Miracle Amadi

HYBRID MODELLING METHODS FOR EPIDEMIOLOGICAL STUDIES

ACTA UNIVERSITATIS LAPPEENRANTAENSIS 1030

Miracle Amadi

HYBRID MODELLING METHODS FOR EPIDEMIOLOGICAL STUDIES

Dissertation for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism in the Auditorium 1316 at Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland on the 23rd of June 2022, at 2:00 pm.

Acta Universitatis
Lappeenrantaensis 1030

Supervisor Professor Heikki Haario
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Reviewers Professor Thomas Goetz
Department of Mathematics
University of Koblenz and Landau
Germany

Associate Professor Erin L. Landguth
School of Public and Community Health Sciences
University of Montana
United States

Opponent Associate Professor Erin L. Landguth
School of Public and Community Health Sciences
University of Montana
United States

ISBN 978-952-335-832-4
ISBN 978-952-335-833-1 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenranta-Lahti University of Technology LUT
LUT University Press 2022

Abstract

Miracle Amadi
Hybrid Modelling Methods for Epidemiological Studies
Lappeenranta 2022
86 pages
Acta Universitatis Lappeenrantaensis 1030
Diss. Lappeenranta-Lahti University of Technology LUT
ISBN 978-952-335-832-4, ISBN 978-952-335-833-1 (PDF), ISSN-L 1456-4491, ISSN 1456-4491

Epidemiological modelling plays an important role in the study of the distribution of a disease and its impact in a given population, and helps to suggest effective control and prevention measures. The complexity of models as well as the modelling approach vary depending on a number of factors such as how much is understood about the disease epidemiology, the objective of the study and the nature of the data available. Compartmental models have shown to capture the macro level dynamics of an infectious disease outbreak and have been utilised to develop control policies and outbreak responses. However, they contain a limited account of the complex processes of dynamics of most infectious diseases. Unlike the continuous modelling framework, the Agent-based modelling (ABM) approach features the simulation of heterogeneous communities subjected to more realistic transmission scenarios and can incorporate complex and stochastic issues affecting diseases. This work provides an example of how to utilise the strength of both kinds of models through a hybrid approach that combines in situ field data with the parameters of a classical malaria model. The ABM simulations provide a computational laboratory for generating data on the impact of some complex factors on malaria prevalence. The ABM results can be extended to continuous time by inserting the values fitted by the classical response surface regression as the key coefficients of compartmental models. Another regression approach presented in this work is a cluster-integrated regression which helps to screen the incidence clusters where the available explanatory variables fail to predict, using a panel data. The cluster-integrated regression method also improves the accuracy of the model by providing more explanatory variables. In addition, the spatial autocorrelation study using global Moran's I, the Geary's C and Moran's scatter plot was made to measure the timely spatial pattern of disease incidence in a country and to form the grouping. This was combined with a proposed metapopulation model that parameterises and reassesses non-pharmaceutical interventions. The uncertainty quantification of model outputs using Markov Chain Monte Carlo (MCMC) techniques was done based on the notion of randomness in the modelling approach.

Keywords: Agent-Based Models, Compartmental Models, Epidemiological Modelling, Regression analysis, Spatial Autocorrelation and Uncertainty Quantification

Acknowledgements

This research was conducted at Lappeenranta University of Technology, Finland, at the School of Engineering Science, division of Computational and Process Engineering. First and foremost, I would like to express my gratitude to the Academy of Finland for providing adequate funding throughout my studies.

My heartfelt appreciation goes to my supervisor, Professor Heikki Haario, who is always willing to contribute his knowledge and time. I found an exceptional academic mentor and true pillar of support in him throughout the research process. He responded to all my questions and enquiries promptly and I consider myself very lucky to have worked with a supervisor who understands my pace.

I would also like to thank the dissertation reviewers, Professor Thomas Goetz and Associate Professor Erin Landguth, for their time and thoughtful comments.

I would also like to thank some of the colleagues I worked closely with, including Anna Shcherbacheva and Karunia Putra Wijaya, for making working with them interesting and smooth. To my fellow graduate students, research technicians and collaborators who helped me in one way or the other with this work, I really appreciate your support.

My sincere gratitude goes to my family for their unwavering support and unending love. I am particularly grateful to my parents who have always supported and encouraged me as well as motivated me to obtain the education and training that was absolutely necessary for this undertaking.

Special thanks go to my lovely husband Emeka Godson for his patience, unwavering love and for encouraging and supporting my personal development. To our kids, Deborah and Samuel, I cherish you.

Miracle Amadi
June 2022
Lappeenranta, Finland

This dissertation is dedicated to my beloved husband and my parents.

Yours, Miracle!

Contents

List of publications

This dissertation is based on the following papers. The rights have been granted by publishers to include the papers in the dissertation.

I. Amadi, M., Shcherbacheva, A., and Haario, H. (2021). Agent-based modelling of complex factors impacting malaria prevalence. Malaria journal, 20(1), pp. 1-5.

II. Wijaya, K.P., Aldila, D., Erandi, K.K., Fakhruddin, M., Amadi, M., and Ganegoda, N. (2021). Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia. Stochastic Environmental Research and Risk Assessment, 35(2), pp. 437-456.

III. Ganegoda, N.C., Wijaya, K.P., Amadi, M., Erandi, K.K., and Aldila, D. (2021). Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany. Scientific reports, 11(1), pp. 1-6.

IV. Ganegoda, N.C., Wijaya, K.P., Chavez, J.P., Aldila, D., Erandi, K.K., and Amadi, M. Reassessment of contact restrictions and testing campaigns against COVID-19 via spatio-temporal modeling. (2021). Nonlinear Dynamics, pp. 1-25.

The publications are numbered throughout the dissertation using Roman numerals. Reprints of each publication are included at the end of this dissertation.

Author's contribution
Miracle Amadi is the principal author in Publication I and a co-author in Publication II, III and IV. In Publication I, the author contributed to the modelling approach, participated in running the simulations, drafted the manuscript and participated in editing the manuscript. In Publication II, the author participated in data interpretation, conducted preliminary analysis and participated in discussions and writing the manuscript. In Publication III and IV, the author participated in running the simulations and participated in drafting and editing the manuscripts.

List of all publications

The publications included in this dissertation are in bold.

I. **Amadi, M., Shcherbacheva, A., and Haario, H. (2021). Agent-based modelling of complex factors impacting malaria prevalence. Malaria journal, 20(1), pp. 1-5.**

II. **Wijaya, K.P., Aldila, D., Erandi, K.K., Fakhruddin, M., Amadi, M., and Ganegoda, N. (2021). Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia. Stochastic Environmental Research and Risk Assessment, 35(2), pp. 437-456.**

III. **Ganegoda, N.C., Wijaya, K.P., Amadi, M., Erandi, K.K., and Aldila, D. (2021). Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany. Scientific reports, 11(1), pp. 1-6.**

IV. **Ganegoda, N.C., Wijaya, K.P., Chavez, J.P., Aldila, D., Erandi, K.K., and Amadi, M. Reassessment of contact restrictions and testing campaigns against COVID-19 via spatio-temporal modeling. (2021). Nonlinear Dynamics, pp. 1-25.**

V. Suandi, D., Wijaya, K.P., Amadi, M., Ganegoda, N.C., Kusdiantara, R., Sidarto, K.A., Syafruddin, D., and Soewono, E. An evolutionary model propounding Anopheles double resistance against insecticides. (2022). Applied Mathematical Modelling, 106, pp. 463-481.

VI. Amadi, M., and Haario, H. Parameter and identification and forecast with a biased model. (2022). ECMI conference proceedings (Accepted for publication).

VII. Amadi, M., Killeen, G., and Haario, H. Models of acquired immunity to malaria: a review. Book chapter on Editorial Book 'Bio-mathematics, Statistics and Nano-Technologies: Mosquito Control Strategies'. (Submitted to Taylor and Francis)

1 Introduction

An epidemiological model is a system that simulates the natural progression of a disease and the outbreak of an epidemic. It incorporates and connects the most important epidemiological factors that influence infection dynamics. A model by definition is a simplification of natural processes. However, if properly constructed, it can simulate the natural dynamics and progression of an epidemic or an endemic situation, allowing the study of the disease dynamics and the impact of deliberate interventions on the natural course of infection transmission and, thus, disease incidence. When solid information on the natural history of a disease is available and the goal of the model is defined, it is important to move on with the identification of the epidemiological features and their relative significance. The spread of a disease is influenced by a number of factors. These factors include the infectious agent's sensitivity and resistance to treatments and vaccinations as well as its mode of transmission and infectious period. Factors within the population (i.e. geographic, economic, demographic and cultural factors) also play a role. If a given factor is thought to play an important role in the dynamics of a disease, a precise quantitative assessment of that factor should be made. It can be difficult to select important factors among the high number of them, and preliminary model simulations may be required in some cases to show their true importance, if any. On the other hand, including all conceivable factors would be inappropriate, as it would make the model too complex without necessarily adding to its efficiency. Moreover, it is doubtful that all the relevant data on a variety of factors would be available or straightforward to obtain in public health practice. Before starting to construct a model of a certain disease, it is necessary to think about the model's purpose and the practical goals it needs to accomplish. While model building may be motivated by scientific curiosity, the only socially meaningful goal of a model is to aid in the improvement of infectious disease control and/or treatment by a more rational application of the existing preventive and curative strategies and available resources. The prerequisites for building the model become obvious once the model's purpose has been determined.

Epidemiological models come in a variety of levels of complexity. Simple deterministic models or complex spatially explicit stochastic simulations are both feasible. Epidemiologists' approach is determined by a number of factors, including how much is known about the disease's epidemiology, the study's objective and the amount and quality of the data available. In the early twentieth century, mathematical modelling was extended into the field of epidemiology by scientists like Anderson Gray McKendrick and Janet-Leigh Claypon (see Rothman et al. (2008)). Mathematical modelling has, since then, become increasingly essential in managing outbreaks and epidemics as well as informing public health decisions.

Infectious diseases spread within populations as a result of the behaviour of both the infectious agent and the population. Models of how epidemics progress in a population are based on series of assumptions and information that are used to create a set of parameters which decide how effective intervention will be (e.g. social distancing or mass vaccination). This may be used to forecast future growth and spread patterns as well as a variety of other characteristics (see Hunter et al. (2020)).

1.1 Contribution of research

The overall contribution of this research is to present some modelling tools that may be used in answering vital questions in epidemiological studies. This work models the significance of some factors that impact disease spread in order to give insight on how they can be taken into consideration in planning intervention measures. The general contribution can be broken down into specific contributions made in each of the publications included in the dissertation as follows:

1.1.1 Computational laboratory by ABM

This work impacts malaria transmission modelling in terms of addressing the common pitfall of obtaining data that could be directly used for model calibration. In Publication I, a technique for combining in situ field data with the parameters of malaria transmission models is provided. The ABM simulations serve as a computational laboratory for generating data on the impact of several complex factors on malaria prevalence. The results of the simulations provides synthetic data for regression analysis, that enable essential parameters of classical malaria models, such as biting rates and vector mortality, to be calibrated.

1.1.2 Screening incidence data based on the available explanatory variables

The localisation of disease incidence that is correctly predicted by the available predictor variables has remarkable significance for public health officials. The clustering-integrated model associated with optimal barriers in Publications II and III aims to discover which ranges of the incidence of dengue fever and COVID-19 respectively, are well predicted by the available explanatory variables (e.g. meteorological factors) rather than looking at which ranges of the explanatory variables predict the incidence. This method aids in the evaluation and screening of incidence clusters where the explanatory factors are unable to predict. For example, if a given level of incidence data could not be explained using the available explanatory variables, it is possible that other factors, may have contributed to such incidence level. Additionally, the need for confounding factors to explain different incidence levels has been mitigated by the use of incidence clustering.

1.1.3 More explanatory variables and data for prediction

The clustering strategy in Publication II and III helps to decompose the data of explanatory variables into several more datasets that correspond to the clustering of the associated response variable. By providing more explanatory variables, this approach may improve statistical fitting (see West et al. (2006); Strand et al. (2011)). Moreover in Publication II, clustering was motivated by the substantial zero values in the incidence data and the limited availability of supporting data in Indonesia, while at the same time, requiring a model with proper fitting and strong predictability which is crucial in assisting decision-makers in the development of an early warning system.

