiol. 2020 Aug 1;34:100360.

142. da Conceição Araújo D, Dos Santos AD, Lima SVMA, Vaez AC, Cunha JO, de Araújo KCGM. Determining the association between dengue and social inequality factors in north-eastern Brazil: A spatial modelling. Geospatial Health. 2020;15(1).

143. Aswi A, Cramb S, Duncan E, Mengersen K. Evaluating the impact of a small number of areas on spatial estimation. Int J Health Geogr. 2020 Sep 25;19(1):39.

144. Jaya IGNM, Folmer H. Identifying Spatiotemporal Clusters by Means of Agglomerative Hierarchical Clustering and Bayesian Regression Analysis with Spatiotemporally Varying Coefficients: Methodology and Application to Dengue Disease in Bandung, Indonesia. Geogr Anal. 2020;

145. Gunderson AK, Kumar RE, Recalde-Coronel C, Vasco LE, Valle-Campos A, Mena CF, et al. Malaria Transmission and Spillover across the Peru–Ecuador Border: A Spatiotemporal Analysis. Int J Environ Res Public Health. 2020;17(20):7434.

146. Tsheten T, Clements ACA, Gray DJ, Wangchuk S, Wangdi K. Spatial and temporal patterns of dengue incidence in Bhutan: a Bayesian analysis. Emerg Microbes Infect. 9(1):1360–71.

147. Wangdi K, Canavati SE, Ngo TD, Nguyen TM, Tran LK, Kelly GC, et al. Spatial and Temporal Patterns of Malaria in Phu Yen Province, Vietnam, from 2005 to 2016. Am J Trop Med Hyg. 2020 Oct 7;103(4):1540–8.

148. Wangdi K, Xu Z, Suwannatrai AT, Kurscheid J, Lal A, Namgay R, et al. A spatio-temporal analysis to identify the drivers of malaria transmission in Bhutan. Sci Rep. 2020 Apr 27;10(1):7060.

149. Carabali M, Harper S, Lima Neto AS, Dos Santos de Sousa G, Caprara A, Restrepo BN, et al. Spatiotemporal distribution and socioeconomic disparities of dengue, chikungunya and Zika in two Latin American cities from 2007 to 2017. Trop Med Int Health. 2020;

150. Puggioni G, Couret J, Serman E, Akanda AS, Ginsberg HS. Spatiotemporal modeling of dengue fever risk in Puerto Rico. Spat Spatio-Temporal Epidemiol. 2020 Nov 1;35:100375.

151. Rue H, Held L. Gaussian Markov random fields: theory and applications. CRC press; 2005.

152. Matérn B. Spatial variation. Vol. 36. Springer Science & Business Media; 2013.

153. Mallya S, Sander B, Roy-Gagnon MH, Taljaard M, Jolly A, Kulkarni MA. Factors associated with human West Nile virus infection in Ontario: A generalized linear mixed modelling approach. BMC Infect Dis. 2018;18(1):141.

154. Gething PW, Elyazar IR, Moyes CL, Smith DL, Battle KE, Guerra CA, et al. A long neglected world malaria map: Plasmodium vivax endemicity in 2010. PLoS Negl Trop Dis. 2012;6(9):e1814.

155. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. Malar J. 2011;10(1):1–16.

156. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, Noor AM, et al. A world malaria map: Plasmodium falciparum endemicity in 2007. PLoS Med. 2009;6(3):e1000048.

157. Alegana VA, Atkinson PM, Lourenço C, Ruktanonchai NW, Bosco C, zu Erbach-Schoenberg E, et al. Advances in mapping malaria for elimination: fine resolution modelling of Plasmodium falciparum incidence. Sci Rep. 2016;6(1):1–14.

158. Fornace KM, Abidin TR, Alexander N, Brock P, Grigg MJ, Murphy A, et al. Association between landscape factors and spatial patterns of Plasmodium knowlesi infections in Sabah, Malaysia. Emerg Infect Dis. 2016;22(2):201–8.

159. Reid H, Vallely A, Taleo G, Tatem AJ, Kelly G, Riley I, et al. Baseline spatial distribution of malaria prior to an elimination programme in Vanuatu. Malar J. 2010;9(1):150.

160. Stensgaard AS, Vounatsou P, Onapa AW, Simonsen PE, Pedersen EM, Rahbek C, et al. Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: Predictors of risk and geographical patterns of co-endemicity. Malar J. 2011;10:298.

161. Amratia P, Psychas P, Abuaku B, Ahorlu C, Millar J, Oppong S, et al. Characterizing local-scale heterogeneity of malaria risk: A case study in Bunkpurugu-Yunyoo district in northern Ghana. Malar J. 2019;18(1):81.

162. Diggle P, Moyeed R, Rowlingson B, Thomson M. Childhood malaria in the Gambia: a case-study in model-based geostatistics. J R Stat Soc Ser C-Appl Stat. 2002;51:493–506.

163. Giorgi E, Sesay SSS, Terlouw DJ, Diggle PJ. Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. J R Stat Soc Ser -Stat Soc. 2015 Feb;178(2):445–64.

164. Craig MH, Sharp BL, Mabaso ML, Kleinschmidt I. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. Int J Health Geogr. 2007;6(1):1–15.

165. Giardina F, Kasasa S, Sié A, Utzinger J, Tanner M, Vounatsou P. Effects of vector-control interventions on changes in risk of malaria parasitaemia in sub-Saharan Africa: A spatial and temporal analysis. Lancet Glob Health. 2014;2(10):e601–15.

166. Ashton RA, Kefyalew T, Rand A, Sime H, Assefa A, Mekasha A, et al. Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys. Am J Trop Med Hyg. 2015;93(1):168–77.

167. Giardina F, Franke J, Vounatsou P. Geostatistical modelling of the malaria risk in Mozambique: Effect of the spatial resolution when using remotely-sensed imagery. Geospatial Health. 2015;10(2):232–8.

168. Aimone AM, Brown P, Owusu-Agyei S, Zlotkin SH, Cole DC. Impact of iron fortification on the geospatial patterns of malaria and non-malaria infection risk among young children: A secondary spatial analysis of clinical trial data from Ghana. BMJ Open. 2017;7(5):e013192.

169. Noor AM, Alegana VA, Kamwi RN, Hansford CF, Ntomwa B, Katokele S, et al. Malaria control and the intensity of Plasmodium falciparum transmission in Namibia 1969-1992. PLoS One. 2013/05/15 ed. 2013;8(5):e63350.

170. Reid H, Haque U, Clements AC, Tatem AJ, Vallely A, Ahmed SM, et al. Mapping malaria risk in Bangladesh using Bayesian geostatistical models. Am J Trop Med Hyg. 2010/10/05 ed. 2010 Oct;83(4):861–7.

171. Nguyen M, Howes RE, Lucas TC, Battle KE, Cameron E, Gibson HS, et al. Mapping malaria seasonality in Madagascar using health facility data. BMC Med. 2020;18(1):1–11.

172. Arab A, Jackson MC, Kongoli C. Modelling the effects of weather and climate on malaria distributions in West Africa. Malar J. 2014;13(1):126.

173. Samadoulougou S, Maheu-Giroux M, Kirakoya-Samadoulougou F, De Keukeleire M, Castro MC, Robert A. Multilevel and geo-statistical modeling of malaria risk in children of Burkina Faso. Parasit Vectors. 2014;7(1):350.

174. Elyazar IR, Gething PW, Patil AP, Rogayah H, Kusriastuti R, Wismarini DM, et al. Plasmodium falciparum malaria endemicity in Indonesia in 2010. PLoS One. 2011/07/09 ed. 2011;6(6):e21315.

175. Gething PW, Patil AP, Hay SI. Quantifying aggregated uncertainty in Plasmodium falciparum malaria prevalence and populations at risk via efficient space-time geostatistical joint simulation. PLoS Comput Biol. 2010;6(4):e1000724.

176. Hanks EM, Schliep EM, Hooten MB, Hoeting JA. Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. Environmetrics. 2015;26(4):243–54.

177. Chiaravalloti-Neto F, da Silva RA, Zini N, da Silva GCD, da Silva NS, Parra MCP, et al. Seroprevalence for dengue virus in a hyperendemic area and associated socioeconomic and demographic factors using a cross-sectional design and a geostatistical approach, state of Sao Paulo, Brazil. BMC Infect Dis. 2019 May 20;19(1):441.

178. Kazembe LN, Kleinschmidt I, Holtz TH, Sharp BL. Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. Int J Health Geogr. 2006;5:41.

179. Ayele DG, Zewotir TT, Mwambi HG. Spatial distribution of malaria problem in three regions of Ethiopia. Malar J. 2013;12(1):207.

180. Noor AM, Clements AC, Gething PW, Moloney G, Borle M, Shewchuk T, et al. Spatial prediction of Plasmodium falciparum prevalence in Somalia. Malar J. 2008/08/23 ed. 2008 Aug 21;7:159.

181. Macharia PM, Giorgi E, Noor AM, Waqo E, Kiptui R, Okiro EA, et al. Spatio-temporal analysis of Plasmodium falciparum prevalence to understand the past and chart the future of malaria control in Kenya. Malar J. 2018;17(1):1–13.

182. Kang SY, Battle KE, Gibson HS, Ratsimbasoa A, Randrianarivelojosia M, Ramboarina S, et al. Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess Plasmodium falciparum endemicity trends between 2011 and 2016. BMC Med. 2018;16(1):1–15.

183. Colborn KL, Giorgi E, Monaghan AJ, Gudo E, Candrinho B, Marrufo TJ, et al. Spatio-temporal modelling of weekly malaria incidence in children under 5 for early epidemic detection in Mozambique. Sci Rep. 2018;8(1):1–9.

184. Ssempiira J, Nambuusi B, Kissa J, Agaba B, Makumbi F, Kasasa S, et al. The contribution of malaria control interventions on spatio-temporal changes of parasitaemia risk in Uganda during 2009–2014. Parasit Vectors. 2017;10.

185. Noor AM, Uusiku P, Kamwi RN, Katokele S, Ntomwa B, Alegana VA, et al. The receptive versus current risks of Plasmodium falciparum transmission in Northern Namibia: implications for elimination. BMC Infect Dis. 2013;13(1):1–10.

186. Matthys B, Vounatsou P, Raso G, Tschannen AB, Becket EG, Gosoniu L, et al. Urban farming and malaria risk factors in a medium-sized town in Côte d'Ivoire. Am J Trop Med Hyg. 2006;75(6):1223–31.

187. Kleinschmidt I, Sharp BL, Clarke GP, Curtis B, Fraser C. Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa. Am J Epidemiol. 2001;153(12):1213–21.

188. Giorgi E, Osman AA, Hassan AH, Ali AA, Ibrahim F, Amran JG, et al. Using non-exceedance probabilities of policy-relevant malaria prevalence thresholds to identify areas of low transmission in Somalia. Malar J. 2018;17(1):1–10.

189. Chirombo J, Lowe R, Kazembe L. Using structured additive regression models to estimate risk factors of malaria: Analysis of 2010 Malawi malaria indicator survey data. PLoS ONE. 2014;9(7):e101116.

190. Guerra CA, Kang SY, Citron DT, Hergott DE, Perry M, Smith J, et al. Human mobility patterns and malaria importation on Bioko Island. Nat Commun. 2019;10(1):1–10.

191. Fornace KM, Alexander N, Abidin TR, Brock PM, Chua TH, Vythilingam I, et al. Local human movement patterns and land use impact exposure to zoonotic malaria in Malaysian Borneo. Elife. 2019;8:e47602.

192. Raso G, Schur N, Utzinger J, Koudou BG, Tchicaya ES, Rohner F, et al. Mapping malaria risk among children in Côte d'Ivoire using Bayesian geo-statistical models. Malar J. 2012;11(1):1–11.

193. Gosoniu L, Vounatsou P, Sogoba N, Smith T. Bayesian modelling of geostatistical malaria risk data. Geospat Health. 2008/08/08 ed. 2006 Nov;1(1):127–39.

194. Giardina F, Gosoniu L, Konate L, Diouf MB, Perry R, Gaye O, et al. Estimating the Burden of Malaria in Senegal: Bayesian Zero-Inflated Binomial Geostatistical Modeling of the MIS 2008 Data. PLOS ONE. 2012 Mar 5;7(3):e32625.

195. Janko M, Goel V, Emch M. Extending multilevel spatial models to include spatially varying coefficients. Health Place. 2019 Nov;60.

196. Riedel N, Vounatsou P, Miller JM, Gosoniu L, Chizema-Kawesha E, Mukonka V, et al. Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modelling of the 2006 Zambia national malaria indicator survey (ZMIS). Malar J. 2010;9(1):37.

197. Gosoniu L, Vounatsou P, Sogoba N, Maire N, Smith T. Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. Comput Stat Data Anal. 2009 Jul;53(9):3358–71.

198. Salehi M, Mohammad K, Farahani MM, Zeraati H, Nourijelyani K, Zayeri F. Spatial modeling of malaria incidence rates in Sistan and Baluchistan province, Islamic Republic of Iran. Saudi Med J. 2008;29(12):1791–6.

199. Raso G, Silué KD, Vounatsou P, Singer BH, Yapi A, Tanner M, et al. Spatial risk profiling of Plasmodium falciparum parasitaemia in a high endemicity area in Côte d'Ivoire. Malar J. 2009 Nov 11;8(1):252.

200. Adegboye OA, Leung DHY, Wang YG. Analysis of spatial data with a nested correlation structure. J R Stat Soc Ser C Appl Stat. 2017;67(2):329–54.

201. Sharmin S, Glass K, Viennet E, Harley D. Geostatistical mapping of the seasonal spread of under-reported dengue cases in Bangladesh. PLoS Negl Trop Dis. 2018;12(11):e0006947.

202. Sow A, Nikolay B, Faye O, Cauchemez S, Cano J, Diallo M, et al. Changes in the Transmission Dynamic of Chikungunya Virus in Southeastern Senegal. Viruses. 2020;12(2):10.

203. Ahmad H, Ali A, Fatima SH, Zaidi F, Khisroon M, Rasheed SB, et al. Spatial modeling of Dengue prevalence and kriging prediction of Dengue outbreak in Khyber Pakhtunkhwa (Pakistan) using presence only data. Stoch Environ Res Risk Assess. 2020 Jul 1;34(7):1023–36.

204. Routledge I, Lai S, Battle KE, Ghani AC, Gomez-Rodriguez M, Gustafson KB, et al. Tracking progress towards malaria elimination in China: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. PLOS Comput Biol. 2020 Mar 23;16(3):e1007707.

205. Sedda L, Taylor BM, Eiras AE, Marques JT, Dillon RJ. Using the intrinsic growth rate of the mosquito population improves spatio-temporal dengue risk estimation. Acta Trop. 2020 Aug 1;208:105519.

206. Martino S, Rue H. Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program. Dep Math Sci NTNU Nor. 2009;

207. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. Stat Med. 2010;29(21):2224–34.

208. Chien LC, Yu HL. Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. Environ Int. 2014;73:46–56.

209. Chien L, Lin R, Liao Y, Sy F, Perez A. Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. BMC Infect Dis. 2018;(1471-2334 (Electronic)):T-epublish.