1.1.4 Assessing the spatial autocorrelation of the incidence data

Evaluating the spatial autocorrelation of disease incidence data provides practical implications to assist decision makers especially in developing countries where medical and economic resources to combat the disease are limited. We use the global Moran's index and Moran's scatter (see Moran (1950); Geary (1954); Anselin (1995)) in Publications II, III and IV to evaluate the timely spatial pattern of disease incidence in a country and to group the data. Prioritising high-risk locations or hotspots is influenced by the cautious use of health care resources. The novelty of the work in Publication IV is the combination of spatial autocorrelation analysis (for hierarchical prioritisation of non-pharmaceutical interventions (NPIs)) with a newly proposed metapopulation model that parameterises and reassesses non-pharmaceutical interventions (e.g. physical distancing, wearing face masks, washing hands, mobility restriction etc.). The metapopulation model is utilised to assess the intra- and inter-cluster contact restrictions together with testing campaigns, given the absence of confounding factors.

1.2 Structure of the thesis

The thesis is organised as follows: Chapter 1 presents the objectives and contributions of this research. The compartmental model paradigm alongside its limitations is discussed in Chapter 2. In Chapter 3, the ABM approach along with the hybrid modelling method are presented. Regression models are discussed in Chapter 4, where the regression for panel data, clustering-integrated regression and Logistic regression was particularly discussed, among others. Measures of global and local spatial autocorrelation are presented in Chapter 5. In Chapter 6, the uncertainty quantification of model outputs was discussed with particular consideration of the MCMC method of uncertainty quantification. In Chapter 7, the summary of the results of each publication is presented. Chapter 8 presents the discussion and conclusion of the thesis. In Appendix A, the discussion of the ABM simulations following a recommended protocol is presented. At the end of this thesis, the references and the original publications that make up this dissertation are presented.

2 Compartmental models

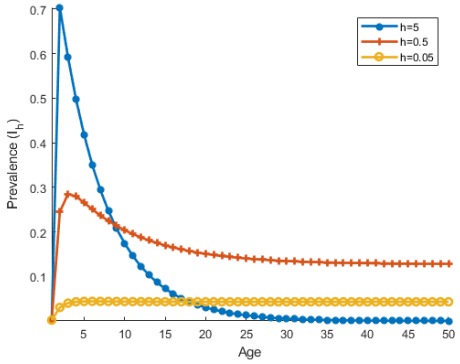
Modelling the transmission of infectious disease has traditionally been done using compartmental models. The models are frequently employed when dealing with large populations and they assign different populations to distinct subgroups. In compartmental models of disease transmission, the population under study is divided into classes, and assumptions are made about the nature and rate of disease transmission from one class to the next. The model is formed by using differential equations since the transition rates from one state to the next are mathematically defined as derivatives (see Brauer et al. (2019)). In a SIR model, for example, the population under study is divided into three compartments named S, I and R. Herewith, $S(t)$ represents the number of people who are susceptible to the disease, or who are not yet affected at time t. $I(t)$ represents the number of infected persons who are presumed to be infectious and capable of spreading the disease to susceptible ones through contact. $R(t)$ denotes the number of people who have been infected and then are no longer at risk of becoming infected or spreading infection. There are many variations of the SIR model, such as SIS model that incorporate births and deaths with no immunity upon recovery, $SIRS$ model where immunity lasts only for a short period of time, $SEIS$ and $SEIR$ model where the disease can be latent and the person is not infectious, and $MSIR$ model where infants may be born with immunity (see Brauer et al. (2019)). Depending on the model's purpose, there is a variety of other modifications.
A typical example of a compartmental model which was used to answer relevant questions related to immunity to malaria is the Aron and May model (see Aron and May (1982)). The model demonstrated how the burden of malaria is age-determined in endemic areas and what happens when transmission is reduced. In their age-specific SIRS model, the rate of loss of immunity γ, depends on the transmission rate h and parameterised as

$$\gamma = \frac{he^{-h\tau}}{1 - e^{-h\tau}}, \tag{2.1}$$

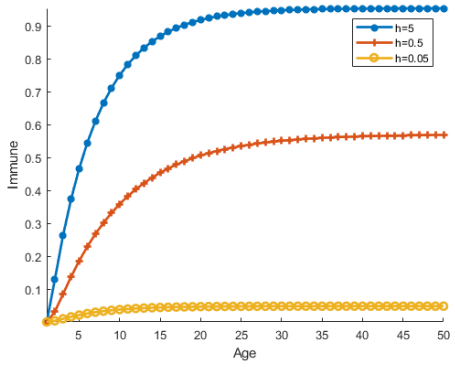
where τ is the duration of immunity. In the model, time is represented via the age of the cohort; this is typical for a population that has attained its equilibrium stage of infection. The model comprises three compartments in humans: *Susceptible* (S_h), *Infected* (I_h) and *Recovered and Immune* (R_h), where the effect of mosquitoes is introduced through the force of infection h (i.e. per capita rate of acquiring new infection per unit time (see Smith et al. (2006))). In this model,

$$\begin{aligned} \frac{dS_h}{d\alpha} &= -hS_h + rI_h + \gamma R_h \\ \frac{dI_h}{d\alpha} &= hS_h - rI_h - qI_h \\ \frac{dR_h}{d\alpha} &= qI_h - \gamma R_h, \end{aligned} \tag{2.2}$$

an infected person can recover at rate r and move directly to the susceptible class, or may slowly acquire immunity at rate q. They introduced immunity factor in their model by subtracting the people who lost immunity, γR_h from the immune class R_h and adding them to the susceptible class S_h. The plot of the simulated solution of the model Equation 2.2, given in Figure 2.1, reveals how prevalence changes with respect to age for different values of force of infection. With a higher rate of infection (h = 5/yr), typical for endemic areas, malaria prevalence rises speedily at a young age up to a peak, from where it gradually declines to a low level in adulthood, as a result of the increase in immunity. Contrarily, prevalence is shown to have an insignificant dependence on age for a low force of infection (h = 0.05/yr). This model predicts that, in highly endemic areas, the prevalence rapidly rises in early childhood and gradually wanes into adulthood as a result of slow acquisition of immunity with age and time. It can also be seen that the prevalence in adults is highest at intermediate infection rates. This is consistent with the infection pattern summarised by Boyd for tropical Africa (see Boyd (1949)) and also with the speculations of some epidemiologist that partial control, which leads to a moderate reduction of transmission from initially high levels, could increase adult prevalence (see Colbourne et al. (1966); Molineaux et al. (1980)). While immunity to malaria generally rises with age, especially in places with the highest forces of infection and stabilises at adulthood, this increase with respect to age is not noticeable at a low force of infection.



(a)



(b)

Figure 2.1: (a) Age prevalence and (b) immune curve simulated using the Aron model (see Aron and May (1982)) for different forces of infection with $r = 0.9$/yr, $q = 0.25$/yr and $\tau = 5$yrs.

2.1 Metapopulation model

Metapopulation models are a sort of spatial model that studies the interactions and movements of distinct subpopulations of the same species through time and space (see Ross (1915); Boyce (2007); Goethert et al. (2009)). It is a variation on more traditional population-level compartment models, which often assume homogeneous mixing and implicit population interactions. The metapopulation model could be modified to capture the explicit dynamics of pathogen migration, replication and degradation since pathogens travel between different hosts and environments. Metapopulation models are commonly written as a set of ordinary differential equations (ODE) that may be solved analytically without requiring a lot of computing power.

Metapopulation models were initially developed in ecology for scenarios when a population can be divided into a number of geographically distant sub-populations (see Levins (1969); Hanski et al. (1999)). It can, however, be applied more broadly to any segmentation of a population into groups that has an impact on infectious disease dynamics.

Mobile individuals having memory over their origin zones and short visits to other zones that allow them to infect other humans or acquire illnesses ex situ can be included in the metapopulation data (see Arino and Van den Driessche (2003); Sattenspiel and Dietz (1995)), or they can be without memory (see Wang and Zhao (2004); Li and Shuai (2009)). The concept of travel time becomes the most important parameter in the first case.
Most epidemiological metapopulation models are currently limited to direct host-to-host transmissions and are solely focused on the host (see Keeling and Rohani (2008); North and Godfray (2017)). In these metapopulation models, pathogens are implicitly considered (typically via direct host-host contact) but rarely quantified (only population sizes of host in various epidemiological states are evaluated). Nonetheless, the inseparable epidemiological triangle of host, pathogen and environment is the core notion and basis of modern epidemiology. As a result, mathematical models that explicitly address pathogens may be able to better characterise the system.
In terms of incorporating movement data into an epidemiological model, the question is whether choosing one mechanistic movement model over another has an impact on the modelled epidemiological outcomes, and whether there are certain epidemiological settings where one kind of movement model is more appropriate than another (see Citron et al. (2021)). In Citron et al. (2021), models of infectious disease dynamics have been demonstrated to be sensitive to the choices made by modelers when considering movement-mediated interactions between subpopulations. The Eulerian and Lagrangian movement models are two of the most basic and widely used movement models. In principle, more extensive and complicated movement models, such as adding stochastic effects, or employing agent-based models to integrate movement behaviour heterogeneity among the travellers, can be employed. More broadly, the question of which model to apply in which situation or how to model disease spatial dynamics deserves more consideration.

2.2 Important thresholds learned from compartmental models

Thresholds are an important concept in compartmental models in epidemiology. These include crucial values, such as the vector density, contact number, and reproduction number. These thresholds are essential in determining whether a disease becomes an epidemic, persists in the population or becomes endemic. These limits also aid in determining how transmissible a disease is.

2.2.1 Basic reproduction number

The basic reproduction number $\mathscr{R}_0$ is a key concept in epidemiological studies. The threshold which is mostly used in compartmental models for the transmission of infectious diseases, represents the mean number of secondary infections produced by a single infective released into a susceptible population (see Diekmann et al. (1990); Perasso (2018)). If $\mathscr{R}_0 < 1$ there is an asymptotically stable disease-free equilibrium, and the infection dies out. When $\mathscr{R}_0 > 1$, the disease usually persists since there is a unique endemic equilibrium that is asymptotically stable. The 'next-generation matrix' approach

is a well-known method for computing $\mathscr{R}_0$ for ODE models. The dominating eigenvalue of the next generation operator $\mathscr{E}$ is designated as the $\mathscr{R}_0$ value (see Diekmann et al. (1990)). According to Perron-Frobenius theorem, $\mathscr{R}_0$ is a simple eigenvalue of $\mathscr{E}$ if $\mathscr{E}$ is irreducible. On the other hand, if $\mathscr{E}$ is reducible, as it typically is for diseases with several strains, $\mathscr{E}$ may contain several positive real eigenvectors correspondingly for the reproduction numbers of each competing disease strain (see Brauer et al. (2019)).
Explicitly, the importance of calculating $\mathscr{R}_0$ include:

- $\mathscr{R}_0$ can be used as a predictive tool for preventing disease outbreaks. The idea is to link a dynamic model with temporal data. By 'to link', we refer to the process of determining the values of the model's parameters such that the solution of the dynamical system matches the data (see Kermack and McKendrick (1927)).

- Calculating $\mathscr{R}_0$ is needed so as to predict the threshold of, say, the vaccination needed for herd immunity (see Ross (1911)).

- $\mathscr{R}_0$ can be used to assess the impact of biodiversity on the transmission of tropically transmitted parasitic diseases (see Baudrot et al. (2016)).

- $\mathscr{R}_0$ can also be used in studying disease extinction and persistence. The study of the temporal symptotic features of the solutions of epidemic dynamical systems is a mathematically interesting topic pertaining to the basic reproduction number. Indeed, when $\mathscr{R}_0 < 1$, the DFE is locally asymptotically stable, under satisfactory assumptions on the next generation matrix for ODE systems. However, when $\mathscr{R}_0 > 1$, it becomes interesting to see how the infected curve behaves, and whether or not this coincides with disease persistence (see Diekmann et al. (1990)).

However, in some cases, the conventional prerequisite of the reproduction number being less than one is only essential, but not sufficient, for disease eradication. This happens when there is a backward bifurcation, a condition where a stable endemic equilibrium overlaps with a stable disease-free equilibrium when the associated $\mathscr{R}_0$ is smaller than one (see Hethcote et al. (1981)). Once $\mathscr{R}_0$ exceeds 1, the disease can spread to a reasonably high endemic level. In this situation, lowering $\mathscr{R}_0$ to its previous level will not necessarily result in the disease disappearing. Another possible behaviour is a forward, or transcritical, bifurcation at $\mathscr{R}_0 = 1$, with an asymptotically stable endemic equilibrium and an equilibrium infective population size continuously depending on $\mathscr{R}_0$. The requirement for $\mathscr{R}_0 < 1$ is both necessary and sufficient for disease elimination in models with forward bifurcation. The behaviour during a bifurcation can graphically be represented by the bifurcation curve (i.e. the plot of the infective population size I at equilibrium as a function of $\mathscr{R}_0$) (see Brauer et al. (2019)).
In publication IV, a metapopulation model for Moran's clusters was designed based on the available panel COVID-19 incidence data from Sri Lanka. This is used to assess the intra- and inter-cluster contact restrictions as well as to test campaigns in the absence of confounding factors. In the model, the concept of memory is employed, but unlike in Arino and Van den Driessche (2003); Sattenspiel and Dietz (1995), just the number of

persons from cluster i who are in cluster j are known, and hence at which cluster the contacts occur is not displayed. To determine the long-term trend of the regressing solution around disease-free and endemic equilibria, we evaluated the role of $\mathscr{R}_0$. This features an analytical bifurcation analysis based on Brouwer Degree Theory and asymptotic expansions around the $\mathscr{R}_0$ as well as associated numerical analyses based on path-following techniques. Herewith, the disease-free equilibrium is disrupted and an endemic branch emerges (through a forward bifurcation) for the strongly connected network between clusters.

2.2.2 Entomological inoculation rate (EIR)

Also with the help of deterministic compartmental models, the entomological inoculation rate (EIR) can be calculated. EIR is commonly evaluated to quantify the intensity of an infected mosquito pool and its potential to spread malaria infection to the human population within a certain time period. This is typically understood as the number of *Plasmodium falciparum* infective bites a person has received during a given period of time (i.e. per night, monthly, seasonally or annually) (see Kilama et al. (2014)). Mathematically, it is calculated by multiplying the human biting rate by the sporozoite rate (see Shaukat et al. (2010))

$$EIR = maS. \tag{2.3}$$

The human biting rate (ma) is described as the number of vectors biting a person over a given period of time. m denotes the number of *Anopheles* per person and a denotes the average number of individuals bitten by an *Anopheles* in one day. The sporozoite rate (S) is the fraction of infectious vector mosquitoes present.
In Publication I, the simulations of the agent-based model are run over a one-night 'snapshot' period. By using the response surface values as the key coefficients of classical compartmental models, the results are extended to continuous time. As a result, the effects of intervention measures or socioeconomic variables were modelled over longer time periods and to steady state. This enables the EIR values to be calculated in a wide range of transmission conditions.

2.3 Complexity reduction: partially observed model

A partial observation of state variables is one of the challenges encountered in compartmental models when there is a need to reveal some hidden dynamics of the system based on the available data. It usually leads to the structural non-identifiability of model parameters even for rather simple models (see Li and Vu (2015); Miao et al. (2011)).
One approach for tailoring the model complexity to the content of a data is to reduce the complexity of the model in accordance with the available data, resulting in a reduction in the dimension of the ODE system (see Wieland et al. (2021)). A method for addressing this problem was suggested by Xia and Moog (see Xia and Moog (2003)), using the Implicit Function Theorem. All latent variables (unobservable system state variables) can be removed after algebraic calculations by taking derivatives of observable system outputs with respect to independent variables, such as time and age, and a finite number

of equations comprising known system inputs, observable system outputs and unknown parameters can then be formulated. This is owing to the practical need for parameters to be stated as functions of the system's known quantities. A simple example is with the Ross model (see Ross (1911)) which has since played a key role in the development of studies on mosquito-borne pathogen transmission and has had a significant impact on the development of malaria control strategies. Using two differential equations for the human and mosquito, the model depicts the time evolution of the fraction of individuals in infected groups (ih, im).

$$\begin{aligned} di_h &= mabi_m(1 - i_h) - i_h r \\ di_m &= aci_h(1 - i_m) - \mu i_m, \end{aligned} \tag{2.4}$$

where i_h and i_m represent the proportions of infected humans and mosquitoes, respectively, m denotes mosquito-to-human ratio, b and c stand for the probabilities of transmission during mosquito contact with the humans, μ_m is the mosquito mortality rate, a denotes the contact rate and r is the recovery rate for humans. In many cases, the data on the infected mosquito population is not available for fitting. Considering that the presence of mosquito dynamics gives an additional degree of freedom, a reduced model which has only the infectious human compartment can be proposed. More detailed analysis to separate the vector dynamics from the host dynamics, which typically requires different time scales, has been discussed in Rocha et al. (2013). The equilibrium solution of infected mosquitoes given as

$$i_m^* = \frac{i_h}{i_h + \kappa}, \text{ where } \kappa = \frac{\mu}{ac}, \tag{2.5}$$

can be plugged into Equation 2.4 and parameterised as

$$\frac{di_h}{dt} = mab\frac{i_h}{i_h + \kappa}(1 - i_h) - i_h r. \tag{2.6}$$

With this, the model can easily be fitted with the data on infectious humans which is commonly available.
In Publication IV, based on the lack of related field data, a complexity reduction for the metapopulation model is provided, resulting in a simple (SI type) model that is rational enough to mediate contact restrictions and testing campaigns.

2.4 Limitations of compartmental models

Although compartmental models have been shown to represent the macro dynamics of infectious disease outbreaks and have been utilised in the creation of control policies and epidemic responses, there are several drawbacks to utilising them (see Hunter et al. (2018)). Compartmental SEIR models, in general, are not sufficient for reproducing the real dynamics of some diseases such as malaria, as they allow only a limited account of the complex process of malaria transmission. They make clearly artificial assumptions that seem to make them conceptually compelling, but they are actually inefficient. One considerable reason is that malaria modelling requires an in-depth study of in-host par-

asite dynamics rather than a mere presence or absence of infection and prevalence in a group of population. Again, the important sources of heterogeneity, spatial and temporal scales of transmission remain inadequately addressed using deterministic models.
There is usually a trade-off between simple, or strategic, models, which exclude most details and are only aimed at emphasising the overall qualitative behaviour, and comprehensive, or tactical, models, usually tailored for specific scenarios including short-term quantitative projections. For instance, simple epidemic models predict that an epidemic will die out after a period of time, leaving a portion of the population disease-free. This qualitative principle is not really helpful in determining which control measures would be most effective in a given circumstance, but it does imply that public health professionals might benefit from a detailed model that describes the situation as accurately as possible.

3 Agent-based models

Agent-based modelling (ABM) is a type of computational modelling approach where individual entities in a complex system are represented as discrete agents that interact independently in a simulated space and time by following a set of established rules in order to fulfil their goals, e.g. survival, reproduction or economic profit (see Gilbert (2019)). As a result, the system frequently displays more complicated behaviour than it would be when agents act individually. The concept of agent-based modelling was developed in 1940 and gained widespread adoption in the 1990s, thanks to significant improvements in computational capabilities and the introduction of more software development environments such as Netlogo, Swarm and Repast, which were designed to allow non-computer programmers develop and understand ABMs.
In recent years, ABMs have gained prominence in epidemiological studies. They have virtually exclusively been used to represent infectious disease transmission and control in populations. An epidemiological agent-based model has four basic components: disease, society, transportation and the environment (see Hunter et al. (2017)). It must be identified how the infectious disease is transferred between agents and how the disease progresses in an infected agent when modelling the disease. Simulating society entails simulating the population, whereas simulating transportation entails simulating how the agents would move through the environment. The process of modelling the environment entails designing the space in which the agents will interact.

3.1 Agents, environment and events

Agents are simplistic depictions of real-world entities that handle problems by perceiving and reacting to stimuli from other agents and their surroundings. An agent could be anything, from a single person to a large organisation or entity like a nation state. They are being programmed to respond to the computing environment in which they are placed. Their activities are governed by a collection of coded rules (see Gilbert (2019)) which may be fixed or learned (see Russell and Norvig (1995)). Agents' perceptions of stimuli can be linked to the condition side of decision rules, which then trigger appropriate behavioural reactions on the action side. An agent decides what it will do at each time step; the actions can be as simple as defining which direction an agent will move to based on some simulated perceived notion, or they can be more complex, such as searching for agents with specific characteristics within a specified range and socially interacting with them. The ability for agents to interact, that is, pass informational signals to one another and act on what they learn from these messages, is a critical aspect of agent-based models. Collective phenomena emerge as a result of the agents' decision-making heuristics. The performance of each agent is typically assessed using a utility function connected to decisions taken by the agents. Actions with higher utility values have a better possibility of being approved.
In Russell and Norvig (1995), agents are categorised into four classes: utility-based agents, goal-based agents, agents with internal states and reflexive agents. These classes of agents are ranked by the sophistication of their decision-making processes. Goal- and

utility-based agents, in particular, are intelligent agents with learning mechanisms that enable them to change their decision rules in response to system dynamics by utilising machine learning algorithms such as neural networks and evolutionary algorithms.
The virtual world in which the agents operate is referred to as the environment. It could be a completely neutral medium with little or no effect on the agents, or it could be as precisely designed as the agents themselves in other models. The interactive environment could be a physical space or a geographic map, with attractions and obstacles for the agents to follow and overcome. A continuous-based method or a grid-based approach can be used to depict the environment. In the continuous-based representation, the environment is regarded as a set of topologically connected geometric features (e.g. points, polygons and lines), with continuous coordinates describing the size, shape and position of the objects. In the latter approach, the environment is divided into a collection of grid cells, where each element of the grid is separated into various state variables connected to the environment of interest (see Gilbert (2019)). Neighbourhood adjacency rules (such as von Neumann and Moore neighbourhood rules) can be used to construct associations between spatial features (see Tang and Bennett (2010)).
ABM models the dynamics of real-world systems using a collection of events that update the state of agents and environments. Characterising agent-agent and agent-environment interactions requires an explicit consideration of events. A sampling event in a real-world system can be equated to each iteration of an agent-based model. For the modelling of dynamics in real-world systems, ABM's temporal extent and resolution must be correctly determined. Internal (e.g. change in the internal state of an agent) or external events (e.g. change in conditions of the environment) can be defined in ABM using time-stamped empirical or simulated data (see Tang and Bennett (2010)). The illustration of an agent and its environment is depicted in Figure 3.1.

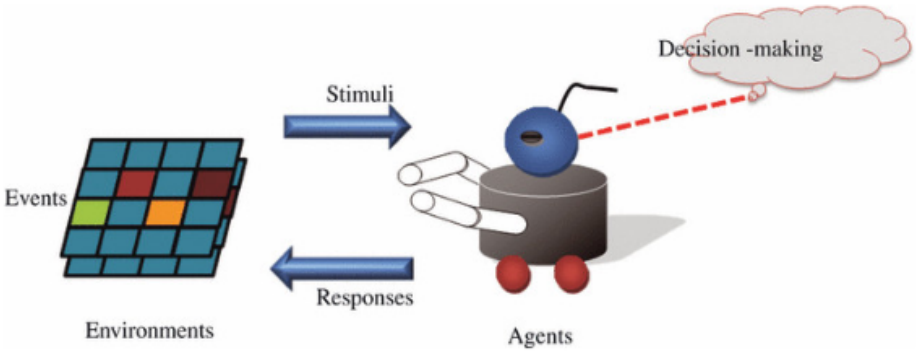


Figure 3.1: Illustration of agent and its environment taken from Tang and Bennett (2010).