210. Yu HL, Lee CH, Chien LC. A spatiotemporal dengue fever early warning model accounting for nonlinear associations with hydrological factors: a Bayesian maximum entropy approach. Stoch Environ Res Risk Assess. 2016;30(8):2127–41.

211. Lowe R, Gasparrini A, Meerbeeck CJV, Lippi CA, Mahon R, Trotman AR, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. PLOS Med. 2018 Jul 17;15(7):e1002613.

212. Prem K, Lau MSY, Tam CC, Ho MZJ, Ng LC, Cook AR. Inferring who-infected-whom-where in the 2016 Zika outbreak in Singapore-a spatio-temporal model. J R Soc Interface. 2019 Jun;16(155).

213. Cauchemez S, Ledrans M, Poletto C, Quénel P, De Valk H, Colizza V, et al. Local and regional spread of chikungunya fever in the Americas. Eurosurveillance. 2014;19(28):20854.

214. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Vol. 1. Springer series in statistics New York; 2001.

215. Akhtar M, Kraemer MUG, Gardner LM. A dynamic neural network model for predicting risk of Zika in real time. Bmc Med. 2019 Sep;17(1).

216. Shi B, Liu J, Zhou X-N, Yang G-J. Inferring Plasmodium vivax Transmission Networks from Tempo-Spatial Surveillance Data. PLoS Negl Trop Dis. 2014 Feb 6;8(2):e2682.

217. Bomfim R, Pei S, Shaman J, Yamana T, Makse HA, Andrade JS, et al. Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. J R Soc Interface. 2020 Oct 28;17(171):20200691.

218. Rossi G, Karki S, Smith RL, Brown WM, Ruiz MO. The spread of mosquito-borne viruses in modern times: A spatio-temporal analysis of dengue and chikungunya. Spat Spatio-Temporal Epidemiol. 2018;26:113–25.

219. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13.

220. Haddawy P, Hasan AHMI, Kasantikul R, Lawpoolsri S, Sa-angchai P, Kaewkungwal J, et al. Spatiotemporal Bayesian networks for malaria prediction. Artif Intell Med. 2018;84:127–38.

221. Marini G, Guzzetta G, Marques Toledo CA, Teixeira M, Rosa R, Merler S. Effectiveness of Ultra-Low Volume insecticide spraying to prevent dengue in a non-endemic metropolitan area of Brazil. PLoS Comput Biol. 2019;15(3):e1006831.

222. Guzzetta G, Marques-Toledo CA, Rosà R, Teixeira M, Merler S. Quantifying the spatial spread of dengue in a non-endemic Brazilian metropolis via transmission chain reconstruction. Nat Commun. 2018;9(1):2837.

223. Miyaoka TY, Lenhart S, Meyer JFCA. Optimal control of vaccination in a vector-borne reaction–diffusion model applied to Zika virus. J Math Biol. 2019;79(3):1077–104.

224. Li R, Xu L, Bjørnstad ON, Liu K, Song T, Chen A, et al. Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. Proc Natl Acad Sci U S A. 2019;116(9):3624–9.

225. Yu HL, Angulo JM, Cheng MH, Wu J, Christakos G. An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan). Biom J. 2014;56(3):428–40.

226. Zhu G, Liu T, Xiao J, Zhang B, Song T, Zhang Y, et al. Effects of human mobility, temperature and mosquito control on the spatiotemporal transmission of dengue. Sci Total Environ. 2019;651:969–78.

227. Silal SP, Little F, Barnes KI, White LJ. Hitting a moving target: A model for malaria elimination in the presence of population movement. PLoS ONE. 2015;10(12):e0144990.

228. Senapati A, Sardar T, Ganguly KS, Ganguly KS, Chattopadhyay AK, Chattopadhyay J. Impact of adult mosquito control on dengue prevalence in a multi-patch setting: A case study in Kolkata (2014–2015). J Theor Biol. 2019;478:139–52.

229. Xue L, Scott HM, Cohnstaedt LW, Scoglio C. A network-based meta-population approach to model Rift Valley fever epidemics. J Theor Biol. 2012;306:129–44.

230. Moulay D, Pigné Y. A metapopulation model for chikungunya including populations mobility on a large-scale network. J Theor Biol. 2013;318:129–39.

231. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci. 2015 Sep 22;112(38):11887–92.

232. Kraemer MUG, Bisanzio D, Reiner RC, Zakar R, Hawkins JB, Freifeld CC, et al. Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan. Epj Data Sci. 2018 Jun;7:17.

233. O'Reilly KM, Lowe R, Edmunds WJ, Mayaud P, Kucharski A, Eggo RM, et al. Projecting the end of the Zika virus epidemic in Latin America: A modelling analysis. BMC Med. 2018;16(1):180.

234. Sun K, Zhang Q, Pastore-Piontti A, Chinazzi M, Mistry D, Dean NE, et al. Quantifying the risk of local Zika virus transmission in the contiguous US during the 2015-2016 ZIKV epidemic. BMC Med. 2018;16(1):195.

235. Zhang Q, Sun KY, Chinazzi M, Piontti APY, Dean NE, Rojas DP, et al. Spread of Zika virus in the Americas. Proc Natl Acad Sci U S A. 2017 May;114(22):E4334–43.

236. Ruktanonchai NW, DeLeenheer P, Tatem AJ, Alegana VA, Caughlin TT, Erbach-Schoenberg E zu, et al. Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data. PLOS Comput Biol. 2016;12(4):e1004846.

237. Barrios E, Lee S, Vasilieva O. Assessing the effects of daily commuting in two-patch dengue dynamics: A case study of Cali, Colombia. J Theor Biol. 2018;453:14–39.

238. Stolerman LM, Coombs D, Boatto S. Sir-network model and its application to dengue fever. SIAM J Appl Math. 2015;75(6):2581–609.

239. Seroussi I, Levy N, Yom-Tov E. Multi-season analysis reveals the spatial structure of disease spread. Phys Stat Mech Its Appl. 2020 Jun 1;547:124425.

240. Fitzgibbon WE, Morgan JJ, Webb GF. An outbreak vector-host epidemic model with spatial structure: the 2015–2016 Zika outbreak in Rio De Janeiro. Theor Biol Med Model. 2017 Mar 27;14(1):7.

241. Mukhtar AYA, Munyakazi JB, Ouifki R. Assessing the role of human mobility on malaria transmission. Math Biosci. 2020;320(108304).

242. Núñez-López M, Alarcón Ramos L, Velasco-Hernández JX. Migration rate estimation in an epidemic network. Appl Math Model. 2021 Jan 1;89:1949–64.

243. Zhu G, Liu J, Tan Q, Shi B. Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China. PLoS Negl Trop Dis. 2016;10(4):e0004633.

244. Moore SM, ten Bosch QA, Siraj AS, Soda KJ, España G, Campo A, et al. Local and regional dynamics of chikungunya virus transmission in Colombia: The role of mismatched spatial heterogeneity. BMC Med. 2018;16(1).

245. Zhu G, Xiao J, Zhang B, Liu T, Lin H, Li X, et al. The spatiotemporal transmission of dengue and its driving mechanism: A case study on the 2014 dengue outbreak in Guangdong, China. Sci Total Environ. 2018;622–623:252–9.

246. Chadsuthi S, Althouse BM, Iamsirithaworn S, Triampo W, Grantz KH, Cummings DAT. Travel distance and human movement predict paths of emergence and spatial spread of chikungunya in Thailand. Epidemiol Infect. 2018;146(13):1654–62.

247. Kim M, Paini D, Jurdak R. Modeling stochastic processes in disease spread across a heterogeneous social system. Proc Natl Acad Sci U S A. 2019;116(2):401–6.

248. Pizzitutti F, Pan W, Barbieri A, Miranda JJ, Feingold B, Guedes GR, et al. A validated agent-based model to study the spatial and temporal heterogeneities of malaria incidence in the rainforest environment. Malar J. 2015;14(1):1030.

249. Karl S, Halder N, Kelso JK, Ritchie SA, Milne GJ. A spatial simulation model for dengue virus infection in urban areas. BMC Infect Dis. 2014;14(1):1–17.

250. Massaro E, Kondor D, Ratti C. Assessing the interplay between human mobility and mosquito borne diseases in urban environments. Sci Rep. 2019;9(1):1–13.

251. Gerardin J, Bever CA, Hamainza B, Miller JM, Eckhoff PA, Wenger EA. Optimal Population-Level Infection Detection Strategies for Malaria Control and Elimination in a Spatial Model of Malaria Transmission. PLoS Comput Biol. 2016;12(1):e1004707.

252. Gardner LM, Bota A, Gangavarapu K, Kraemer MUG, Grubaugh ND. Inferring the risk factors behind the geographical spread and transmission of Zika in the Americas. PLoS Negl Trop Dis. 2018 Jan;12(1):25.

C.2 Supplementary figures

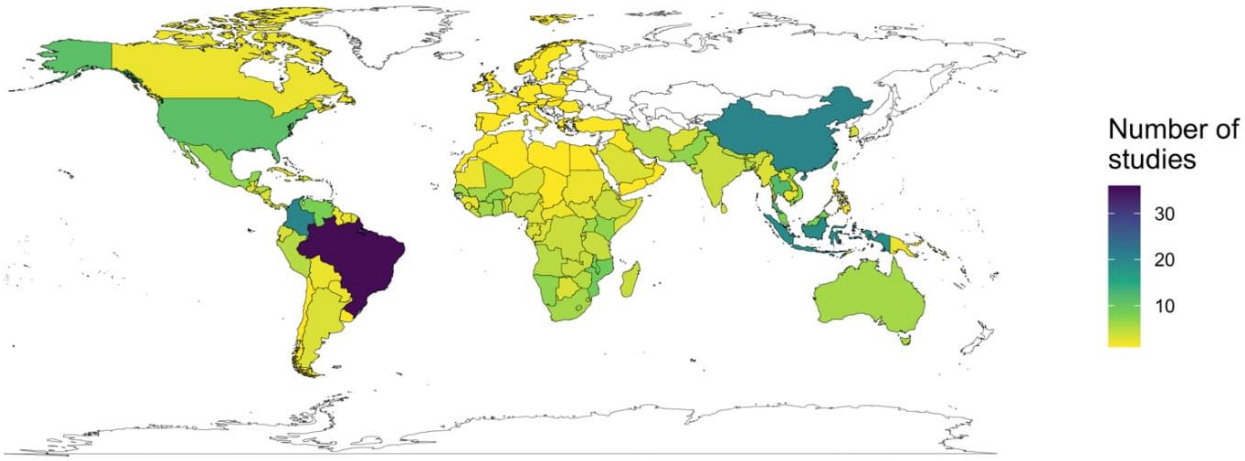


Figure S1. Spatial distribution of studies included in this review. Number of studies modelling mosquito-borne diseases by country. Note that studies included data from more than one country or an entire region (e.g. Latin America or EU/EEA).

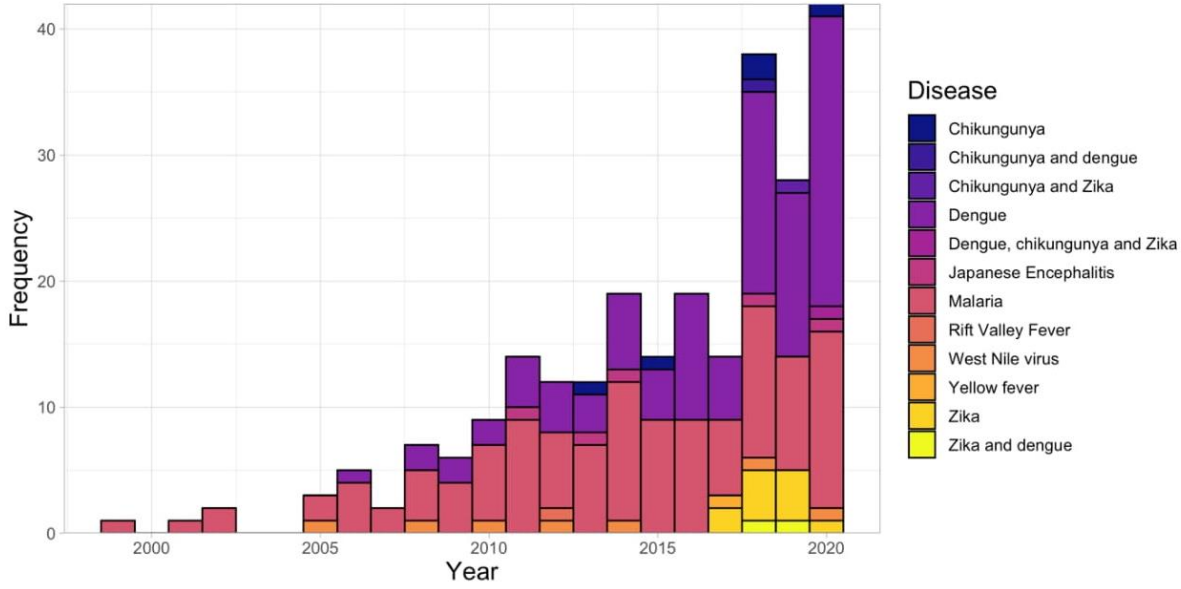


Figure S2: Number of spatial modelling studies published per year by mosquito-borne disease.