3.2 Epidemiological agent-based modelling

ABMs are becoming increasingly used in infectious disease epidemiology as they have become an attractive alternative in epidemiological models in recent times. This is because they allow the simulation of heterogeneous communities subjected to more realistic

transmission scenarios and can incorporate complex and stochastic issues affecting diseases. The models can describe disease propagation dynamics as well as heterogeneous agent mixing and social networks. Thus, any kind of heterogeneity (such as heterogeneous intervention measures, host movement, multiple parasite variants) and stochasticity (such as inter-patient variability in duration of infection and virulence) can easily be modelled. Another strength of this modelling approach is that it enables the simulation of the interaction of individuals and vectors in the domain of interest, which can in itself be heterogeneous, and stores information about each agent. Protective efficacy of immunity varies between individual humans (see Hunter et al. (2017)), so it may be misleading to ignore individual variations in biologically important parameters while this can be efficiently addressed by ABM simulations. The flexibility of agent-based approaches in modifying model attributes to reflect local individual features and geographical factors, allows the construction of models that can handle realistic questions relating to disease control in specific local contexts.

Disease models for agent-based models are divided into two categories: transmission between hosts and progression within hosts. When a susceptible agent comes into contact with an infective agent, between host transmission occurs, and the between host transmission component of a disease model depicts how a disease is transmitted when this happens (see Hunter et al. (2017)). When an agent becomes infected, the inside host progression component of a disease model simulates how they move between infection stages (for example exposed, infective and recovered). Both aspects of the disease model are essential for accurately modelling the spread of a disease (see Figure 3.2).

The transmission dynamics play an important role in how the disease spreads among people (see Linard et al. (2009)). A disease can be transmitted from person to person, from food or drinking water to humans, or between hosts of different species, such as mosquitos and humans. A probability distribution is used to determine transmission when an agent comes in contact with an infected agent or contaminated food or drink.

The parameters of the modelled population have a significant impact on how a disease spreads across hosts. A densely populated location will result in more agent contact and, as a result, a higher risk of infection. The disease spread is also influenced by the social networks of the agents. In public places and gatherings, such as churches and schools, the majority of infections occur. The higher the social networks of agents in a model, the higher the rate of spread of infection (see Perez and Dragicevic (2009)).

The behaviour of agents can also affect the rate of transmission of the disease between hosts. For example, if the agents escape in the face of an outbreak or a potential outbreak, they may transmit the disease at a faster rate than if they stayed at home in isolation. Agents who opt to isolate themselves after being ill lower the number of contacts they establish and consequently the number of agents to whom the disease spreads (see Dunham (2005)). Agents who engage in preventative behaviours such as getting vaccinated or taking medicine, such as flu prophylaxis, minimise their risk of contracting an infection (see Mao (2014)). Disease progression within a host has fewer external factors than disease transmission between hosts. The Society's nature and other agents' behaviours have no bearing on how an agent progresses from exposed to infected or infected to recovered. The behaviour of infected agents, on the other hand, can affect the advancement. Vaccina-

tion and other preventative behaviours can lower the chances of an agent moving from the susceptible to infected state (see Mao (2014)). The factors that influence transmission or progression vary with the model, however they generally fall into the areas of progression dynamics, behaviours, and societal influences.
A considerable disadvantage of ABM is the higher computational burden, especially with increasing population size. The level of complexity in agent-based modelling has surpassed that of previous decades due to recent breakthroughs in computing capacity (see Taylor et al. (2014)). Another historical challenge with the ABM model is that of reproducibility. ABMs have usually been subject to criticism for being irreproducible despite its numerous advantages. However, in recent times, a standard approach of description called the 'Overview, Design concepts, Details (ODD) protocol' has been established with the primary objectives of making ABM model descriptions more understandable and complete (see Grimm et al. (2010)). We followed the recommended documentation protocol for ABMs in summarising the ABM done in Publication I, such that our model descriptions are more understandable and reproducible (see Appendix A).

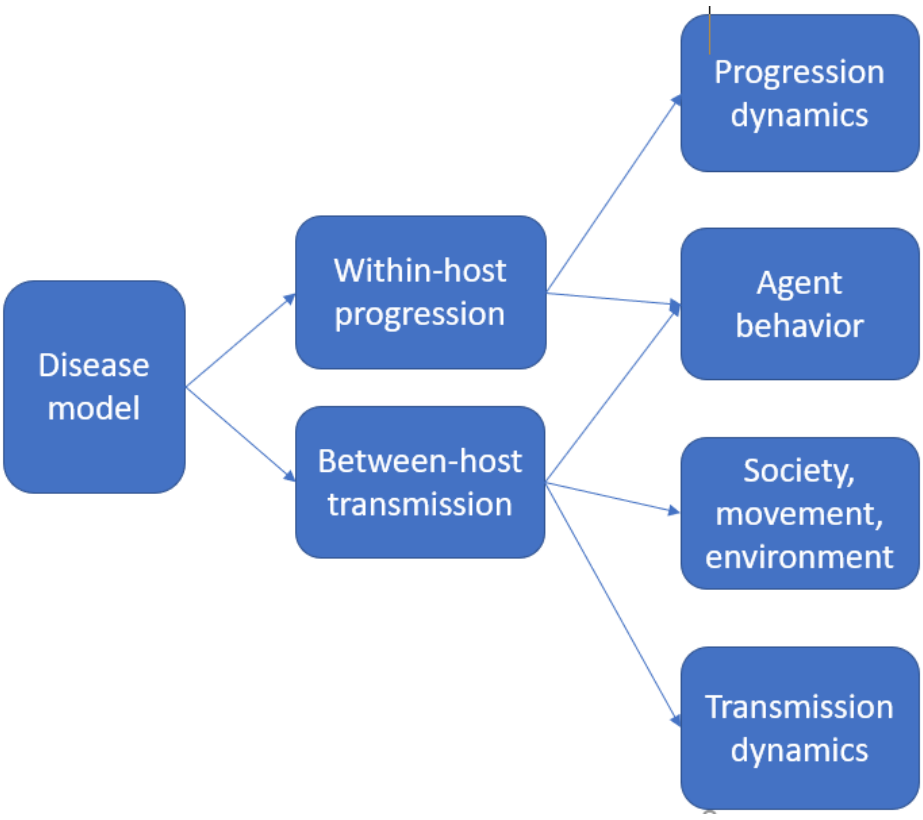


Figure 3.2: Illustration of the components of the disease model recreated from Hunter et al. (2017).

3.3 Hybrid model: from ABM to a continuous approach

Agent-based models are a technique of capturing the heterogeneity of a system and using it to drive the system's dynamics. The heterogeneity, on the other hand, can lead to larger models that require more computing power and time. Hybrid models allow you to keep such heterogeneity while decreasing the amount of computing power required to execute the model. The hybrid model enables the further scaling and simulation of a larger population while maintaining a heterogeneous population (see Hunter et al. (2020); Binder et al. (2012); Bobashev et al. (2007)). However, it is necessary to ascertain when the models can be shifted between compartmental and agent-based paradigms during the modelling process; otherwise, the model may lose performance. While making the disease element of the model compartmental-based can save time and computing power, it overlooks individual agent activities as well as the role of contacts and different contact patterns between agents in disease spread. Switching between agent-based and compartmental-based models allows for contact patterns to help drive the infectious disease spread in the early stages of the outbreak, but saves time once the outbreak has grown large enough for a few individual movements and interactions to no longer have as much of an impact because there are enough other agents infected.

In Publication I, a hybrid model was used to address the well-known issue of continuous modelling in epidemiological applications, which is the absence of experimental data required for model calibration. The hybrid method aids the transition from ABM to continuous modelling. The ABM simulations serves as a 'computational laboratory,' allowing data to be generated that reflects the impact of numerous complex factors on malaria prevalence. The outputs of the ABM are then integrated to basic dynamic transmission models to enable public health predictions. It should be noted that the author chose a low-level CUDA-based implementation of the model over one of the regularly used software frameworks specifically intended for ABM simulations because of the flexibility and efficiency of parameter identification afforded by the MCMC toolkit (see Haario et al. (2006, 2001)).

4 Regression models

In analytical epidemiology (see Silman et al. (2018); Dicker et al. (2006)), regression modelling is one of the most essential statistical approaches. Using regression models, the impact of one (simple regression) or several (multiple regression) explanatory variables (e.g. risk factors, exposures, subject characteristics) on a response variable (e.g., mortality, occurrence of a disease) can be examined (see Suárez et al. (2017)). Given the p predictor variables and N observations, the multiple linear regression model (MLRM), which is an extension of the simple linear regression model (SLRM), is given as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i; \text{ for } i = 1, \ldots, N. \tag{4.1}$$

From the above equation, Y_i is the response variable, X_{ij} denotes the explanatory variables, the slope β_j and the intercept β_0 are unknown constants, and e_i is the error term. The MLRM may include interaction terms generated by the product of explanatory variables. MLRM is a flexible and powerful method that can handle a wide range of data formats. The nature of the regression model depends on the type of data available. It is possible to analyse only cross-sectional data or time series data or both, with regression methods.

4.1 Regression for panel data

Individual data, both between persons and across time, can be collected in a panel data. Panel data are cross-sectional as well as time-series in structure. Panel data comprise N individuals observed over T time periods. They can either be balanced (if all individuals are observed in all the time periods) or unbalanced (if individuals are not observed in all the time periods) (see Nguyen (2022); Sheytanova (2015)). Examples of panel data include an aggregated weekly panel data of Dengue incidence level and meteorological factors (rainfall, average temperature, humidity) from 2009 to 2017, for 5 districts in Jakarta, Indonesia (see Publication II) and data of COVID-19 incidence and meteorological factors (average temperature, relative humidity) from March-December 2020, for the 16 states in Germany (see Publication III).
Panel data regression models describe the individual behaviour both across individuals and across time. In Frees et al. (2004), the considerable benefits of using panel data are extensively discussed. The three types of panel data regression models include: the pooled model, the fixed effects model and the random effects model (see Nguyen (2022); Sheytanova (2015)).

4.1.1 Pooled model

The pooled model does not differ from the simple regression model as they share the same assumptions. Thus, panel information is not used in the pooled model. This kind of panel regression model is employed based on the assumption that the individuals behave in the same way, with homoscedasticity and no autocorrelation. The pooled model sets constant

coefficients, which is the typical assumption for cross-sectional analysis, and thus does not consider individual specific effects, and is defined as

$$y_{it} = \alpha + x'_{it}\beta + e_{it} \text{ for } t = 1, \ldots, T \text{ and } i = 1, \ldots N. \tag{4.2}$$

This is the most restrictive panel data model and is rarely employed in the literature. When the fixed effect is inappropriate, however, the pooled model should be employed. If fixed effects were employed instead, the pooled ordinary least squares (OLS) estimates would be inconsistent.

4.1.2 Individual-specific effects model

We assume that there is unobserved heterogeneity across persons represented by α_i, such as unobserved features or activities of a German state that influence the frequency of new COVID-19 instances in that state. The fundamental question is whether the individual-specific effects α_i and the regressors x are correlated. We have a fixed effects model if they are correlated. The random effects model is used if they are not correlated.

Fixed effect (FE) model The FE model allows the regressors x to be correlated with the individual-specific effects α_i. Intercepts are included as α_i. The intercept term are distinct for each individual, but the slope term is the same for all individuals, defined as

$$y_{it} = \alpha_i + x'_{it}\beta + e_{it} \text{ for } t = 1, \ldots, T \text{ and } i = 1, \ldots N. \tag{4.3}$$

Individual specific effects are recovered as the leftover variability of the dependent variable that cannot be explained by the regressors

$$\hat{\alpha_i} = \bar{y_i} - \bar{x_i}\hat{\beta}_i. \tag{4.4}$$

The FE model estimates are always consistent, but inefficient when compared to random effects model estimates.

Random effect (RE) model Individual-specific effects α_i are assumed to be distributed independently of regressors in the RE model. In the error term we include α_i. Individuals have the same slope parameter and a composite error term $\varepsilon_{it} = \alpha_i + e_{it}$, expressed as

$$y_{it} = x'_{it}\beta + (\alpha_i + e_{it}). \tag{4.5}$$

Here, the interclass correlation of the error term is represented as $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$. It is the fraction of the error variance owing to the individual-specific effects. This term approaches 1 when the individual effects σ_α^2 dominate the idiosyncratic error σ_e^2. The random effects model provides the best linear unbiased results when utilised correctly.

4.2 Deciding which model is appropriate: panel regression

Choosing between the Pooled OLS and FE/RE model The Breusch-Pagan Lagrange Multiplier (LM) test (see Sheytanova (2015)), which is based on the OLS residual, tests whether the random effect variance σ_e^2 or equivalently $cor(e_{it}, e_{is})$ is significantly different from zero (i.e. whether or not every individual should have the same intercept). The random effect model is chosen instead of the OLS model if the LM test is significant.

Choosing between the FE and RE model The Hausman test (see Sheytanova (2015)) checks whether the fixed and random effects estimators differ significantly. The associated test statistic can be computed only for the time-varying regressors. It tests whether the correlation between the error term of the RE model and the regressors $cov(\alpha_i, x_{it})$ is significantly different from zero. It follows a chi-square distribution with degrees of freedom equal to the number of time-varying regressors' parameters. The random effect model is chosen over the fixed effect model if the Hausman test is insignificant.
After conducting the relevant test, a random effect model with a clustering strategy was used in Publication II and III to perform the regression. Rather than looking at which ranges of weather components predict the incidence, the clustering-integrated model associated with optimal barriers aims to determine which ranges of the incidence are well predicted by the existing weather components. Our clustering is based on the stratification of incidence data into an arbitrary number of groups separated by barriers. In both Publications II and III, the meteorological data were also grouped in accordance with the incidence data clustering. This not only improves fitting by adding more explanatory factors, but it also identifies incidence clusters that the meteorological components are unable to predict.

4.3 Clustering-integrated regression

Clustering is performed on the response data while the associated explanatory factors are classified following the places of the response data levels. The clustering technique encourages the use of only certain hypothetical variables to predict a given response variable, especially when explanatory variables are limited. Thus, the main goal of clustering is to accurately assign explanatory factors in situations where they should never have predicted a specific response variable. For example, if a high level of COVID-19 incidence data could not be predicted using the available weather data, it is probable that other factors, such as super-spreading, played a role in the high levels. Thus, unlike the usual regression approach, the clustering-integrated regression seeks to determine which ranges of the response variable are well predicted by the available explanatory variables. By including additional explanatory variables, this method may also improve fitting (see West et al. (2006); Strand et al. (2011)). In fact, the major motivation for employing the clustering-integrated regression in Publication II, is to properly handle the substantial zero values in the incidence data. Instead of the usual treatment of dropping time points corresponding to zero response (incidence) levels, the zeros are classified under low incidence cluster, but are left unpredicted. This allows for utilising more data for fitting the model

as compared to the case of dropping the time points with zeros.
Considering Equation 4.1, the clustering idea is based on categorising the response data into M clusters $(\Omega_k)_{k=1}^{M}$ separated by barriers $\theta := (\theta_k)_{k=1}^{M-1}$. The clusters are given in closed forms as $\Omega_k = \{y : \max\{0, \theta_{k-1}\} \leq y < \min\{\theta_k, \max_{i,j} y_{ij}\}\}$. Giving the function $(Y;\theta) := (1_{\Omega_k} y_{ij})$, where 1_{Ω_k} is the characteristic function, which takes value 1 in case y_{ij} belongs to Ω_k or 0 otherwise. Let us denote $P \circ Q = (p_{ij} q_{ij})$ as the Hadamard product between two matrices and define $X_{i1}^k = X_{i1}^k(\theta) := (Y;\theta) \circ X_{i1}$, $X_{i2}^k = X_{i2}^k(\theta) := (Y;\theta) \circ X_{i2}, \ldots, X_{ip}^k = X_{ip}^k(\theta) := (Y;\theta) \circ X_{ip}$. If the pairing response cases belong to the associated cluster, the latter returns the original entries of the matrices $X_{i1}, X_{i2}, \ldots, X_{ip}$; otherwise, they return 0. It then holds that $\sum X_{ij}^k = X_{ij}$, under this composition. Classifying the response data into three clusters ($M = 3$) on the basis of practicality to call for a lower, middle and upper cluster, model (4.1) can be rewritten as

$$Y_i = \beta_0 + \sum_{n=1}^{3} \beta_1^n X_{i1}^{(n)} + \sum_{n=1}^{3} \beta_2^n X_{i2}^{(n)} + \ldots + \sum_{n=1}^{3} \beta_p^n X_{ip}^{(n)} + e_i; \text{ for } i = 1, \ldots, N. \quad (4.6)$$

In theory, the number of clusters specified does not have to be limited to a small number because better fitting would be achieved with more explanatory factors. When adopting a large number of clusters, however, concerns about complexity and practical interpretations may arise. For example, if $X_{i1}^{(2)}$ is removed as a result of insignificance, it essentially implies that X_{i1} fails to predict response cases in the range specified by the middle cluster Ω_2. This approach thus permits this single case to be "unexplained by X_{i1}".
Considering the fact that as X_{ij}^k change with the lower and upper barrier $\theta = (\theta_\text{l}, \theta_\text{u})$, the pooled estimator $\hat{\beta} = \hat{\beta}(\theta)$ also does the same, it is of interest to find the optimal barriers such that the squared error between the data $Y = (y_{ij})$ and the model $Y[\hat{\beta}](\theta)$ is as small as possible. This problem can be presented mathematically as

$$\min_{\theta} \quad \sum_{i,j} (y_{ij}[\hat{\beta}](\theta) - y_{ij})^2 \quad (4.7a)$$

$$\text{subject to} \quad \min_{i,j} y_{ij} \leq \theta_\text{l} \leq \theta_\text{u} \leq \max_{i,j} y_{ij}. \quad (4.7b)$$

The above problem can be solved by employing optimisation methods such as the brute-force or particle swarm optimisation methods.

4.4 Logistic regression

Multiple linear regression can be used to determine the relationship between a single continuous response variable and a number of continuous, dichotomous or categorical explanatory variables. Logistic regression is similar to multiple linear regression in many ways, but it is used when the response variable is dichotomous. The outcome could be dichotomous by conception (an individual is either a university graduate or not), or it could be a dichotomisation of a continuous or categorical variable, for instance, although blood pressure is measured on a continuous scale, people may be categorised as

having high blood pressure or not for the purposes of analysis. In logistic regression, response variables are commonly coded as 0-1, with 0 indicating the absence of a feature and 1 indicating its presence. It uses the log function to predict the probabilities of outcomes. It converts the outcome into a categorical value using an activation function (sigmoid/logistic function). The logistic function is used to characterise growth rates that follow an S-shaped pattern

$$y = \frac{1}{1 + e^{-z}}. \tag{4.8}$$

This function first grows exponentially, gradually slows down and levels off as $|z| \to \infty$. As a result, if the value of z approaches positive infinity, the expected value of y becomes 1; if it approaches negative infinity, the predicted value of y becomes 0. If the sigmoid function's result is greater than 0.5, we may categorise the label as positive class or class 1, and if it is less than 0.5, we can classify it as negative class or label as class 0.
In Publication I, the effect of a spatial repellent is simulated by using an accept/reject technique using a logistic equation to determine the probability of rejection. Furthermore, the ABM simulation data show that, with respect to coverage, contact rates behave like logistic functions, with a specific threshold coverage required for the contact rate to start reducing, hence the logistic functions is applied.
Other variants of regression, in general, are better suited to specific types of data or to express specific relationships among the data. For example, the ABM simulation data in Publication I demonstrated a nonlinear, quadratic relationship between death rate and long-lasting insecticidal net (LLIN) coverage, but no significant dependence on household size. As a result, the death rates are only fitted with a second degree polynomial with respect to coverage.

4.5 Interpretation of coefficients in MLRM

When the explanatory variables of the MLRM are not correlated, the β_i coefficient represents the expected change in the variable y per unit change in the independent variable x_i, when all other explanatory variables are held constant. If β_i is positive (negative), the expected value of the response variable increases(decreases) by the number of units specified by the coefficient. One common issue in MLRM is that explanatory variables are sometimes correlated. Thus, it is advisable to assess their correlations, evaluate the changes in the regression coefficients as other variables are added or removed from the model and use collinearity diagnostics to detect multicollinearity (see Jewell (2003)).
Furthermore, interpreting parameter estimates in a linear regression when some variables are transformed is not always straightforward. After the transformation of variables, the marginal effects forthwith do not represent the true marginal effects. Three notions mitigate the problems of having variable true marginal effects (see Cameron and Trivedi (2010); Leeper (2017)):

- Marginal effects at representative values (MERs)
- Marginal effects at means (MEMs)

- Average marginal effects (AMEs)

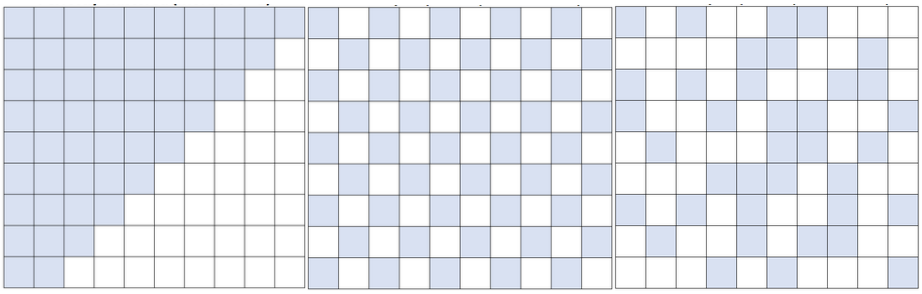
MERs compute the marginal effect of each variable at a theoretically interesting combination of X values. MEMs assess the marginal effects of each variable at the means of the explanatory variables. AMEs compute marginal effects at each observed value of X and then average across the resulting effect estimates. AMEs are especially better because unlike MEMs, AMEs generate a single quantity summary that represents the full distribution of X rather than an arbitrary prediction. Together with MERs, AMEs have the potential to reveal a significant amount of information regarding the influence of each explanatory variable on the response variable. AMEs can capture variability better than MEMs because they average across the variation in the fitted outcomes. AMEs provide a summary method that respects the distribution of the original data and does not depend on summarising a considerably unobserved or unobservable X value, while MERs provide a way to interpret model estimates at theoretically interesting combinations of predictor values (as in MEMs).
In Publication II, the incidence and rainfall datasets have exponential distributions, while the temperature and humidity datasets have normal distributions. We normalise the incidence and rainfall datasets for greater compatibility with dummy or coded variables. We obtain the average marginal effect in the context of the marginal effect of rainfall on dengue incidence.

5 Spatial autocorrelation

The study of spatial autocorrelation is an integral aspect of the research in Publications II, III and IV. The term spatial autocorrelation means the existence of systematic spatial variation in a mapped variable. In Publication II, spatial analyses were investigated by utilising Moran's autocorrelation coefficient and a LISA (Geary C), to explore geographical clustering in the incidences and to identify high-risk districts in Jakarta. In Publication III, both global Moran's I and global Geary's C were used to extract the spatial autocorrelation between 16 German States, using the 'queen case' contiguity based spatial weights. Furthermore, the results of the spatial auto-correlation provide the allocation of the 16 German states in the four quadrants from Moran's scatter plot as well as proper policy addressing travel restrictions. In Publication IV, spatial auto-correlation analysis using the global Moran's index (using distance-based spatial weights) and Moran's scatter is presented to aid in the modulation of hierarchical-based priority for health care capacity and interventions (such as possible vaccination) as well as finding a route for the corresponding deployment and landmarks for appropriate border controls in Sri Lanka. In general, the practical point of view of studying spatial autocorrelation for infectious disease incidence data is to understand how to curtail the disease. The methods employed for studying the spatial autocorrelation in the aforementioned publications are explained in this chapter.