	Chikungunya	Chikungunya and dengue	Chikungunya and Zika	Dengue	Dengue, chikungunya and Zika	Japanese Encephalitis	Malaria	Rift Valley Fever	West Nile fever	Yellow fever	Zika	Zika and dengue	Total
Total	5 (100%)	1 (100%)	1 (100%)	95 (100%)	1 (100%)	5 (100%)	118 (100%)	1 (100%)	7 (100%)	1 (100%)	11 (100%)	2 (100%)	248 (100%)
World region													
African Region	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	72 (61%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	75 (30.2%)
Eastern Mediterranean Region	0 (0%)	0 (0%)	0 (0%)	5 (5.3%)	0 (0%)	0 (0%)	5 (4.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	10 (4%)
European Region	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Multiple regions	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (4.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (2.4%)
Region of the Americas	2 (40%)	0 (0%)	1 (100%)	33 (34.7%)	1 (100%)	0 (0%)	11 (9.3%)	0 (0%)	7 (100%)	0 (0%)	10 (90.9%)	2 (100%)	67 (27%)
South-East Asia Region	1 (20%)	0 (0%)	0 (0%)	27 (28.4%)	0 (0%)	2 (40%)	11 (9.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	41 (16.5%)
Western Pacific Region	1 (20%)	0 (0%)	0 (0%)	29 (30.5%)	0 (0%)	3 (60%)	14 (11.9%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	48 (19.4%)
Spatial scale													
Country	0 (0%)	1 (100%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (1.2%)
Country and district	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	2 (18.2%)	0 (0%)	4 (1.6%)
District	3 (60%)	0 (0%)	1 (100%)	71 (74.7%)	1 (100%)	5 (100%)	65 (55.1%)	1 (100%)	7 (100%)	1 (100%)	5 (45.5%)	2 (100%)	162 (65.3%)
Cluster	0 (0%)	0 (0%)	0 (0%)	4 (4.2%)	0 (0%)	0 (0%)	21 (17.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	25 (10.1%)
Patch	1 (20%)	0 (0%)	0 (0%)	9 (9.5%)	0 (0%)	0 (0%)	11 (9.3%)	0 (0%)	0 (0%)	0 (0%)	3 (27.3%)	0 (0%)	24 (9.7%)
Individual	0 (0%)	0 (0%)	0 (0%)	10 (10.5%)	0 (0%)	0 (0%)	19 (16.1%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	30 (12.1%)

	Chikungunya	Chikungunya and dengue	Chikungunya and Zika	Dengue	Dengue, chikungunya and Zika	Japanese Encephalitis	Malaria	Rift Valley Fever	West Nile fever	Yellow fever	Zika	Zika and dengue	Total
Total	5 (100%)	1 (100%)	1 (100%)	95 (100%)	1 (100%)	5 (100%)	118 (100%)	1 (100%)	7 (100%)	1 (100%)	11 (100%)	2 (100%)	248 (100%)
Model class													
Machine learning	0 (0%)	1 (100%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	2 (1.7%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	5 (2%)
Machine learning and mechanistic	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Mechanistic	3 (60%)	0 (0%)	0 (0%)	17 (17.9%)	0 (0%)	0 (0%)	5 (4.2%)	1 (100%)	1 (14.3%)	0 (0%)	6 (54.5%)	0 (0%)	33 (13.3%)
Statistical	2 (40%)	0 (0%)	1 (100%)	75 (78.9%)	1 (100%)	5 (100%)	111 (94.1%)	0 (0%)	6 (85.7%)	1 (100%)	4 (36.4%)	2 (100%)	208 (83.9%)
Statistical and machine learning	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Model type													
Fixed effect	0 (0%)	0 (0%)	0 (0%)	27 (28.4%)	0 (0%)	3 (60%)	17 (14.4%)	0 (0%)	2 (28.6%)	1 (100%)	1 (9.1%)	0 (0%)	51 (20.6%)
Mixed effect	1 (20%)	0 (0%)	1 (100%)	48 (50.5%)	1 (100%)	2 (40%)	94 (79.7%)	0 (0%)	4 (57.1%)	0 (0%)	2 (18.2%)	2 (100%)	155 (62.5%)
Machine learning	0 (0%)	1 (100%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	2 (1.7%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	5 (2%)
Compartmental	1 (20%)	0 (0%)	0 (0%)	15 (15.8%)	0 (0%)	0 (0%)	4 (3.4%)	1 (100%)	1 (14.3%)	0 (0%)	6 (54.5%)	0 (0%)	28 (11.3%)
Mixed	0 (0%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.8%)
Other	3 (60%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	7 (2.8%)

	Chikungunya	Chikungunya and dengue	Chikungunya and Zika	Dengue	Dengue, chikungunya and Zika	Japanese Encephalitis	Malaria	Rift Valley Fever	West Nile fever	Yellow fever	Zika	Zika and dengue	Total
Total	5 (100%)	1 (100%)	1 (100%)	95 (100%)	1 (100%)	5 (100%)	118 (100%)	1 (100%)	7 (100%)	1 (100%)	11 (100%)	2 (100%)	248 (100%)
Spatial model													
GLM	0 (0%)	0 (0%)	0 (0%)	8 (8.4%)	0 (0%)	2 (40%)	2 (1.7%)	0 (0%)	1 (14.3%)	1 (100%)	0 (0%)	0 (0%)	14 (5.6%)
GWR	0 (0%)	0 (0%)	0 (0%)	9 (9.5%)	0 (0%)	1 (20%)	9 (7.6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	19 (7.7%)
GAM	0 (0%)	0 (0%)	0 (0%)	9 (9.5%)	0 (0%)	0 (0%)	6 (5.1%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	16 (6.5%)
Autoregressive distributed lag model	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
GLMM	1 (20%)	0 (0%)	1 (100%)	44 (46.3%)	1 (100%)	2 (40%)	80 (67.8%)	0 (0%)	3 (42.9%)	0 (0%)	1 (9.1%)	2 (100%)	135 (54.4%)
GAMM	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	11 (9.3%)	0 (0%)	1 (14.3%)	0 (0%)	0 (0%)	0 (0%)	13 (5.2%)
DLNM	0 (0%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	3 (1.2%)
GLMM and GLM	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (14.3%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
GLMM and GWR	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
GLMM, GLM and GAM	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
GLMM and GAMM	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Neural network	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	3 (1.2%)
Boosted regression trees	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Bayesian network	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Compartmental	0 (0%)	0 (0%)	0 (0%)	4 (4.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	5 (2%)
Metapopulation	3 (60%)	0 (0%)	0 (0%)	11 (11.6%)	0 (0%)	0 (0%)	3 (2.5%)	1 (100%)	1 (14.3%)	0 (0%)	4 (36.4%)	0 (0%)	23 (9.3%)

Agent-based model	0 (0%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)	2 (1.7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (1.6%)
GIIM	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	1 (0.4%)
Bespoke	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	2 (0.8%)
Neural network and metapopulation	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Random forest and GLM	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Spatial assumption													
Distance-based	1 (20%)	0 (0%)	1 (100%)	71 (74.7%)	1 (100%)	4 (80%)	103 (87.3%)	0 (0%)	6 (85.7%)	0 (0%)	3 (27.3%)	2 (100%)	192 (77.4%)
Human movement	4 (80%)	0 (0%)	0 (0%)	21 (22.1%)	0 (0%)	0 (0%)	9 (7.6%)	0 (0%)	1 (14.3%)	1 (100%)	5 (45.5%)	0 (0%)	41 (16.5%)
Vector movement	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Human and vector movement	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	2 (18.2%)	0 (0%)	4 (1.6%)
Mixed	0 (0%)	1 (100%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	5 (4.2%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	8 (3.2%)
Not given	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.8%)

	Chikungunya	Chikungunya and dengue	Chikungunya and Zika	Dengue	Dengue, chikungunya and Zika	Japanese Encephalitis	Malaria	Rift Valley Fever	West Nile fever	Yellow fever	Zika	Zika and dengue	Total
Total	5 (100%)	1 (100%)	1 (100%)	95 (100%)	1 (100%)	5 (100%)	118 (100%)	1 (100%)	7 (100%)	1 (100%)	11 (100%)	2 (100%)	248 (100%)
Spatial data													
Adjacency	0 (0%)	0 (0%)	1 (100%)	49 (51.6%)	1 (100%)	3 (60%)	37 (31.4%)	0 (0%)	5 (71.4%)	0 (0%)	2 (18.2%)	2 (100%)	100 (40.3%)
Distance	0 (0%)	0 (0%)	0 (0%)	16 (16.8%)	0 (0%)	1 (20%)	60 (50.8%)	0 (0%)	1 (14.3%)	0 (0%)	0 (0%)	0 (0%)	78 (31.5%)
Distance and direction	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
Coordinates	0 (0%)	0 (0%)	0 (0%)	9 (9.5%)	0 (0%)	1 (20%)	6 (5.1%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)	0 (0%)	17 (6.9%)
Mobility data	0 (0%)	0 (0%)	0 (0%)	5 (5.3%)	0 (0%)	0 (0%)	3 (2.5%)	0 (0%)	1 (14.3%)	0 (0%)	2 (18.2%)	0 (0%)	11 (4.4%)
Mathematical model	1 (20%)	0 (0%)	0 (0%)	4 (4.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (2%)
Other covariates	0 (0%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.8%)
Mixed	3 (60%)	1 (100%)	0 (0%)	8 (8.4%)	0 (0%)	0 (0%)	5 (4.2%)	1 (100%)	0 (0%)	1 (100%)	4 (36.4%)	0 (0%)	23 (9.3%)
No data	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (18.2%)	0 (0%)	3 (1.2%)
Not given	1 (20%)	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	6 (5.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	8 (3.2%)

Appendix D: Supplementary Material Chapter 4

D.1 Supplementary material

Supplementary material to support Chapter 4: *A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping*. Contains additional simulation studies and a sensitivity analysis. Taken from https://doi.org/10.1098/rsif.2022.0440.

D.1.1 Comparison of spatial smooth and random effect models: a single source of distance-based connectivity

In this section, we provide a comparison between the proposed penalised regression spline modelling approach and the more conventional neighbourhood-based conditional autoregressive (CAR) model. Full details of data generation and model structure are provided in section 3 of the main text. We compared the spatial smooth approach to the BYM2 random effect models fitted using integrated nested Laplace approximations (INLA) in the main text. However, as the spatial smooth model was fitted using Markov chain Monte Carlo (MCMC) methods, we refitted the BYM2 random effects model using MCMC via NIMBLE [1]. We compared this to our spatial smooth approach to ensure that any differences between our new approach and the conventional BYM2 random effects approach were not caused by differences in inferential methods.

D1.1.1 Modelling approach

Briefly, we applied two Poisson models to fictional count data with an intercept of 0 and a known underlying spatial structure:

$$y_i \sim Poisson\big(E(y_i)\big)$$

$$log\big(E(y_i)\big) = log(\xi_i) + \alpha + u_i + v_i \quad (1)$$

$$log\big(E(y_i)\big) = log(\xi_i) + \alpha + \frac{1}{\tau}\big(\sqrt{\phi}u_{*i} + \sqrt{1-\phi}v_{*i}\big) \quad (2)$$

In model (1), u_i is a spatially structured term, formulated using a thin plate regression spline on the coordinates of the centroid of each municipality, and v_i is a spatially unstructured term, assumed to follow a zero-mean normal distribution, representing heterogeneity between regions. Model (2) represents the BYM2 model, a scaled version of the BYM approach that includes a spatially structured random effect u_{*i}, assuming a CAR neighbourhood-structure, and an unstructured Normal random effect v_{*i} [2,3]. Model (2) contains a mixing parameter ϕ which measures the contribution of each random effect to the marginal variance $\frac{1}{\tau^2}$ [2,4]. Here, $\phi = 0$ indicates no spatial structure within the data and $\phi = 1$ represents a purely spatial model. Both models were fitted using MCMC simulations via the NIMBLE package [1], model (2) was also fitted using the R-INLA package [5].

Model comparison was based on mean absolute error (MAE) and WAIC [6]. The relative contribution of each spatial term in model (1) to the overall random structure was estimated using the proportion of the overall random term variance explained by u ($var(u)/var(u+v)$) for the spatial smooth model based on simulations from the posterior distribution of u and v. This was compared to estimates of the ϕ parameter from model (2) and the known mixing parameter from data generation. All analyses were carried out using R version 4.1.1 [7]. Code used to simulate data and perform analyses is available here: https://github.com/sophie-a-lee/spatial_smooth_framework.

D1.1.2 Results

All three models were able to accurately estimate the true intercept coefficient value of 0, with all models capturing 0 within the 95% credible interval (Figure D1).

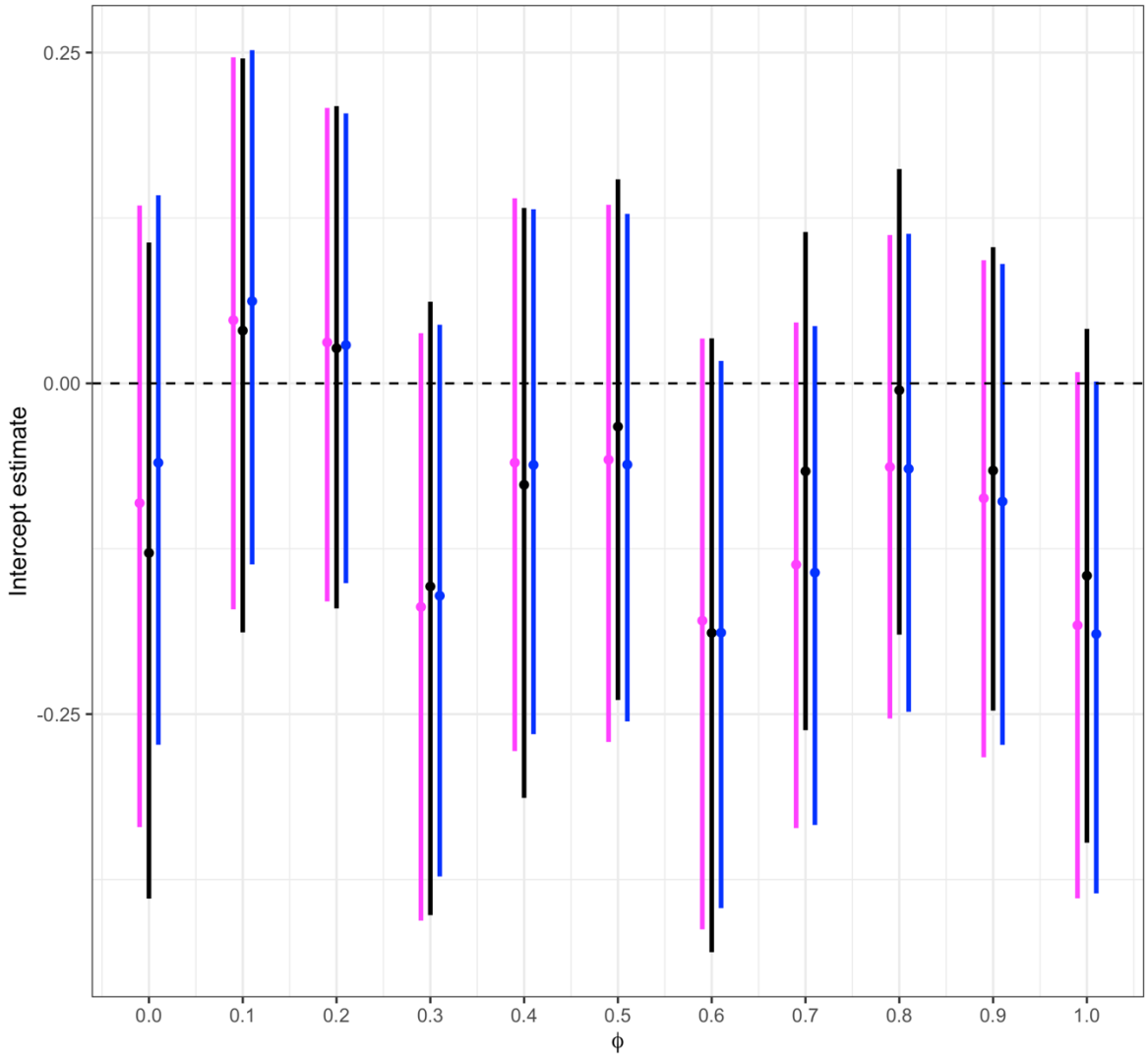


Figure D1: Mean and 95% credible interval of the intercept coefficient estimates from the smoothing spline (black) and BYM2 models fitted using INLA (blue) and MCMC simulations (pink), compared to the true simulated value, 0.

Figure D2 shows that the BYM2 random effects model produced comparable estimates of the mixing parameter when it was fitted via INLA or MCMC approaches. The spatial smooth model provided estimates that were closer to the true value of ϕ than the BYM2 models for most simulations, although all models could detect changes to the relative contribution of the spatial structure within the data.

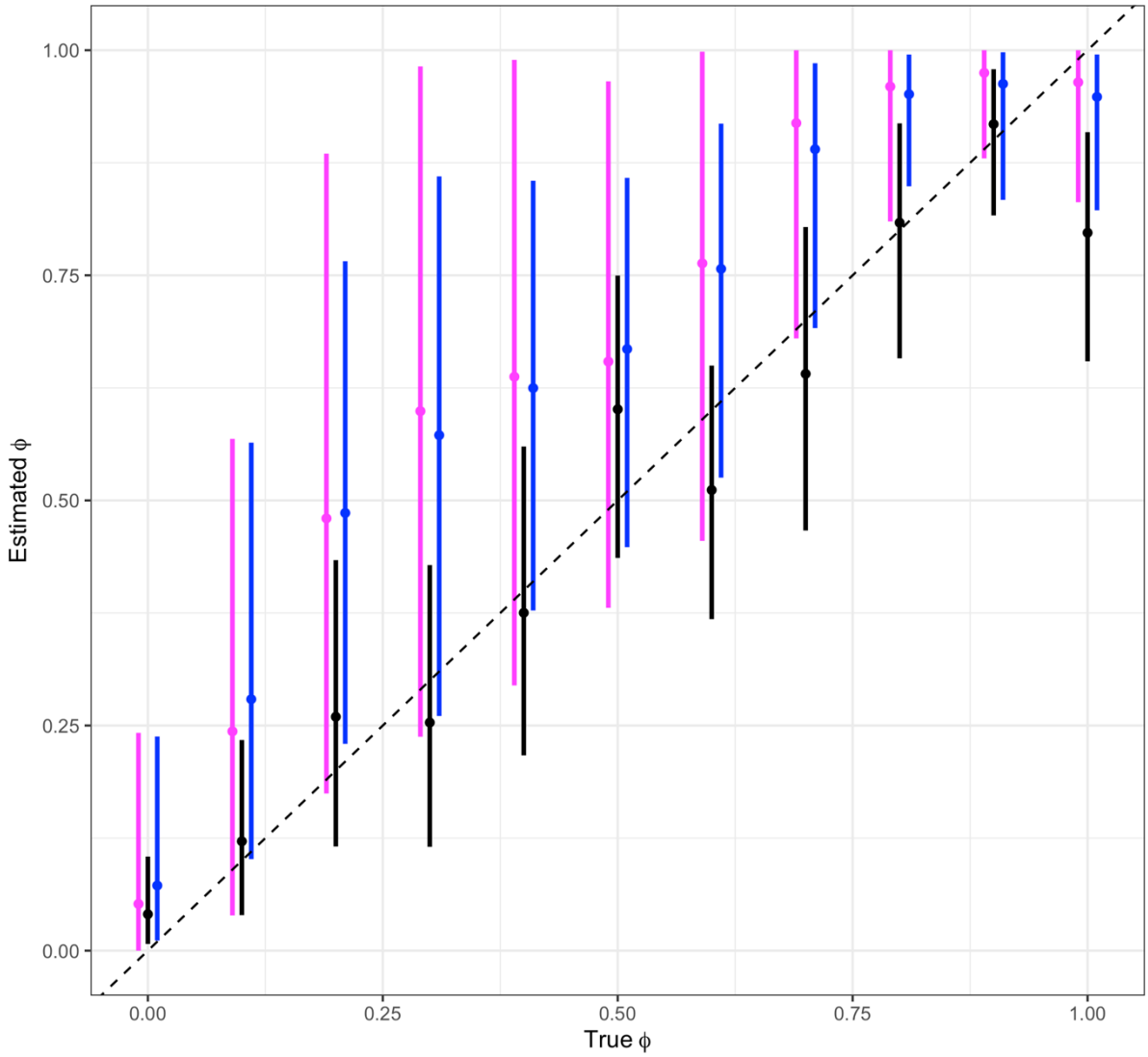


Figure D2: The mean and 95% credible interval of estimated ϕ values extracted from models including a smoothing spline (black) and BYM2 models fitted using INLA (blue) and MCMC simulations (pink), compared to the known value (dashed line).

Mean absolute errors and WAIC values were similar between all 3 models (Table D1). In particular, the results from the BYM2 model fitted using MCMC are almost identical to the spatial smooth model showing that this approach performs as well as the standard approach (BYM2). We noticed some systematic discrepancies between the INLA model goodness of fit statistics and those extracted from NIMBLE. However, as these differences were not apparent between the models fitted using MCMC, they appear to be a result of using different software packages rather than the differences in model formulations. The objective of these comparisons was to show that the spatial smoothing model performs as well as the current standard (the

BYM2 approach) which, given the very small differences between the goodness of fit statistics, is supported by this analysis.

Table D1: Model comparison statistics and mean estimates of the mixing parameter, ϕ, from the smoothing spline and BYM2 models fitted using INLA and MCMC methods. Mean absolute error (MAE), the mean absolute difference between observed and predicted outcomes, and the Watanabe-Akaike information criterion (WAIC), the negative of the average log pointwise predictive density, calculated for the spatial spline and BYM2 models for each simulated dataset. The lowest MAE and WAIC, and the ϕ estimate closest to the value used in each simulation are highlighted in bold.

ϕ	Smoothing spline model			INLA BYM2 model			NIMBLE BYM2 model		
	MAE	WAIC	ϕ estimate	MAE	WAIC	ϕ estimate	MAE	WAIC	ϕ estimate
0	1.51	**996.94**	**0.041**	**1.04**	1005.79	0.072	1.51	997.96	0.052
0.1	1.54	**1030.64**	**0.121**	**1.11**	1034.29	0.279	1.54	1032.92	0.244
0.2	1.33	932.42	**0.260**	**1.04**	**931.79**	0.486	1.37	935.52	0.480
0.3	1.27	**909.42**	**0.253**	**0.93**	912.50	0.572	1.30	912.79	0.599
0.4	1.39	**961.67**	**0.375**	**1.08**	976.12	0.625	1.46	974.02	0.637
0.5	1.54	**935.09**	**0.601**	**1.21**	954.34	0.668	1.58	953.01	0.654
0.6	1.50	**881.09**	**0.512**	**1.13**	973.61	0.757	1.55	900.05	0.763
0.7	1.45	**931.85**	**0.641**	**1.17**	989.24	0.890	1.50	935.66	0.919
0.8	1.63	947.51	**0.808**	**1.37**	983.96	0.951	1.70	**944.76**	0.960
0.9	1.59	876.37	**0.918**	**1.37**	922.29	0.963	1.69	**876.18**	0.975
1	1.48	875.42	0.797	**1.25**	924.14	0.948	1.57	**872.03**	**0.964**

D1.2 Human movement coordinates

To create a smooth surface describing the spatial structure of data under a given connectivity assumption, we apply penalised smoothing splines to coordinates describing this relative connectivity. For example, when describing distance-based connectivity, we apply smoothing splines to the coordinates of observations. However, when a coordinate system does not currently exist that describes the relative connectivity between observations, we must create one. Multidimensional scaling (MDS) is a mathematical approach that translates continuous measures of distance (in this case, connectivity) onto an abstract cartesian space and returns a set of coordinates [8]. When considering human movement-based connectivity, we can apply MDS to a continuous measure of human movement, i.e., the number of people moving between areas, and use the resulting coordinates to construct the smooth spatial surface. Note that MDS requires the continuous measure of connectivity to be symmetric. In the case of human

movement, this means the number of people moving in one direction is assumed to be equal to the number travelling in the opposite direction.

Continuous measures of human movement-based connectivity can include observed data, such as the number of air travel passengers, or can be estimated using movement models, such as gravity and radiation models [9,10], that assume the number of people moving between areas is a function of population and distance. To imitate these movement models, we used smoothing splines to investigate the relationship between 'connectedness', distance and population. In Brazil, the Regiões de Influência das Cidades [Regions of Influence of Cities] (REGIC) study aims to recreate the Brazilian urban network which explains the movement of people, goods, and services around the country [11]. Based on a survey of residents, the study produced a binary matrix classifying all cities within Brazil as either 'connected' or not in 2018 (Figure D3). We used a logistic generalised additive model to estimate the relationship between the binary measure of connectedness, and the Euclidean distance and population in each city:

$$c_{i,j} \sim Bernoulli(p_{i,j})$$

$$logit(p_{i,j}) = \alpha + f(d_{i,j}, r_i, r_j)$$

Where $p_{i,j}$ is the probability that cities *i* and *j* are connected, $c_{i,j}$ was the binary connectedness indicator taken from the REGIC study, $f(d_{i,j}, r_i, r_j)$ is a 3-dimensional tensor smoothing spline applied to the Euclidean distance between cities ($d_{i,j}$) and the populations in city *i* (r_i) and city *j* (r_j). Tensor smoothing splines allow interactions between covariates measured on different scales (in this example, population and distance) [12]. The model was implemented using the mgcv package [13] using restricted maximum likelihood (REML), allowing for Bayesian interpretation of the results [14,15]. The predicted probability that cities are connected ($p_{i,j}$) was extracted from the model by taking simulations from the posterior distribution. MDS was applied to the predicted probabilities to produce a coordinate system describing the relative connectivity of municipalities based on human movement.

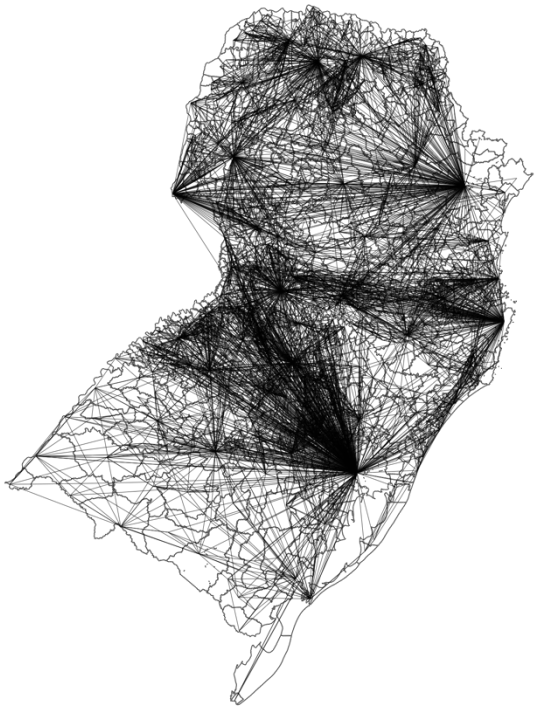


Figure D3: Connections between municipalities in South Brazil extracted from the REGIC 2018 study [11].

D1.3 Simulation study: a single source of human movement-based spatial structure

In this section, we present a simulation study with a single source of spatial connectivity in the data, arising due to human movement.

D.1.3.1 Data generation

Fictitious disease data was generated from a Poisson distribution for each of the 1,013 municipalities in South Brazil:

$$y_i \sim Poisson\big(E(y_i)\big)$$

$$log(E(y_i)) = log(\xi_i) + \alpha + S_i \quad (3)$$

Where y_i is the number of cases in municipality i,, $E(y_i)$ is the expected count, and ξ_i is an offset term set to the population divided by 100,000 so that y_i/ξ_i is the incidence rate per

100,000 residents for municipality i. α = 0 is the intercept, or baseline risk, and S_i is the underlying spatial structure of the data:

$$S_i = \sqrt{\phi} \cdot sm(x_i, z_i) + \sqrt{(1-\phi)} \cdot \varepsilon_i \quad (4)$$

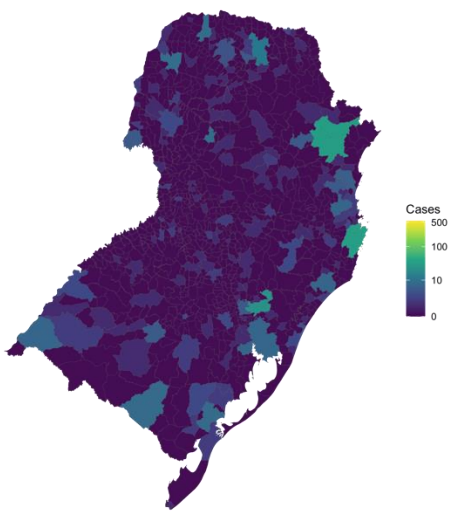
Where ϕ is a mixing parameter that determines the relative contribution of a spatially structured surface, $sm(x_i, z_i)$, and an unstructured random term, $\varepsilon_i \sim N(0,1)$. $sm(x_i, z_i)$ is a smooth function applied to human movement-based connectivity coordinates, defined above, which emulates a spatially structured surface (taken from [12]):

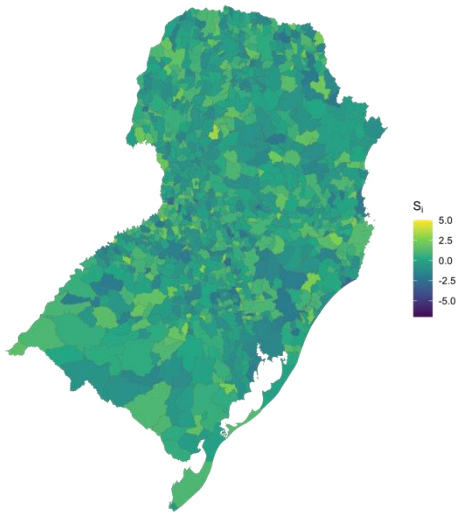
$$sm(x,z) = \pi\sigma_x\sigma_z\left(1.2e^{-(x-0.2)^2/\sigma_x^2-(z-0.3)^2/\sigma_z^2} + 0.8e^{\frac{-(x-0.7)^2}{\sigma_x^2}-\frac{(z-0.8)^2}{\sigma_z^2}}\right) \quad (5)$$

$$\sigma_x = 0.3, \ \sigma_z = 0.4$$

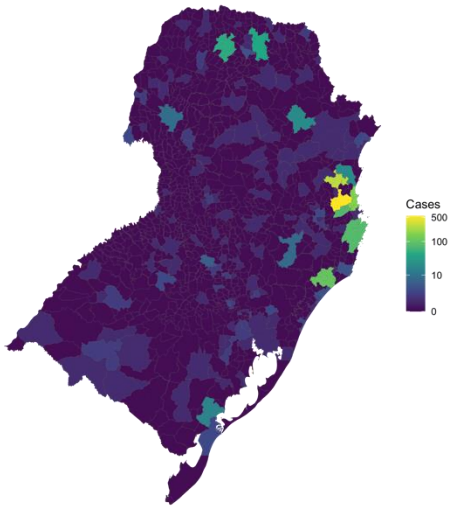
The smooth function was centred around 0 by subtracting the overall mean from each value. 11 simulated datasets were produced using equation (3), setting values of ϕ between 0 and 1 at intervals of 0.1 (Figure D4).