5.1 Spatial patterns

Positive spatial autocorrelation occurs when geographically close values of a variable tend to be similar on a map (i.e. high values tend to be situated near high values, low values are near low values). The map shows negative spatial autocorrelation where nearby observations tend to have remarkably different values. Thus, the spatial pattern is usually considered as lying somewhere between three extremes: locally clustered, randomly distributed and locally dispersed. Locally clustered refers to a condition in which neighbouring states have similar levels of, say, daily new cases of disease. Locally dispersed is then the inverse spatial dependency where neighbouring states are not similar. Something in the middle is then seen as random. A chessboard can be used to understand how these spatial patterns are represented (see Figure 5.1). The spatial pattern is absolutely locally dispersed if the spatial structure of daily cases in all states resembles a chessboard. If all of the black cells had gathered in one location, the spatial pattern would be fully locally clustered. The random spatial pattern is then recognised by the way the white and black cells are randomly located on the board. There are several statistical methods for identifying the presence of spatial autocorrelation (see Diniz-Filho et al. (2003)) such as the Global Moran's I and the Geary's C.



(a) (b) (c)

Figure 5.1: Spatial auto-correlation patterns (a) perfectly clustered (b) perfectly dispersed (c) random (see Bobbitt (2021)).

5.2 Spatial weights

The spatial weight is a main ingredient in spatial auto-correlation. It is a measure of the spatial relationships amongst the features in a given dataset. The matrix that expresses interconnectivity between spatial units i and j is called the spatial weight matrix and takes the form

$$W = (w_{i,j})_{\substack{i=1,\dots,n \\ j=1,\dots,n}} = \begin{bmatrix} 0 & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & 0 & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \dots & 0 \end{bmatrix}, \tag{5.1}$$

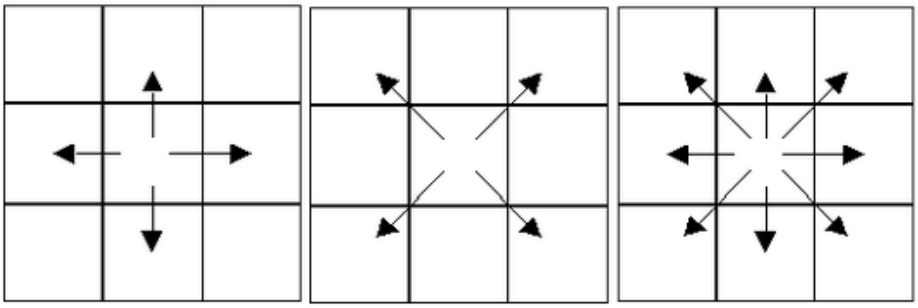
where diagonal elements take the value of 0. These spatial relationships are conceptualised based on the nature of the data. Considering the spatial weights matrix imposes a pattern on the data, it is necessary to choose a conceptualisation that best portrays how features actually interact with each other, while also thinking about what it is you are trying to measure. Spatial weights can be contiguity-based, distance-based or even population-based (see Song and Kulldorff (2005); Griffith (2020)).

5.2.1 Contiguity-based weights

Contiguity refers to the presence of a non-zero length shared border between two spatial units. In analogy to the moves allowed for the corresponding pieces on a chess board, these neighbourhood relations are described as rooks case, bishops case, or queens (kings) case (see Figure 5.2. The rook criterion defines neighbours as two spatial units that share a common edge. The Bishop criterion solely evaluates the diagonals of the relationships. The queen criterion is a little broader, defining neighbours as geographical units that share a common edge or vertex. As a result, the queen criterion's number of neighbours will always be at least as large as the rook and bishop criteria. Using regular grids (square polygons) to quantify these differences, the rook and bishop criterion will produce four neighbours, whereas the queen criterion will yield eight (see Sawada (2001); Kang et al. (2014)). The contiguity-based weights is typically defined based on nearby neighbours

and is written as

$$w_{ij} = \begin{cases} 1, & \text{if i and j are adjacent neighbours} \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$



(a) (b) (c)

Figure 5.2: Contiguity neighbourhood relations (a) rook case (b) bishop case (c) queen case (see Sawada (2001)).

5.2.2 Distance-based weights

In distance-based proximity spatial weight measure, central locations play an important role. Hence, a definition of 'center' is needed to determine the distances, for instance a city center, a main junction, a main administrative/commercial building and a transport hub. The distanced based matrix can be computed with different types of functions such as power function type, exponential type, uniform type and k-nearest neighbour type (see Kondo (2018)). Using an instance of the power functional type, the distance-based weights take the form

$$w_{ij} = \begin{cases} \frac{d_{ij}^{-\delta}}{\sum_{j=1}^{n} d_{ij}^{-\delta}}, & \text{if } d_{ij} < d,\ i \neq j \\ 0, & \text{otherwise,} \end{cases} \tag{5.3}$$

where the distance decay parameter $\delta > 0$ scales the influence of the distance, d is a threshold distance which cuts the inessential interconnectivity and d_{ij} is the distance between spatial units indexed by i and j.

5.2.3 Distance-band weights

In this case, i and j are regarded as neighbours whenever j falls within a critical distance band from i, expressed as

$$w_{ij} = \begin{cases} 1, & \text{when } d_{ij} \leq \delta \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

where δ denote a predefined critical distance threshold. To avoid isolates such as islands, resulting from overly stringent critical distance, the distance should be chosen in a way that each location has at least one neighbour. Such a distance measure satisfies the max-min criterion, (i.e. taking the largest of the nearest neighbour distances).

5.3 Moran's I

Using a row standardised spatial weight, the Moran's I for spatial autocorrelation is defined as

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{5.5}$$

where n denotes the number of features (regions), z_i represents the deviation of an attribute for feature (region) i from its mean $(x_i - \bar{x})$, w_{ij} represents the spatial weight.
The Moran's I falls within the range $[1, -1]$ (usually when the spatial weight is row-standardised) (see Kondo (2018)). When Moran's I approaches zero, it suggests that the data is spread randomly in space. When Moran's I has a positive (negative) value, it means that there is positive (negative) spatial autocorrelation across the regions; that is, the regions adjacent to a region with a high (low) value also have a high (low) value. The null hypothesis of spatial randomisation (i.e. each value is equally likely to occur at any region) can be used to conduct hypothesis testing for spatial autocorrelation in order to arrive at a decision based on the values of the statistics. The statistic which asymptotically follows a standard normal distribution can be defined as

$$z(I) = \frac{I - E[I]}{\sqrt{Var[I]}}, \tag{5.6}$$

where $E[I] = -\frac{1}{n-1}$ and $Var[I] = E[I^2] - [E[I]]$ (see details in (see Sokal et al. (1998); Moran (1950)). A p-value and the associated z-score are computed in a standard way. A p-value less than a specified significance level α rejects the null hypothesis, implying that either a locally clustered (if z-score is positive) or locally dispersed (if z-score is negative) spatial pattern exists (see Sokal et al. (1998)).

5.4 Geary's C

Geary's C is defined as

$$C = \frac{n-1}{2\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right]} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{5.7}$$

The value of Geary's C ranges from 0 to some undefined number greater than 1 (see Geary (1954)). Values considerably lower than 1 indicate increasing positive spatial autocorrelation, values that are significantly higher than 1 show increasing negative spatial autocorrelation and values close to 1 illustrate no spatial autocorrelation (see De Jong et al. (1984)). The significance of Geary's C can be tested in a similar way to that for Moran's I.
Moran's I and Geary's C are inversely related. Moran's I, on the other hand, is a global measure of spatial autocorrelation, whereas Geary's C is more sensitive to local autocorrelation.

5.5 Moran scatter plot

One of the local indicators of spatial association (LISA) is the Moran Scatter plot (see Anselin (1995)). It determines if a location is a 'hot spot,' a 'cold spot,' or somewhere in between. The Moran scatterplot's interpretation is based on the fact that many global association statistics are matrix form

$$\hat{\beta} = \frac{z^T W z}{z^T z}, \tag{5.8}$$

where W is a matrix of known elements, and z is a vector of observations (in deviation from the mean). This form of statistic can be visualised as the slope of a linear regression of Wz on z defined as

$$Wz = \beta z + residuals, \tag{5.9}$$

where Wz represents the spatial lag of the variable z.The scatter plot is divided into four quadrants, each of which represents one of the four types of spatial associations (see Figure 5.3). The lower left (upper right) quadrants present spatial clustering of similar values, whereby low values are bordered by low values (high values are bordered by high values) The upper left (lower right) quadrants represent spatial association of contrasting values, whereby low values are bordered by high values (high values are bordered by low values). With Moran's I statistics, no distinction between these patterns of association is conceivable because both kinds of association in the former case result in a positive sign and both forms of association in the latter case result in a negative sign.
As an illustration, we made a Moran scatter plot of the 7-day incidence COVID-19 data per 100,000 population for the 16 states in Germany as shown in Figure 5.4. While the overall pattern of spatial association seems to be positive, as suggested by the slope of the regression line given by Moran's I, four observations indicate an association between dissimilar values: two in the upper left quadrant, and two in the lower right quadrant. This may indicate the existence of different systems of spatial association, which may be more of randomness as indicated by the calculated Moran index value of 0.0003.
From the practical point of view, being a hot spot or cold spot (positive autocorrelation) can rely on health-care capacity to reduce disease burdens without imposing further limitations on travel between neighbouring states, with the exception of individuals who cross the border between dispersed hot spots and cold spots. This is due to the fact that they do

not experience differing degrees of incidence in their spatial lags. A state with a high-low or low-high (negative autocorrelation) spatial pattern, on the other hand, needs additional caution while travelling to nearby states since the disease may spread (in the case of high-low) or be absorbed (in the case of low-high).

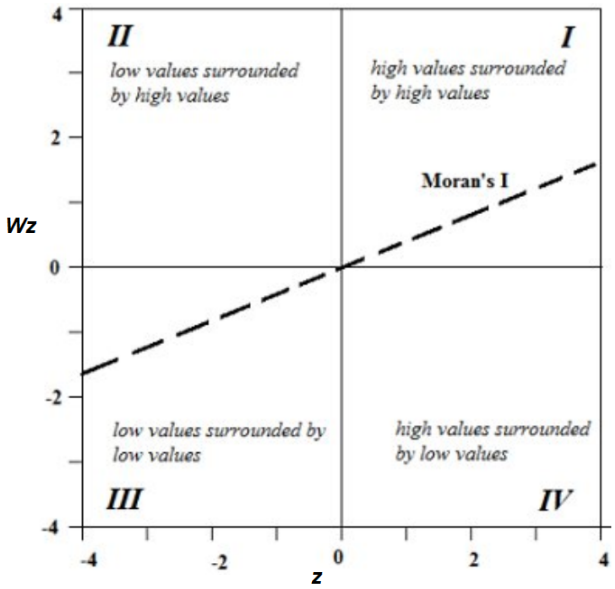


Figure 5.3: Moran Scatter plot definition (see Guţoiu and Pandelea (2016)).

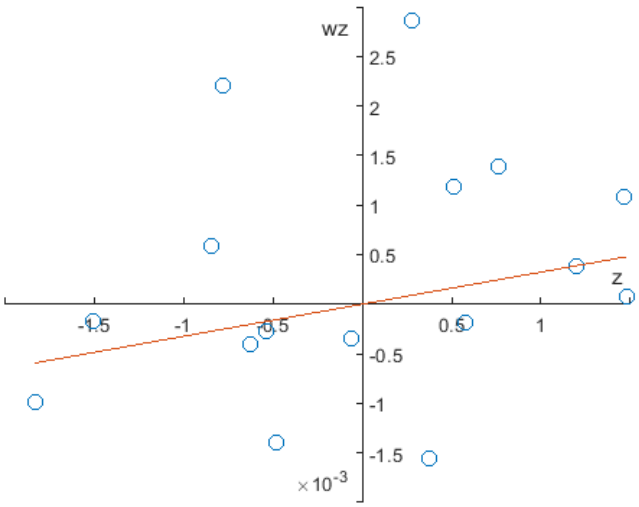


Figure 5.4: Moran Scatter plot of the 7-day incidence COVID-19 data per 100,000 population for the 16 states in Germany.

6 Uncertainty quantification

The correctness and reliability of the output of a model determine its usability. However, because all models are imperfect simplifications of reality, and because the acquired data used to calibrate the models is noisy, the output values are prone to error. As a result, we explain some causes of imprecision in model outputs while pinpointing the source inherent in our model, and thereby provide methods for assessing and expressing the uncertainties inherent in the outputs.