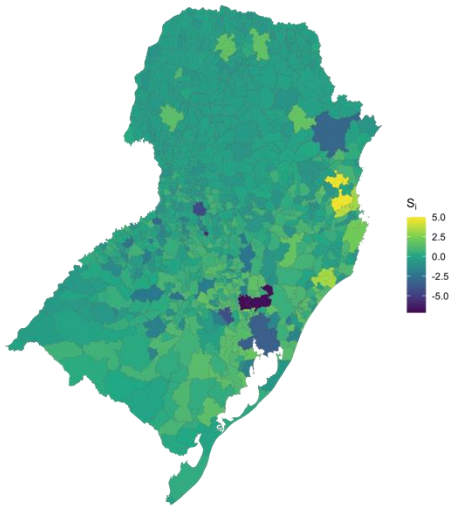
a)


Cases
500
100
10
0


S_i
5.0
2.5
0.0
-2.5
-5.0

b)


Cases
500
100
10
0


S_i
5.0
2.5
0.0
-2.5
-5.0

c)

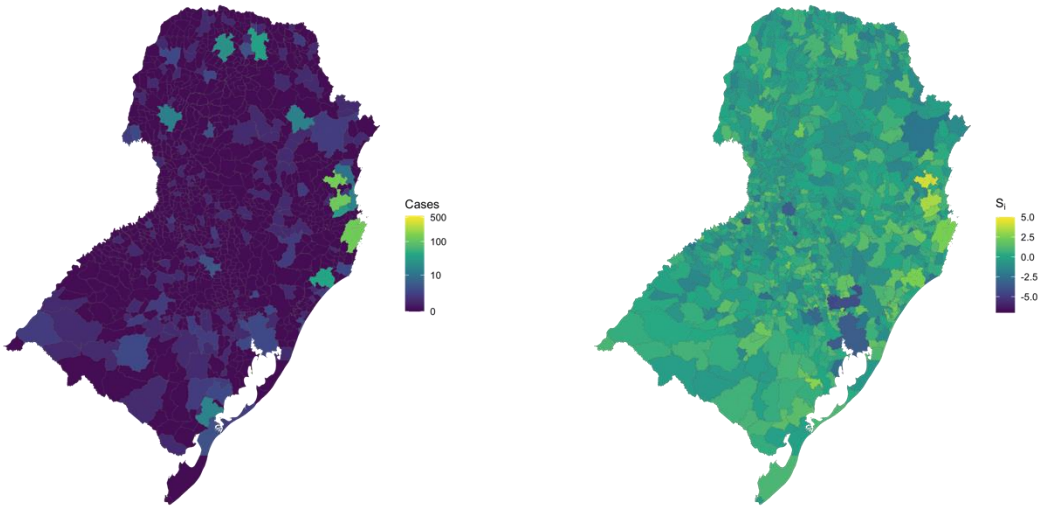


Figure D4: Simulated disease counts (left) and spatial random effects (right) under a human movement-based structure using different spatial structure combinations. The number of cases simulated from a Poisson model and the underlying spatial structure where the data has a) no spatial structure ($\phi = 0$), b) a human movement-based structure only ($\phi = 1$), and c) equal contribution of both structures ($\phi = 0.5$). Note that the number of cases is shown on the log scale.

D.1.3.2 Modelling approach

A spatial Poisson model containing spatially structured and unstructured random effects was applied to each simulated dataset:

$$y_i \sim Poisson\big(E(y_i)\big)$$

$$log\big(E(y_i)\big) = log(\xi_i) + \alpha + u_i + v_i \quad (6)$$

u_i is a spatially structured random term, created by applying a thin plate regression spline to human movement-based coordinates generated using MDS (see section S2). v_i is a spatially unstructured term, expected to follow a zero-mean normal distribution, representing heterogeneity between regions. The smooth surface used to structure u_i was generated using the mgcv package and extracted using the jagam function [13,16]. Models were implemented using MCMC via NIMBLE [1,17]. The relative contribution of each random term to the overall

marginal variance was calculated from simulations and compared to the known value of ϕ used to generate the data (see equation (4)).

D.1.3.3 Results

The 95% credible interval of the estimated intercept contained the true value of zero for all but one of the simulations (Figure D5). The proportion of the random effect variance explained by the structured term provided an accurate estimate of the known contribution from each simulation (Figure D6), showing that alternative spatial structures can be considered within the proposed modelling framework. Multidimensional scaling allows coordinate systems to be derived from any continuous measure of connectivity, allowing assumptions of connectivity beyond distance-based to be included and tested.

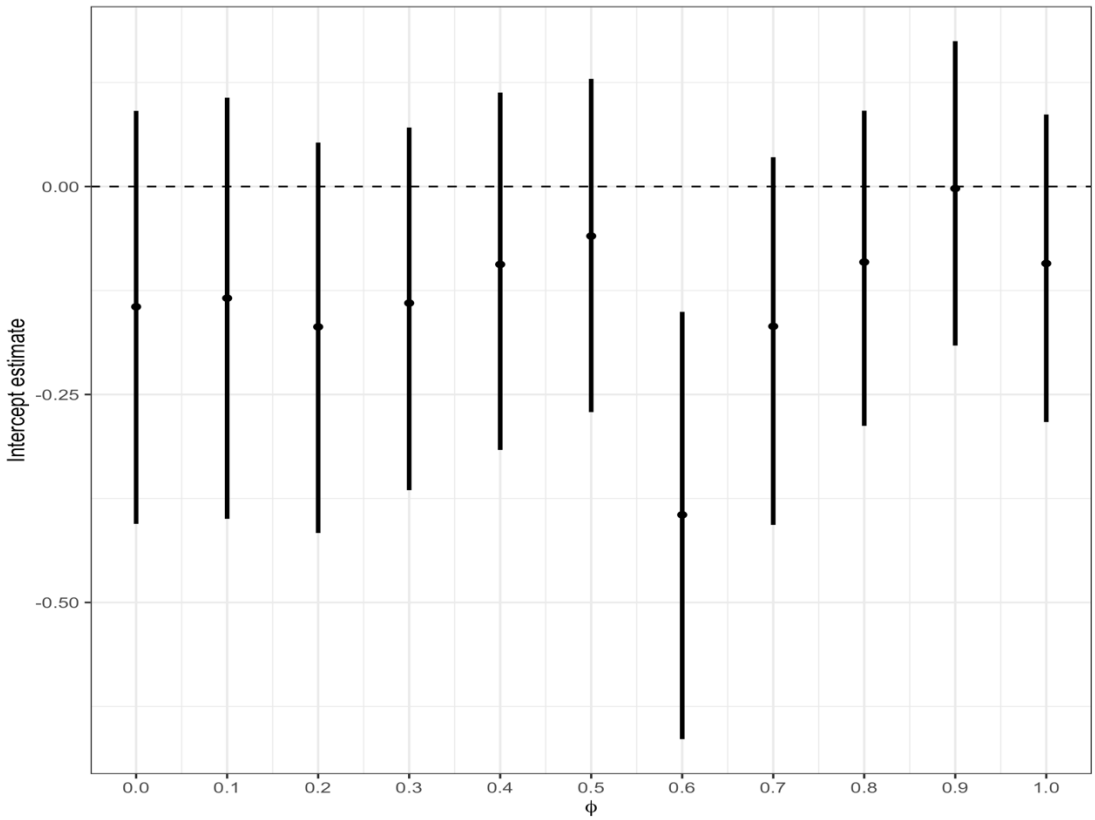


Figure D5: Mean and 95% credible interval of intercept coefficient estimates compared to the true simulated value, 0.

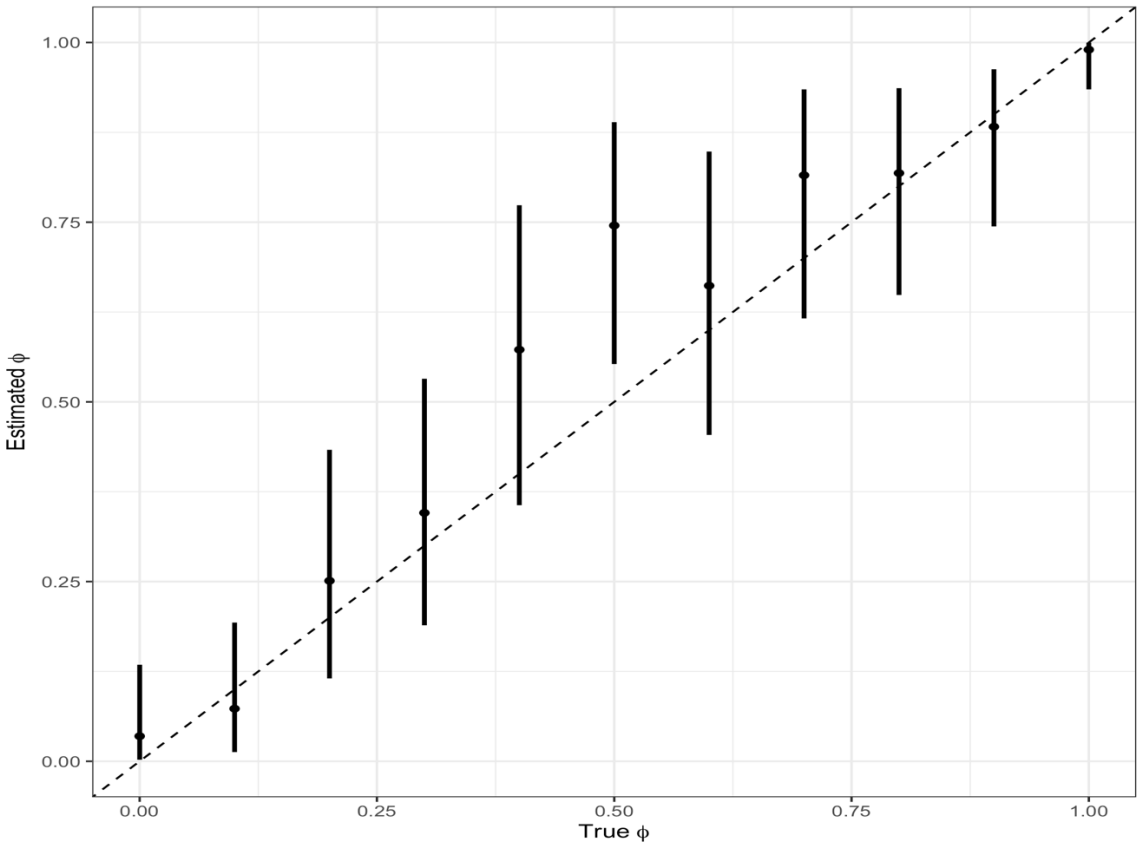


Figure D6: The mean and 95% credible interval of estimated ϕ values compared to the known value. Estimated ϕ values are calculated using the proportion of the random effect variance explained by the spatially structured term.

D.1.4 Simulation study: spatial modelling of binary data

Although the primary purpose of this paper is to develop a modelling framework compatible with count data, the smoothing spline approach could be used to structure random terms for other models, such as logistic models for binary outcomes. In this simulation study, we compare the proposed penalised smoothing spline approach to INLA's BYM2 model applied to a generated binary response.

D.1.4.1 Data generation

Binary response data, y_i, was generated for each of South Brazil's 1,013 municipalities from a binomial distribution with 20 trials or events, with a distance-based spatial structure:

$$y_i \sim Binomial(p_i, 20),$$

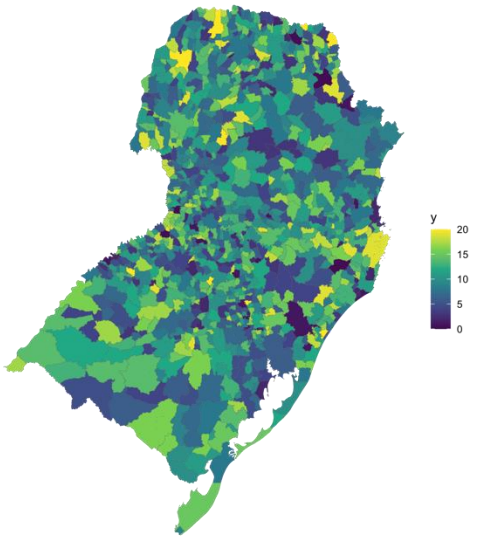
$$log\left(\frac{p_i}{1-p_i}\right) = \beta + S_i \quad (7)$$

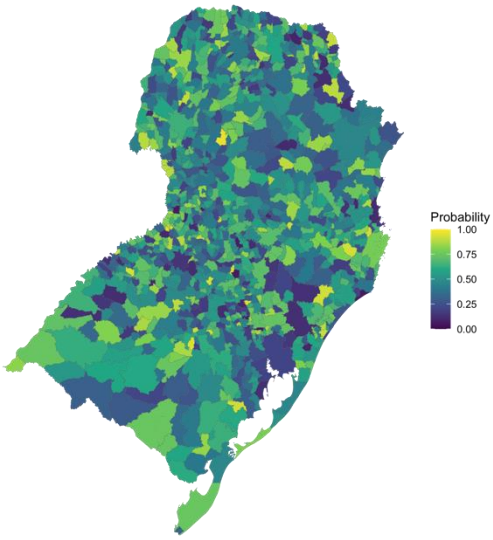
Where β is an intercept set to 0, p_i is the probability of an event, and S_i is the spatial structure of the data, defined as:

$$S_i = \sqrt{\phi} \cdot sm(x_i, z_i) + \sqrt{(1-\phi)} \cdot \varepsilon_i$$

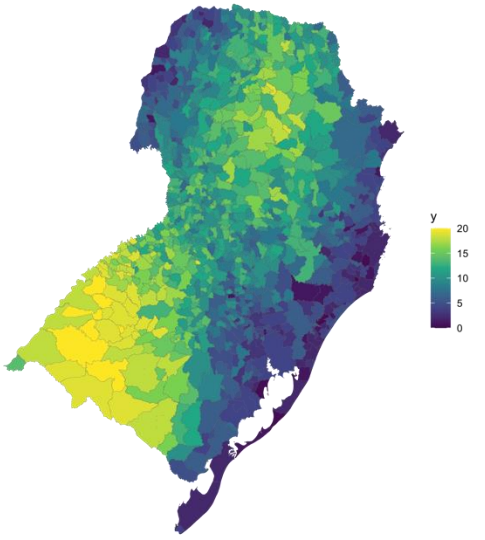
Where ϕ is a mixing parameter defining the relative contribution of a spatially structured term, $sm(x_i, z_i)$, and an unstructured term, $\varepsilon_i \sim N(0,1)$. A smooth function, sm (see equation (5)), was applied to the coordinates of the centroid of municipalities that were scaled to take values between 0 and 1. $sm(x_i, z_i)$ was centred around 0 by subtracting the overall mean from each value. 11 simulated datasets were produced by setting ϕ values between 0 and 1 at intervals of 0.1 (Figure D7).

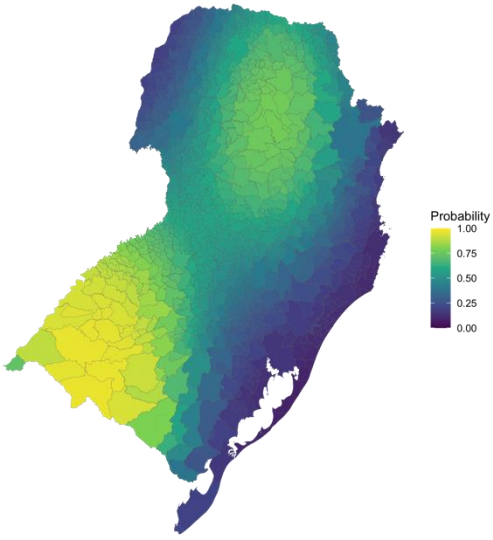
a)


y
20
15
10
5
0


Probability
1.00
0.75
0.50
0.25
0.00

b)


y
20
15
10
5
0


Probability
1.00
0.75
0.50
0.25
0.00

c)

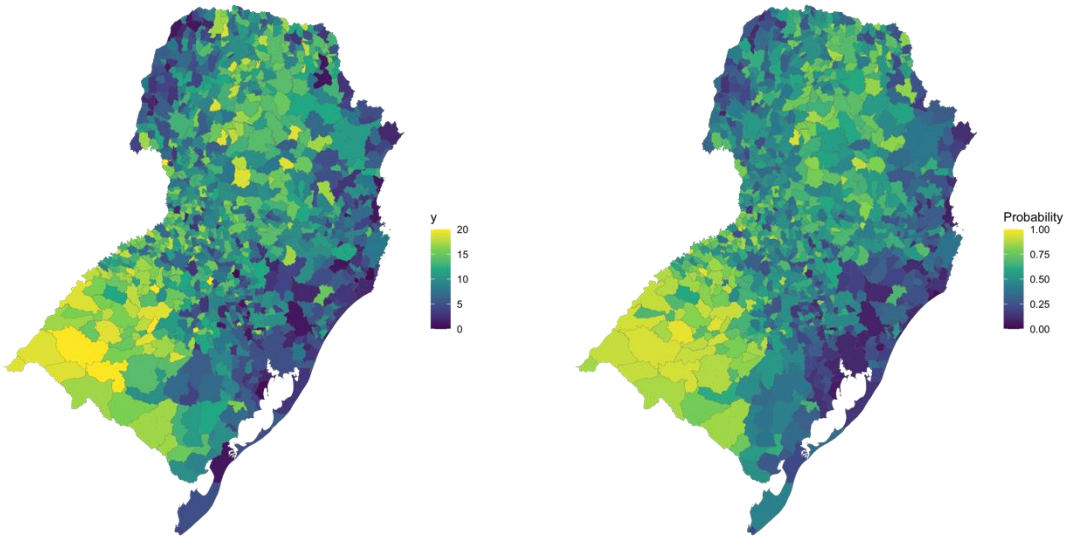


Figure D7: Simulated data (left) and probability of an event (right) under a distance-based structure using different spatial structure combinations. The number of events simulated from a binomial model and the probability of an event where the data has a) no spatial structure ($\phi = 0$), b) a distance-based structure only ($\phi = 1$), and c) equal contribution of both structures ($\phi = 0.5$).

D.1.4.2 Modelling approach

Two logistic models containing spatially structured and unstructured random effects were applied to each simulated dataset:

$$y_i \sim Binomial(p_i, 20)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \alpha + u_i + v_i \quad (8)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \alpha + \frac{1}{\tau}\left(\sqrt{\phi}u_{*i} + \sqrt{1-\phi}v_{*i}\right) \quad (9)$$

In model (8), u_i is a spatially structured random term representing spatial connectivity within the data, created be applying a thin plate regression spline to coordinates of the centroid of each municipality. v_i is a spatially unstructured term, expected to follow a zero-mean normal distribution, representing heterogeneity between regions. This spatial smooth model was compared to a CAR-based random effect model using R-INLA's BYM2 model specification,

given in equation (9) [4,18,19]. Here, u_{*i} is a spatially structured random effect assuming a CAR structure, where municipalities are considered connected if and only if they share a border. v_{*i} are unstructured random effects, and ϕ are mixing parameters that measure the relative contribution of the structured and unstructured random effects to the overall marginal variance $\left(\frac{1}{\tau^2}\right)$ of the random effect. The penalised smoothing spline structure was generated using R's mgcv package [20] and extracted using the jagam function [16]. Spatial smooth models were implemented using Markov chain Monte-Carlo (MCMC) simulations in R via the NIMBLE package [1]. The CAR model was created using INLA's BYM2 model [18,21].

We compared model performance between the spatial smooth and BYM2 models using receiver operating characteristic (ROC) curves, a comparison of the true positive and true negative rates, the Brier score, the mean squared difference between the predicted probability and observed outcomes, and WAIC. Higher values of area under the ROC (AUROC) curve and lower values of the Brier score and WAIC indicate a better model fit. The proportion of the random variance attributed to the spatially structured term was calculated from the smoothing spline model and compared to estimates of ϕ extracted from the INLA model output and the known values for each simulation.

D.1.4.3 Results

Model comparison statistics AUROC were almost equivalent between the spatial smooth and BYM2 models, while the Brier scores and WAIC values preferred the spatial smooth models (Table D2), indicating that the spatial smooth model performs as well (if not better) than one of the current standard approaches. This shows that the smoothing spline structure provides an alternative to CAR-based structures in binomial models as well as models for count data, particularly when the full spatial structure may not be fully understood or where the structure is neither stationary nor isotropic. Both models were able to accurately estimate the intercept coefficient (Figure D8).

Table D2: Model comparison statistics and mean estimates of the mixing parameter, ϕ, from the smoothing spline and INLA BYM2 models. Area under the receiver operating characteristic curve (AUROC), Brier score, WAIC, and ϕ estimates extracted from logistic models fitted to simulated data using smoothing splines or INLA's BYM2 model to structure random effects. The optimal goodness-of-fit statistic is given in bold, that is, the highest AUROC, the lowest Brier score and WAIC, and the ϕ estimate closest to the value used in each simulation.

ϕ	Smoothing spline model				INLA BYM2 model			
	AUROC	Brier score	WAIC	ϕ estimate	AUROC	Brier score	WAIC	ϕ estimate
0	0.764	**0.199**	**4937.1**	**0.008**	**0.764**	0.246	5084.4	0.014
0.1	0.764	**0.199**	**4932.8**	**0.077**	**0.764**	0.247	5075.3	0.126
0.2	0.762	**0.2**	**4926.1**	**0.204**	**0.762**	0.246	5053.6	0.283
0.3	0.763	**0.199**	**4909.4**	**0.278**	**0.764**	0.246	5031.1	0.358
0.4	0.759	**0.201**	**4906.1**	**0.373**	**0.759**	0.246	5002.8	0.457
0.5	0.758	**0.201**	**4884.1**	**0.487**	**0.758**	0.247	4967.6	0.533
0.6	0.749	**0.204**	**4849.7**	**0.615**	**0.749**	0.247	4906.3	0.682
0.7	0.749	**0.204**	**4806.5**	**0.71**	**0.75**	0.248	4857.2	0.747
0.8	0.754	**0.202**	**4733.2**	**0.797**	**0.754**	0.245	4756.5	0.934
0.9	0.751	**0.203**	**4645.8**	**0.877**	**0.754**	0.246	4626.3	0.976
1	0.733	**0.208**	**4289.2**	0.995	**0.748**	0.246	4324.6	**0.998**

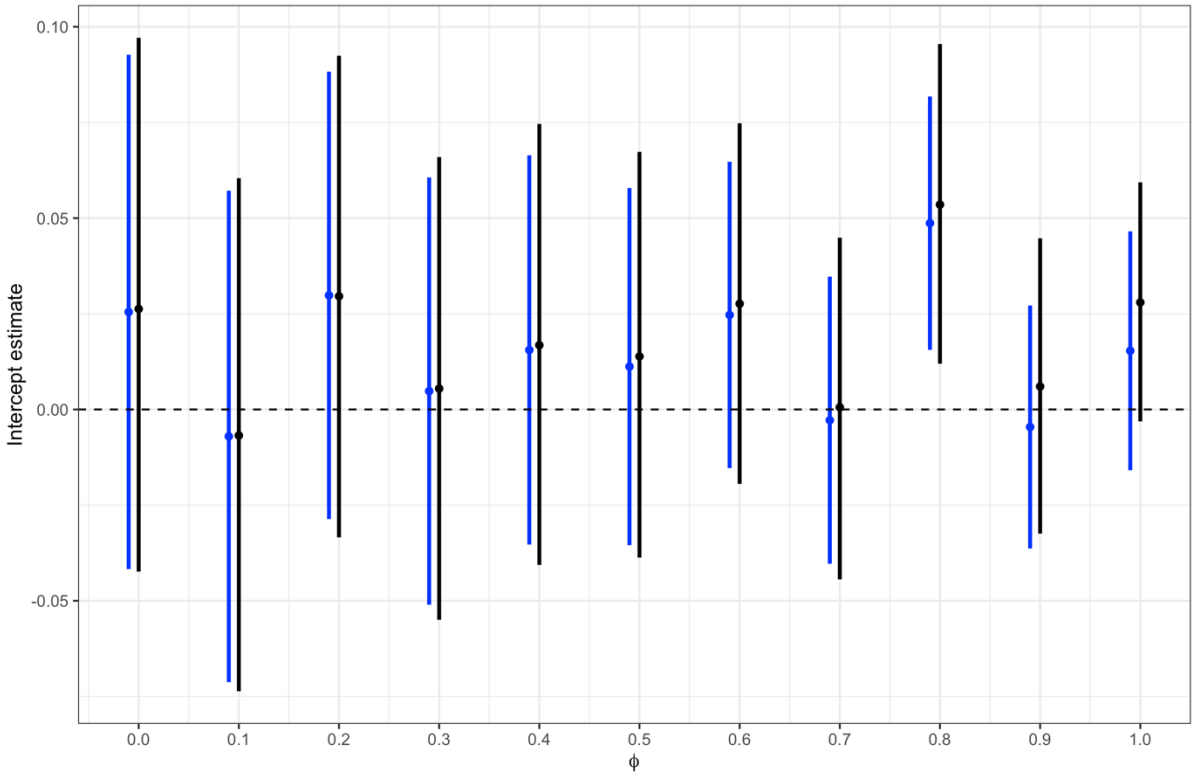


Figure D8: Mean and 95% credible interval of the intercept coefficient estimates from the smoothing spline (black) and BYM2 (blue) models, compared to the true simulated value, 0.

The proportion of the random effect variance explained by the spatially structured term in the smoothing spline model provided an accurate estimation of the true value from simulations (Figure D9). This value was comparable to the ϕ hyperparameter extracted from the BYM2 model using INLA and can therefore be interpreted in a similar way.

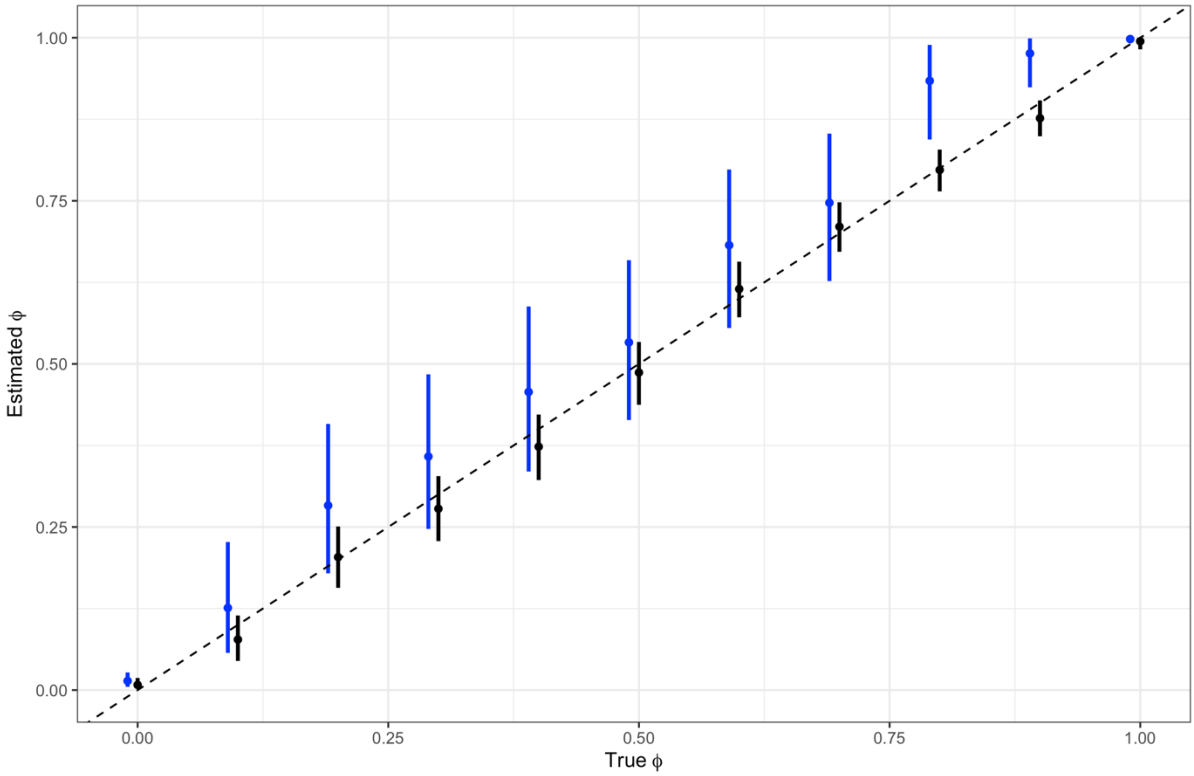


Figure D9: The mean and 95% credible interval of estimated ϕ values extracted from the smoothing spline (black) and INLA BYM2 (blue) models compared to the known value. Estimated ϕ values for the smoothing spline model were calculated using the proportion of the random effect variance explained by the spatially structured term and were extracted from INLA output for the BYM2 model.

D1.5 Simulation study: binary data with two sources of spatial structure

The binomial simulation study above was extended to also include a source of human movement-based connectivity.

D1.5.1 Data generation

Binary response data, y_i, was generated for each of South Brazil's 1,013 municipalities from a binomial distribution with 20 trials or events, using the same process as simulation study in Section S3, but with an extended spatial structure:

$$S_i = \sqrt{\phi_1} \cdot sm(a_i, b_i) + \sqrt{\phi_2} \cdot sm(c_i, d_i) + \sqrt{\phi_3} \cdot \varepsilon_i \quad (10)$$

$$\phi_1 + \phi_2 + \phi_3 = 1$$

sm is a smoothing function (equation (5)), applied to scaled coordinates of the centroid of municipalities, a_i and b_i, to create a distance-based smooth term, and applied to human movement-based connectivity coordinates, c_i and d_i, derived using multidimensional scaling and the Brazilian REGIC study [11] as described previously. ε_i is a random draw from a normal distribution ($\varepsilon_i \sim N(0,1)$), representing heterogeneity between municipalities. ϕ_1, ϕ_2 and ϕ_3 are mixing parameters which define the relative contribution of the spatially structured and unstructured terms to the underlying spatial surface. $\phi_3$was fixed at 0.2, $\phi_1$and ϕ_2 took values between 0 and 0.8 at intervals of 0.1 to create 9 simulated datasets (Figure D10).