6.1 Sources of uncertainty

There could be a variety of explanations for the imprecision and uncertainty associated with model outputs. The following are the different sources of uncertainty:

- Error in the measurement that was taken. The variability of experimental measurements causes this error, which is also known as observation error.
- Inadequate representation of processes in a model (i.e. model structure) in comparison to the real system as well as approximations made by numerical simulation methods.
- Lack of understanding of the fundamental features (parameters) of a model that is needed for portraying reality.
- Specifying the values of parameters connected with the model structure with inaccuracy. Other parameter values would be obtained if the model calibration process was performed using different data sets. These parameters would result in various simulated model behaviours and, as a result, different model outputs.
- Variability arising from observed input and output values over an area and over time that differ from the model's geographical and temporal scale.
- Errors in the algorithm for solving the model.
- Uncertainty due to the natural randomness inherent in a process.

The above-mentioned sources of uncertainty, according to Sullivan (2015), are divided into two main categories, which are discussed below.

- **Epistemic uncertainty** It is also known as systematic uncertainty, and it is caused by some things that are known in theory but not in practice. These sources of uncertainty can be reduced and assessed by gaining a better understanding of the system under investigation.
- **Aleatory variability** It is also known as statistical uncertainty, and refers to the intrinsic randomness in a process which commonly originates from the unknowns that change each time the same experiment is run. Quantifying aleatoric uncertainty can be reasonably simple, e.g. by utilising Monte Carlo techniques to estimate the mean and standard deviation of model outputs.

In general, owing to the fact that the future is, for the most part, exceedingly unclear, there might be significant uncertainty in the model's results. The models take into account what we know with reasonable certainty about the future, such as vaccination rollout plans. Many factors, however, are unknown, such as whether a new variation will develop and what traits it will have. Thus, the further into the future the models are projected, the higher the uncertainty. As a result, each model output (or set of model outputs) must be accompanied by a measure of uncertainty. This uncertainty interval presents the range of values where the observed results are most likely to fall. That is not to say that the actual outcomes data may not be outside the interval; it merely means that it will be less likely in such scenarios. The resulting range of possibilities can help decision makers decide whether to adopt a certain course of action by selecting alternative combinations of model parameters and simulating different trajectories. It can also be used to create a range of thorough contingency plans to deal with various scenarios.

6.2 Uncertainty quantification with MCMC

Uncertainty quantification is usually conducted using the Markov Chain Monte Carlo (MCMC) approach. This method is based on Bayesian inference and may be used to determine the reliability of parameter estimations as well as provide precise parameter confidence quantification (see Browning et al. (2020)). Thus, by providing distributions of parameter values that conform with the available data, this technique gives accurate estimates of model parameters, with their uncertainties. The MCMC method enables the generation of a possible combination of parameter values which fit the data within a specified tolerance with respect to the measurement noise. This method discards the notion of accepting a single best fit to the data in favour of identifying all the regions of the parameter space that agree with the observations. The MCMC method involves selecting candidate parameter values from a predetermined proposal distribution, and then accepting or rejecting these candidate points according to a level of probability, which takes into account the model output's similarity to the data Haario et al. (2006). Adaptive MCMC is used in particular in the study in Publication I since it may not be possible to clearly identify a well-working proposal distribution at first. The Adaptive MCMC algorithm is an upgraded variant of the Metropolis algorithm that uses information from previously sampled points to adjust the proposal covariance during the MCMC run (see Haario et al. (2006, 2001)). To evaluate the fit with the data, the cost function is utilised, which returns the sum of squared differences between the observations and the model outputs while accounting for measurement error variance, given as

$$\mathbb{S}_{sum} = \sum_{i=1}^{N_r} (Y_i - \hat{Y}_i)^2, \tag{6.1}$$

where N_r is the number of replies for which the sum of squares is computed, Y_i denotes the observed data, and $\hat{Y}_i$ denotes the simulated model outputs. The structure of the posterior distribution will reveal whether the observables constrain the model parameters in a unique way. Making a plot of the parameter chain's autocorrelation functions, from

which one can see the extent to which samples that are k steps away correlate with each other, is a useful way to visualise how well the chain is mixing (see Haario et al. (2006)). We would expect successive points to correlate more with each other than points further apart in MCMC because the upcoming points are dependent on the previous points. If autocorrelation is still high for greater values of k, it indicates inadequate mixing. The model parameters are identifiable if the parameter values that are in the best agreement with the data are confined to a small region of the parameter space. The display of parameter chains can also reveal whether or not a parameter is correctly identified.

Publication I studied the uncertainty quantification of model outputs, which involves the concept of randomness as a source of uncertainty. The hut experiment data from Kitau et al. (2012) employed in this publication includes the percentages of exited, blood-fed and dead mosquitoes for each tested chemical. The results were obtained for two species of mosquito: *An. gambiae* and *An. arabiensis*. To minimise the uncertainty due to variances in an individual host's attractiveness, the results in each case were averaged over six repetitions of the experiment. The sum of squares cost function is used to compare the simulated and experimental data, see Equation 6.2. Given the repeated measurements, the likelihood is considered to be Gaussian. Considering that no prior estimate of the parameters is available, all of the cases investigated use uninformative uniform priors for sampling. To approximate the posterior distribution of the model parameters, the likelihood is sampled using adaptive MCMC from Haario et al. (2006, 2001). The model evaluations that produce values within the data noise level are represented by the sampled parameter sets. The sum of squared difference between the model outputs and the observations is parameterised as

$$\mathbb{S}_{sum} = \sum_{i=1}^{N_r} \frac{(Y_i - \hat{Y}_i)^2}{\sigma_i^2}. \tag{6.2}$$

The standard deviations were chosen to correspond to the confidence intervals in the data from Kitau et al. (2012). The number of measured responses, the exit, fed and mortality rates for each of the two mosquito species considered are given by $N_r = 6$. The data from Kitau et al. (2012) is relative and, therefore, the size of the mosquito swarm employed in the simulations can be determined by numerical efficiency. All of the findings are averaged over several iterations because the model is stochastic. The combination of 6 repetitions and a swarm of 600 mosquitoes produced a rather low variance with minimum CPU time. On a CPU core-i7 2500K and GPU GetForce GTX TITAN, the total wall-clock time for one evaluation of the cost function employing parallel GPU calculations was roughly 2 seconds.

As an illustration, the results of the model calibration featuring uncertainty quantification with MCMC, for both mosquito species, when confronted with LLIN treated with Alphacypermethrin kit is presented in Figures 6.1 and 6.2. It can be seen that the mortality rate for *An. gambiae* with Alphacypermethrin produced with the sampled values of parameters displays a high variation from the sampled mean of the model output, which in some cases is even comparable with that of *An. arabiensis*. However, the mean values of the model outputs are all within the confidence bounds. In general, the model calibration

results overall show a good fit to the response measurements, and the variability of the simulated model outputs match with the error bars of the measurement. The pair-wise correlation plots show that the sampled parameters are more or less well identified. The details of the model calibration are given in Appendix A.

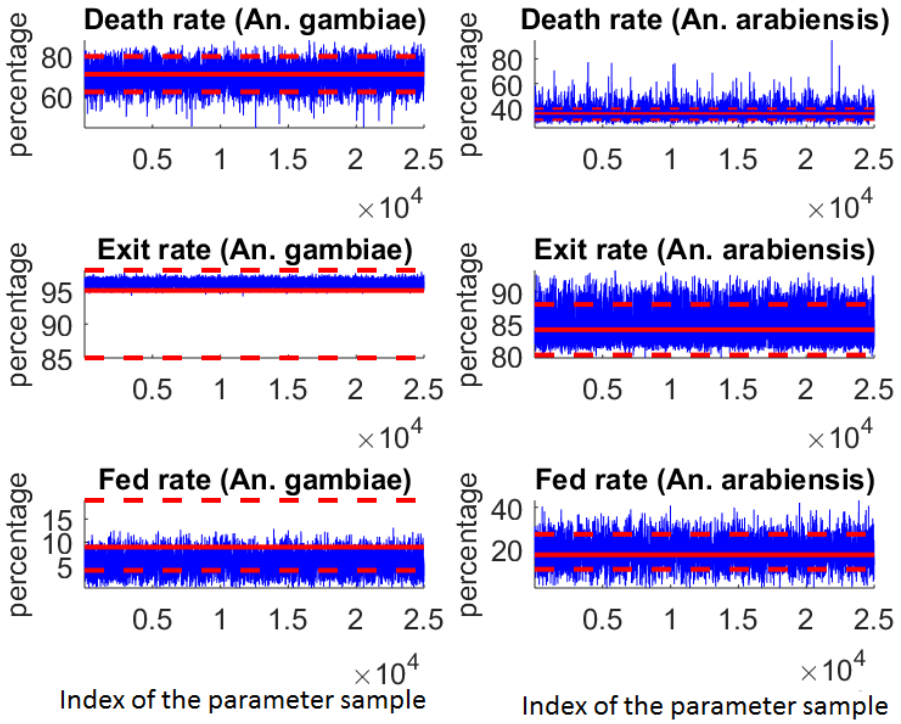
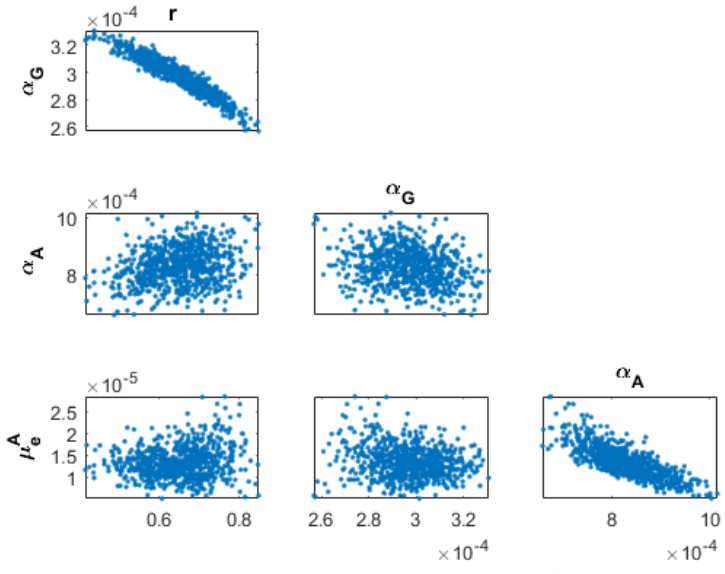


Figure 6.1: Results of MCMC model calibration with the experimental data reported in Kitau et al. (2012), where LLIN is impregnated with Alphacypermethrin. Model outputs obtained using the posterior distributions of parameters (blue trace lines) versus the data (mean values in red solid and 95% confidence intervals in red dashed lines).



(a)

Figure 6.2: Pairwise marginal posterior distributions of model parameters resulting from the MCMC model calibration with the experimental data reported in Kitau et al. (2012), where the LLIN is impregnated with Alphacypermethrin.

7 Summary of the publications

7.1 Summary of Publication I

The aim of this paper is to offer an approach for combining *in situ* field data with the parameters of malaria transmission models. This was motivated based on the common pitfall in modelling, which is to obtain data that can be directly used for model calibration. There is usually a gap between what can be measured and the conceptual factors used in epidemiological models. For instance, we measure the impact of LLIN coverage on malaria prevalence impacting factors such as mosquito-to-human contact rate and mosquito mortality rate (see Kitau et al. (2012)), whereas the conventional epidemiological models consider the fraction of people carrying the infection. Thus, the study seeks to make a bridge, enabling data to be obtained in a way that it can be directly utilised for model calibration. The ABM modelling is firstly done for a single host in a hut, then for a household with some people sleeping under the same roof. The household model is then extended to community-level scenarios, allowing simulations of variation in mosquito-to-human contact rates due to partial net coverage or varying household sizes. The ABM simulations operate as a 'computational laboratory,' allowing data to be generated that reflects the impact of a variety of complex factors. The ABM results can be utilised after repeated simulations, as synthetic data to create regression models for the factors considered (i.e. LLIN coverage, household size and changes in mosquito behaviour caused by the malaria parasite). The agent-based model simulations are run over a one-night 'snapshot' time period. By substituting the values fitted by the response surfaces as the key coefficients of the traditional Ross model, the results are extended to continuous time. As a result, the effects of intervention measures or socioeconomic factors can be replicated over extended time periods and up to a steady state. This enables the EIR values to be estimated in a wide range of transmission scenarios. The work-flow used in this study is presented as a schematic representation in Figure 7.1.