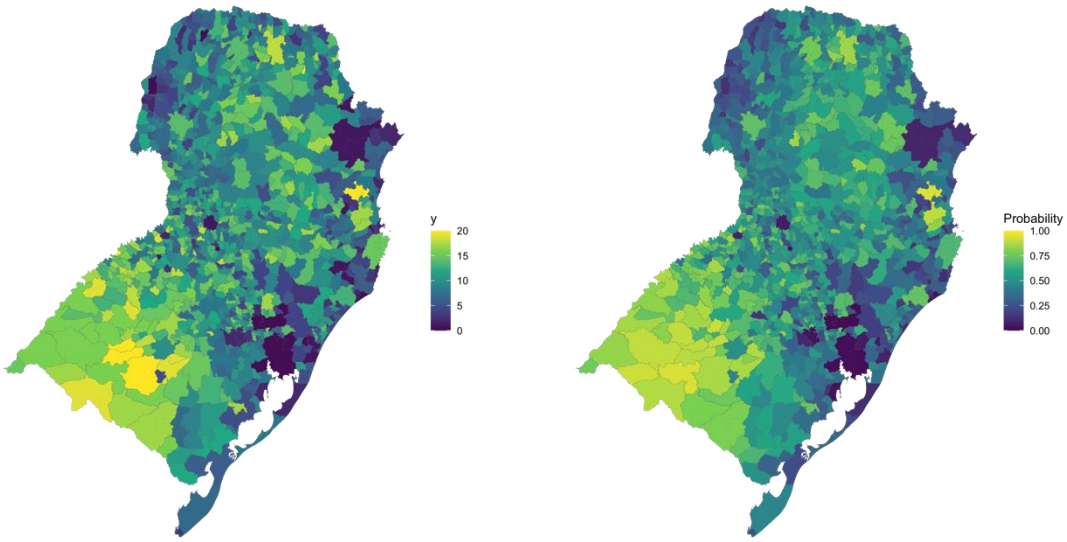


Figure D10: Simulated data (left) and probability of an event (right) containing two sources of spatial structure. An example of simulated data used to test the model, where $\phi_1 = 0.4$, $\phi_2 = 0.4$, and $\phi_3 = 0.2$.

D1.5.2 Modelling approach

A logistic model containing two spatially structured random terms and one unstructured random terms was applied to each simulated dataset:

$$y_i \sim Binomial(p_i, 20)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \beta + u_{1,i} + u_{2,i} + v_i \quad (11)$$

Where $u_{1,i}$ is a distance-based spatially structured random term, created by applying a thin plate regression spline to the coordinates of the centroid of municipalities. $u_{2,i}$ is a human movement-based spatially structured random term, created by applying a thin plate regression spline to coordinates describing relative connectivity between municipalities based on human movement (see Section D1.2). v_i is a spatially unstructured term, expected to follow a zero-mean normal distribution, representing heterogeneity between municipalities. Spatially structured terms were generated using R's mgcv package [20] and extracted using the jagam function [16]. Spatial smooth models were fitted using Markov chain Monte-Carlo (MCMC) simulations in R via the NIMBLE package [1].

D1.5.3 Results

The extended spatial model was able to accurately estimate the intercept coefficient for each simulated dataset (Figure D11). The ability for each model to detect the contribution of each spatial structure to the random effects varied across simulated datasets and between assumptions of connectivity (Figure D12). Although the model estimates were able to detect increases in the contribution of these structures, the estimates and 95% credible intervals did not always contain the true mixing parameter values from the simulation. In particular, this model underestimated the contribution of human movement to the spatial structure and attributed this to either distance-based or independent, unstructured terms for larger values of ϕ_2 (Figure D12). Therefore, as with the INLA BYM2 models compared previously, care should be taken when interpreting the estimates of the mixing parameters.

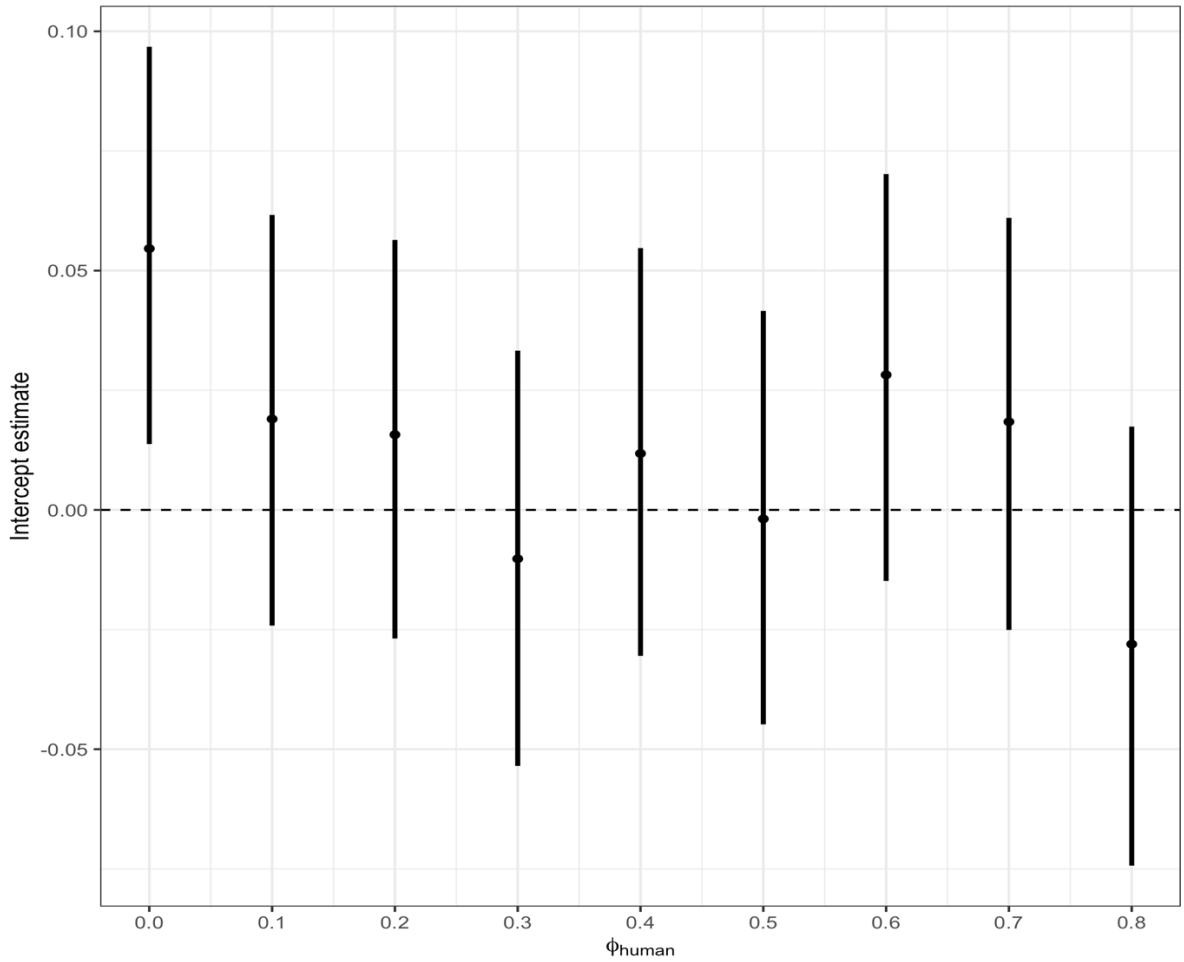
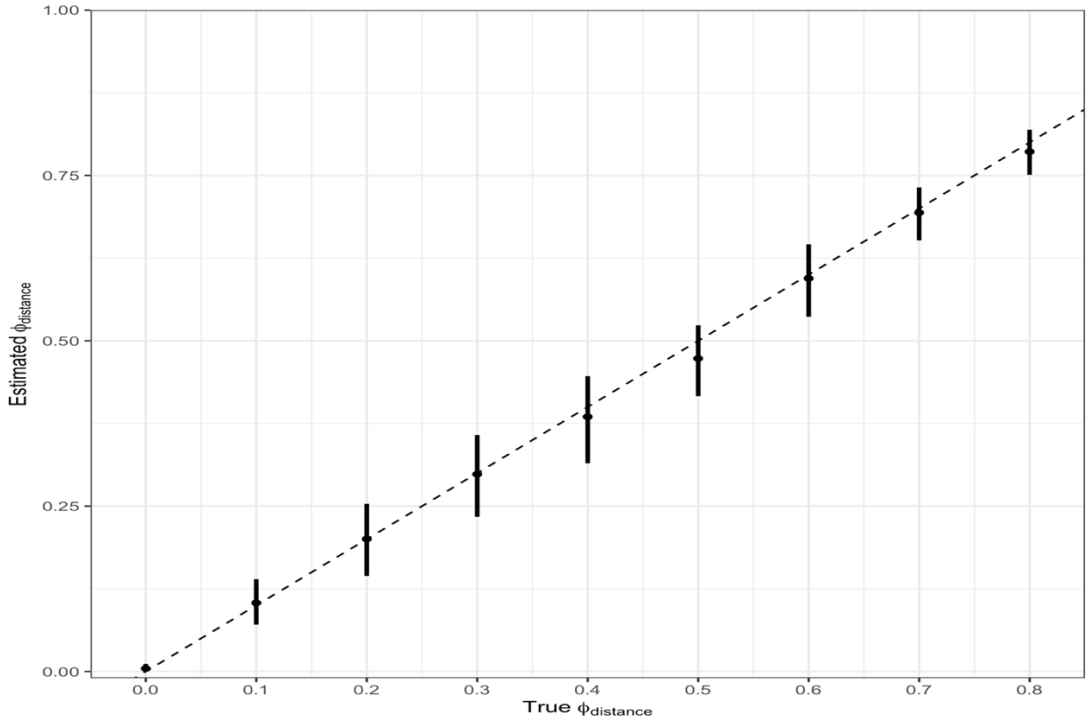
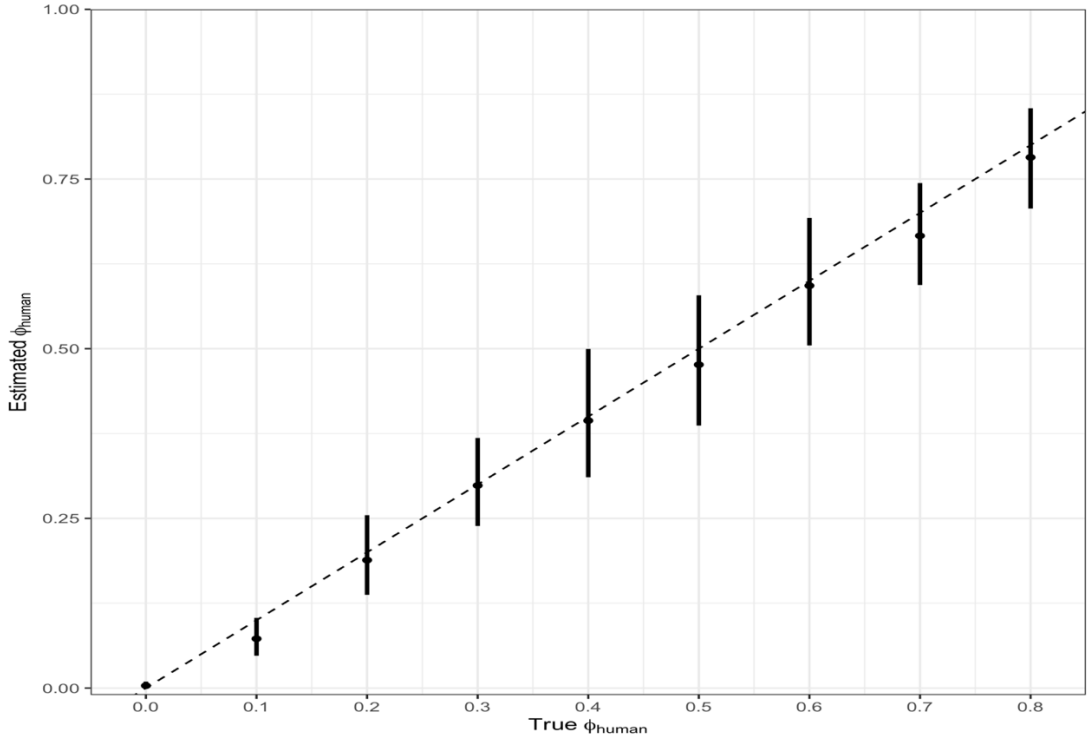


Figure D11: Mean and 95% credible interval of the intercept coefficient estimates, compared to the true simulated value, 0 (sorted by ϕ_{human}).

a)


1.00
0.75
0.50
0.25
0.00
Estimated $\phi_{distance}$
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
True $\phi_{distance}$

b)


1.00
0.75
0.50
0.25
0.00
Estimated ϕ_{human}
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
True ϕ_{human}

c)

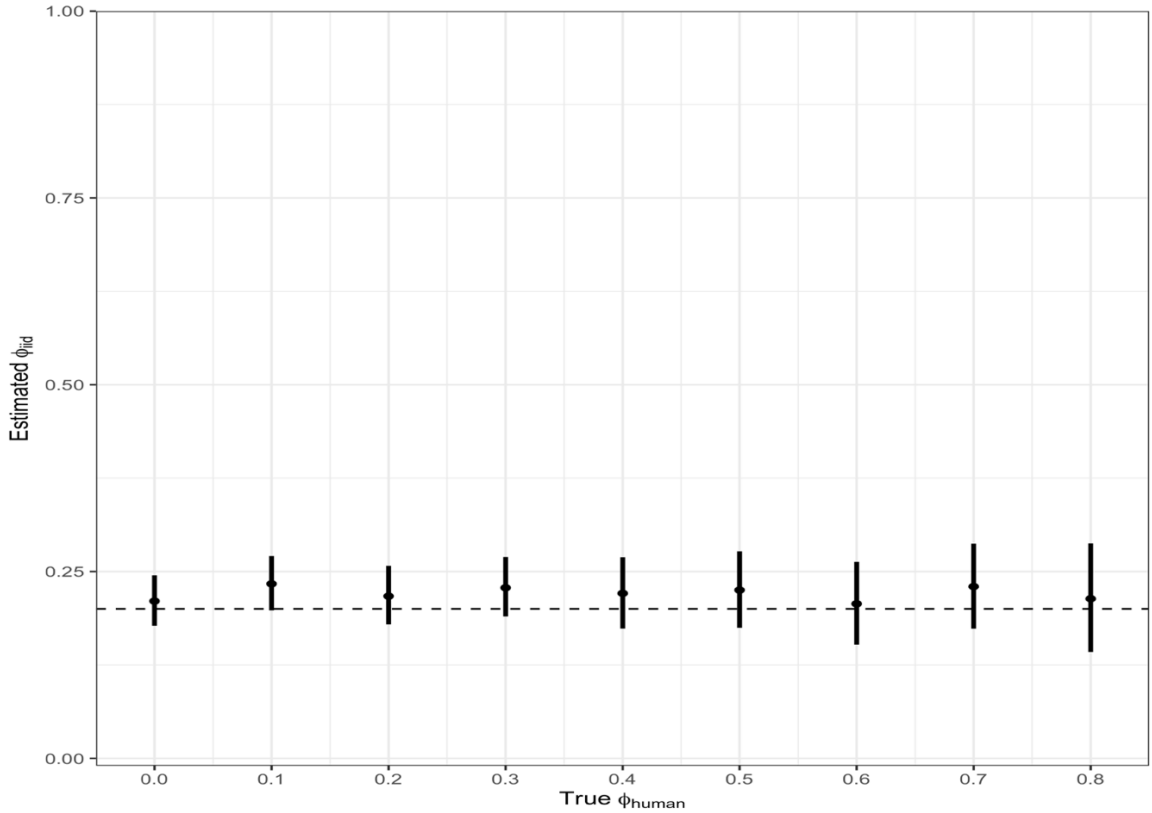


Figure D12: Mean and 95% credible interval of the proportion of variance of the random effects explained by a) the distance-based structure, b) the human movement-based structure, and c) unstructured random terms. Dashed lines represent the true value from simulations.

References

1. De Valpine P *et al.* 2021 Nimble: MCMC, particle filtering, and programmable hierarchical modeling. *R Package Version 011* **1**.

2. Riebler A, Sørbye SH, Simpson D, Rue H. 2016 An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25**, 1145–1165. (doi:10.1177/0962280216660421)

3. Besag J, York J, Mollié A. 1991 Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20.

4. Simpson D, Rue H, Riebler AI, Martins TG, Sørbye SH. 2017 Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. (doi:10.1214/16-STS576)

5. Martino S, Rue H. 2009 Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program. *Dep. Math. Sci. NTNU Nor.*

6. Gelman A, Hwang J, Vehtari A. 2014 Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016.

7. R Core Team. 2021 *R: A language and environment for statistical computing*. 4.1.1. Vienna, Austria. See https://www.R-project.org/.

8. Cox MA, Cox TF. 2008 Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347. Springer.

9. Simini F, González MC, Maritan A, Barabási A-L. 2012 A universal model for mobility and migration patterns. *Nature* **484**, 96–100.

10. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V. 2014 On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716. (doi:10.1371/journal.pcbi.1003716)

11. Estatística IB de G e. 2020 *Regiões de influência das cidades 2018*. IBGE Rio de Janeiro.

12. Wood SN. 2006 Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics* **62**, 1025–1036. (doi:10.1111/j.1541-0420.2006.00574.x)

13. Wood S, Wood MS. 2015 Package 'mgcv'. *R Package Version* **1**, 729.

14. Wood SN. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36. (doi:10.1111/j.1467-9868.2010.00749.x)

15. Wood SN. 2017 *Generalized additive models: an introduction with R*.

16. Wood SN. 2016 Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75**, 1–15. (doi:10.18637/jss.v075.i07)

17. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. 2017 Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.* **26**, 403–413.

18. Riebler A, Sørbye SH, Simpson D, Rue H. 2016 An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25**, 1145–1165. (doi:10.1177/0962280216660421)

19. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. 2017 Bayesian computing with INLA: a review. *Annu. Rev. Stat. Its Appl.* **4**, 395–421.

20. Wood SN. 2017 *Generalized additive models: an introduction with R*. CRC press.

21. Martino S, Rue H. 2009 Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program. *Dep. Math. Sci. NTNU Nor.*

D.2 Supplementary figures

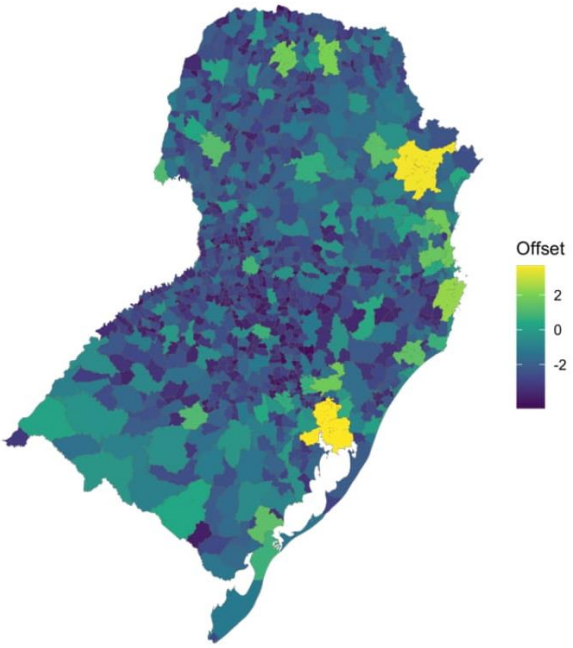


Figure S1: Offset, $log(\xi_i)$, the log of the population of each municipality divided by 100,000 in South Brazil.

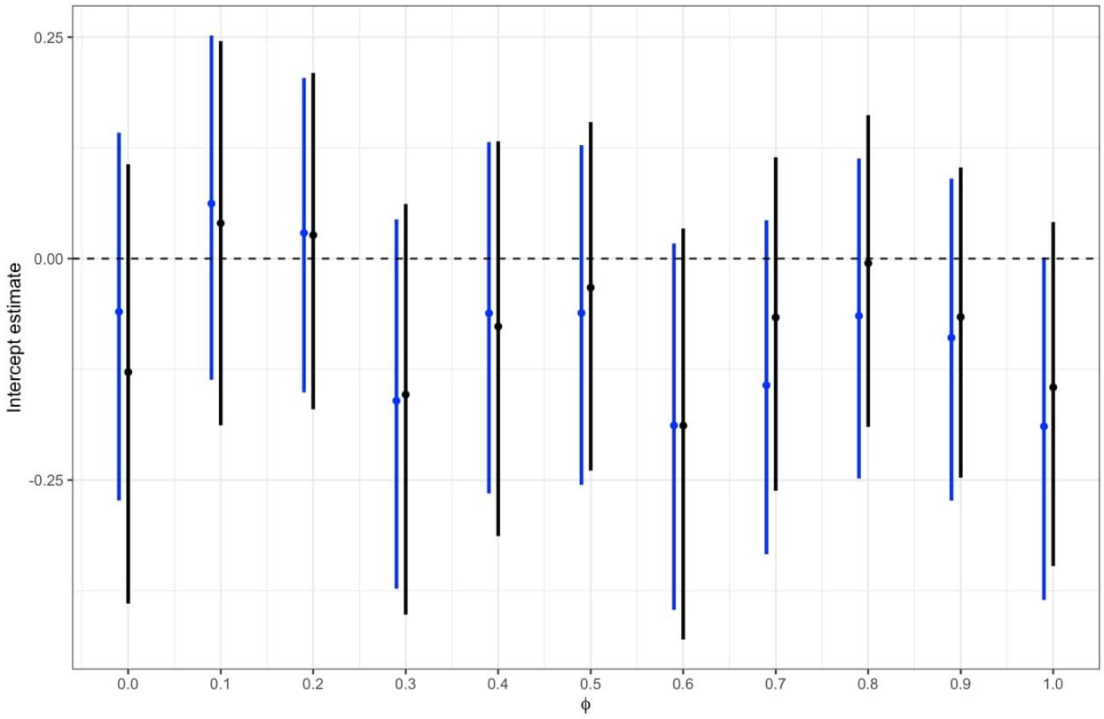


Figure S2: Mean and 95% credible interval of the intercept coefficient estimates from the smoothing spline (black) and BYM2 (blue) models, compared to the true simulated value, 0.

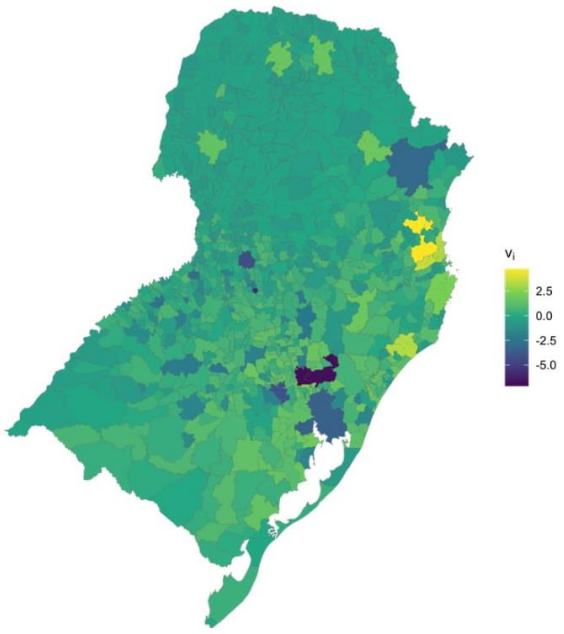


Figure S3: Human movement-based smooth function created by applying the function $sm(x_i, z_i)$ to coordinates describing the connectivity between municipalities in South Brazil arising from human movement.

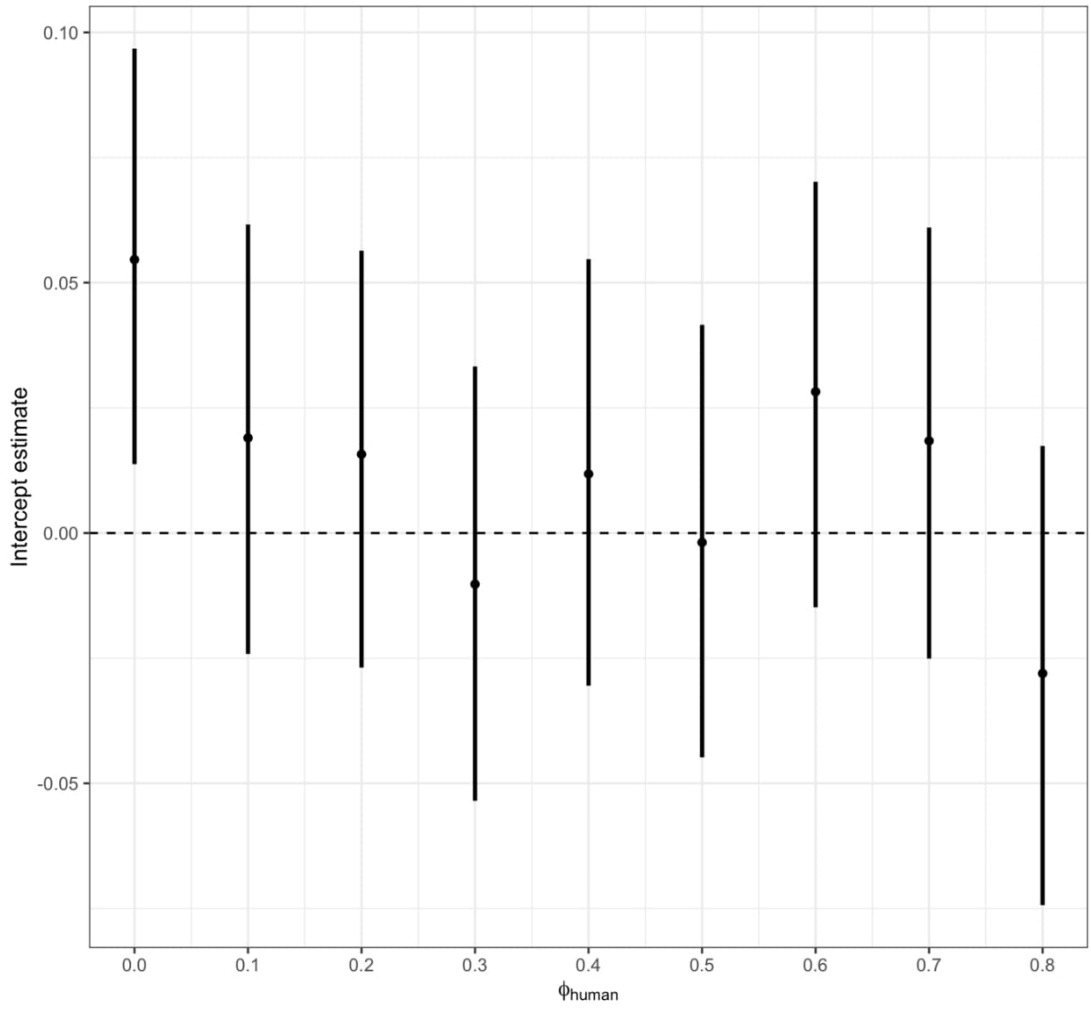


Figure S4: Mean and 95% credible interval of the intercept coefficient estimates, compared to the true simulated value, 0 (sorted by ϕ_{human}).

Appendix E: Supplementary Material Chapter 5

E.1 Supplementary figure

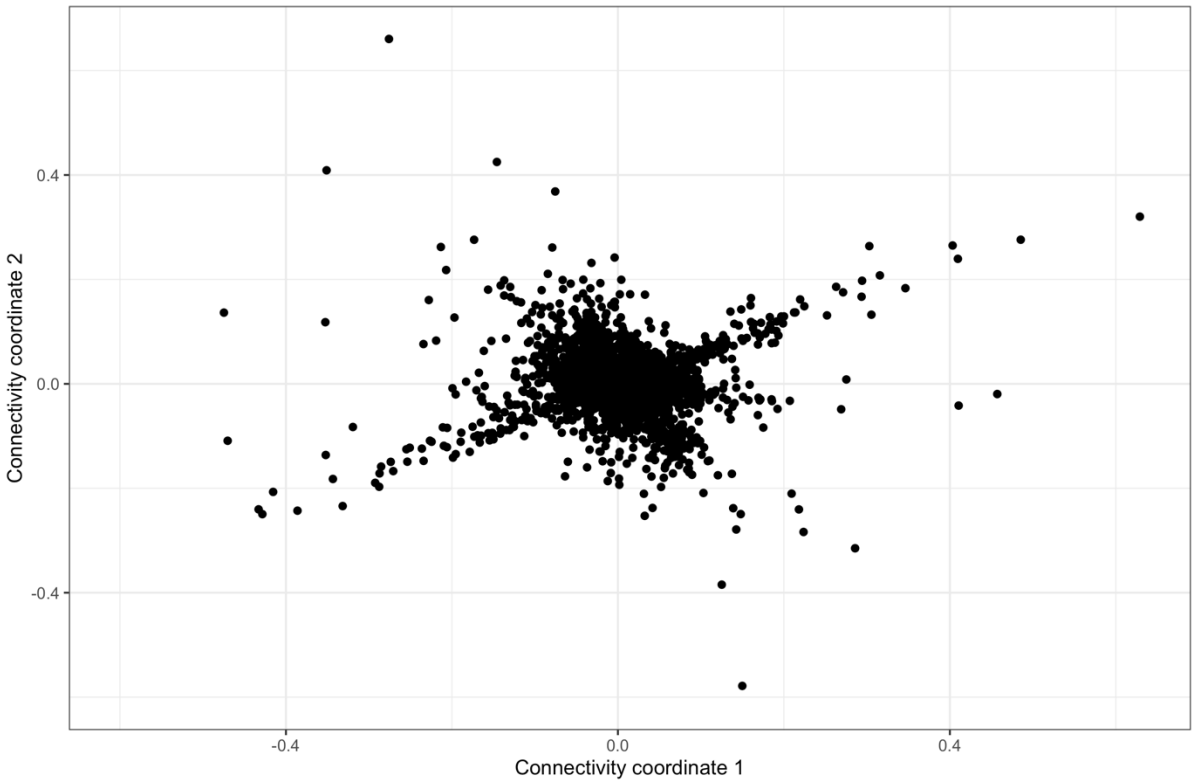


Figure E1: Human movement-based coordinates describing the connectivity between municipalities based on the number of residents moving between them. A coordinate system created by applying multidimensional scaling to the proportion of residents moving between municipalities. Dots closer together represent municipalities that experience more movement between them.