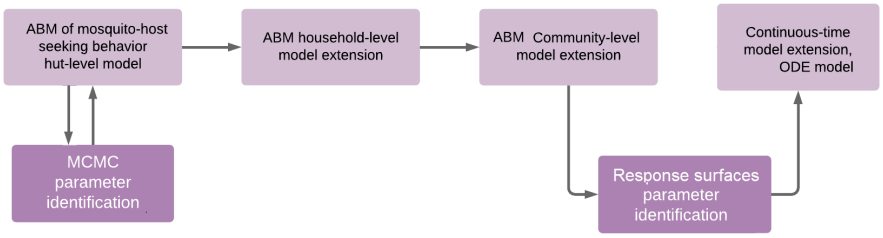


Figure 7.1: A diagram describing the transition from ABM of host-seeking behaviour of mosquito to continuous modelling. The technique is carried out separately for each mosquito species and each chemical under investigation.

While the current study should be viewed as a proof of concept based on a single set of field data, it does draw some important conclusions. For smaller households, a lower LLIN coverage is sufficient to achieve a certain reduction in the biting rate. When as-

suming behavioural change, the contact rates are higher, but with high LLIN coverage, the contact rates are about the same, indicating that the influence of changes in mosquito behaviour due to the presence of the Plasmodium parasite is insignificant. There is also a distinction between the mosquito species. The coverage necessary to achieve a similar reduction in the number of infectious bites is higher for *An. arabiensis* compared to *An. gambiae*, basically as a result of the lower death rate of *An. arabiensis*. When considering the alterations in behaviour, the death rate rises. Intuitively, this may be because of the increased attempts to feed on several hosts during the night resulting in the higher insecticide exposure for infected mosquitoes. Moreover, by incorporating various factors that influence the EIR and malaria incidence, such as reducing mosquito-human contact rates and increasing mortality through control efforts or socio-economic factors, the general transmission characteristics may be evaluated.
The current research can naturally be extended in a number of ways. Other mosquito species in addition to *An. gambiae* and *An. arabiensis* as well as other intervention measures other than LLINs. The mosquito density m is considered to be constant in this study, despite the fact that it varies periodically owing its dependence on rainfall and temperature. By calibrating the corresponding parameters to be site dependent, spatial features such as the local disposition of mosquito-breeding sites can be added. This allows the modelling to be scaled up to cover bigger geographic areas. In this study, all mosquito-human contacts with an infected mosquito are modelled to be equally infectious, even though some persons may have developed partial immunity to the parasite either by frequent exposure to the parasite or through vaccinations (see Filipe et al. (2007)). By characterising the hosts as a population of agents and making the transmission parameter b dependent on the individual immunity level, the impact of naturally acquired immunity can be incorporated into the model. Furthermore, the current research is limited to in-house biting circumstances at night. The model can, however, be extended to cover outdoor biting scenarios (see Sherrard-Smith et al. (2019)). All of these extensions are technically possible, but they require enough field data to allow for a reliable calibration of the underlying ABM model.

7.2 Summary of publication II

Dengue fever continues to ravage developing countries with climates favourable for mosquito breeding and moderate-to-weak health systems, according to WHO medical statistics. This paper presents a study of dengue incidence in relation to meteorological factors in Jakarta, Indonesia, from 2009 to 2017 to gain a better understanding on the influence of the latter on dengue incidence. A clustering-integrated multiple regression model was constructed for aggregated weekly panel data of incidence level and meteorological factors (average temperature, rainfall, humidity) for the case of Jakarta, Indonesia. The clustering was motivated by the zero-inflated problem (i.e. substantial zero values) in the incidence data and the limited availability of supporting data in Indonesia, while at the same time, requiring a model with proper fitting and strong predictability which is crucial in assisting decision-makers in the development of an early warning system. We choose a random effect model with the pooled estimator after sequential tests. Given that

changing the clustering barriers leads to varied modelling outcomes, the optimal barriers in terms of minimising the mean squared error is sought. The use of constant coefficients in the model for multiple districts results in a unified model in which the coefficients indicate the marginal impacts of climatic factors on the incidence level in a broad sense, i.e. district-independent. For computational efficiency, unlike earlier studies, the lag for the meteorological components linked with the maximum Spearman's correlation coefficients is of interest.

A one-step vector autoregression model using just incidence data is proposed as a framework for evaluating risk. The obtained risk measure can be used to investigate disease persistence as well as which districts fall into the low and high risk categories. In terms of geographical competition, we use Moran's and Geary's autocorrelation coefficients to perform global and local spatial autocorrelation analyses, respectively. The findings reveal how all of the districts either increase the incidence level simultaneously or show considerable discrepancies without providing clear evidence as to which district should be prioritised at any particular time. As a result, we use spatial periodicity analysis to establish how the maximum peak of incidences from all districts propagates across time. Using Fourier analysis, we can find that the greatest peak in the propagation resembles annual periodicity throughout all districts. The recurrence of the peak combined depicted by the risk measures may suggest the districts that need the most to the least attention during epidemics.

7.3 Summary of publication III

This study used panel COVID-19 incidence data from Germany and examined their association with meteorological data, which was motivated by an increase in morbidity throughout the autumn and winter seasons as well as considering that most respiratory diseases are seasonal (see Dowell and Ho (2004); Shi et al. (2020)). The average of daily average temperature and relative humidity from three representative meteorological stations in Germany from 31 January 2020 to 15 December 2020 is employed. Moran's I and Geary's C statistics were then thoroughly investigated in order to explore spatial autocorrelation and its practical consequences. The distinction between this study and earlier ones is that the temporal progression of the statistics is presented. Following that, this research proposed a random-effects model with a clustering technique. The overall concept is that the meteorological components may accurately predict the incidence ranges. This is in contrast to finding the ranges of meteorological components that can be used to predict the incidence. Our clustering is based on the stratification of incidence data into arbitrary number of groups separated by barriers. The temperature and relative humidity data were also grouped in accordance with the incidence data clustering. This not only improves fitting by adding more explanatory factors, but it also identifies incidence clusters that the meteorological components are unable to explain. To construct appropriate variables in the regression models, the lags from the cross-correlation between average temperature and relative humidity were obtained.

Case-specific auto-correlation backs up the model specification, indicating that lag-3 and lag-4 days incidence would not be significant predictors of current incidence. The global

spatial measures show random spatial patterns most of the time, with the exception of recent observations from 1 November to 15 December 2020, when there were either local clusters or dispersion. The distribution of hot spots and cold spots altered with time, according to Moran's scatter plot, which was utilised to reveal the local behaviour of the spatial pattern. The random geographical pattern justifies the model specification, where individual- or state-specific effects that would have provided constant weighting factors to specific states were removed.
The clustering-integrated model associated with optimal barriers demonstrates good fit with the data whereby weather components dominate lag incidence cases in the prediction. The fixed-effects estimator was the only seemingly consistent estimate that also dealt with the panel effect in this case. Every explanatory variable competes with the others in order to be a significant predictor in all models. As a result, the decision-maker is solely responsible for model selection and its implications. When a model is chosen *a priori*, marginal effects can provide guidance. When R^2 and BIC are essential, the clustering-integrated model with lag incidence cases and lag weather components is recommended. It was discovered that temperature and relative humidity have relatively small negative marginal effects on the cases in the lower cluster. However, temperature has a large positive marginal effect on the cases in the middle cluster but no marginal effect on the upper cluster. Relative humidity, on the other hand, has a large positive marginal effect on the upper cluster and no effect on the middle cluster. When weather takes precedence over lag incidence cases, the clustering-integrated model with only weather components is advised. Our findings support the cross-correlation study's findings that temperature has negative marginal effects on incidence, whereas relative humidity has positive marginal effects in all clusters. Temperature has the smallest marginal effect on the middle cluster, whereas relative humidity has the smallest marginal effect on the lower cluster. This suggests that temperature can only predict incidence cases during hot (summer) and cold (winter) seasons, where cases clearly differentiate each other from the data, and not during transitional seasons (such as spring and autumn). During the summer season, however, relative humidity is less likely to predict sinking cases. The modelling not only determines the degree of the prediction via marginal effects, but it also allows for precautionary actions in the event of impending weather.

7.4 Summary of publication IV

This study reassesses non-pharmaceutical interventions (NPIs) such as contact restrictions and testing campaigns againt COVID-19 using a spatio-temporal modelling. COVID-19 has disrupted life in many areas worldwide since its first outbreak. The study presents a mathematical framework for improving NPIs during the new normal before herd immunity is achieved. The method is designed to assist decision-makers in developing nations, where medical and economic resources to combat the disease are limited. This research not only prioritises health care capacity and NPIs among spatial units, but also maps out a more robust approach for them than incidence-driven techniques. To measure and group COVID-19 incidence in Sri Lanka, we use the global Moran's index and Moran's scatter. Prioritising high-risk locations or hotspots is influenced by the efficient use of health care

resources, especially in developing nations. As a result, the priority of intra-cluster NPIs remains the same within a cluster, but they are ordered in priority across clusters. This method is critical for developing countries, such as Sri Lanka, which have yet to benefit from a comprehensive spatial analysis of this scale. The clustering analysis can bear the locations for border restrictions, which in this case are those in the major inter-cluster mobility streams, in addition to prioritising and route. However, there is one drawback with these techniques: they are unable to numerically parameterise ongoing government decisions and, therefore, they are unable to convey how sensitive the incidence is to changes in those decisions.

Thus, a metapopulation model for Moran's clusters that addresses available panel COVID-19 incidence data from on Sri Lanka is proposed. The dynamic model was chosen over functional regression models because of the integrable mechanistic components that underpin COVID-19 infection and the lack of spatiotemporal data on confounding factors. Based on the lack of related field data, a complexity reduction is presented, resulting in a basic model that is logical enough to mediate contact limitations and testing campaigns. The possibility that the incidence will endure for a long period was studied. The model solution is compared to particular local equilibria, with the basic reproduction number and effective reproduction numbers playing a crucial role therein. The fitting of a metapopulation model will provide a proxy for not just approximate reproduction numbers, but also non-observable dynamics such as contact matrix and current government testing campaign decisions.

Finally, the bifurcation analysis is extended numerically utilising a path-following approach for the instance where the clusters are not strongly connected according to the fitting. Furthermore, the effectiveness of government decisions on contact restrictions and testing campaigns during the observations is evaluated using the maximal *average policy effect*, which measures the average number of individuals per 1,000,000 inhabitants who could have been saved from COVID-19 infection if better interventions had been implemented. Scenarios for cost-to-benefit analysis are also included.

Forward bifurcation for strongly connected networks among clusters was discovered around $\mathscr{R}_0 = 1$. A numerical study was carried out for the case where the network is not strongly connected, as determined by rounding small β-values (infection rates) to zero. Time-varying effective local reproduction numbers for the four clusters are also computed. Their appearance outperforms clueless cases data when it comes to pinpointing the time at which the current transmission is high ($\mathscr{R}_0 > 1$), suggesting immediate interventions.

One-parameter continuation of equilibria produces an intriguing result. From the analytical framework, the baseline direction of the continuum of endemic equilibria at $\mathscr{R}_0 = 1$ is the Perron vector of the next generation matrix. Owing to the fact that the network associated to the contact matrix β or the next generation matrix is not strongly connected (see Figure 7.2, Perron-Frobenius Theorem only ensures the nonnegativity of the components

of the Perron vector ψ_1. For this study, we obtain

$$\psi_1 \approx \begin{pmatrix} 0 \\ 0 \\ 0.9553 \\ 0.2957 \end{pmatrix}.$$

Two conclusions can be drawn from this Perron vector: (1) When $\mathscr{R}_0$ immediately exceeds 1, the clusters Q1 and Q2 stay disease-free; nevertheless, (2) the long-term number of active cases in the cluster Q3 increases to a greater extent than that in cluster Q4. However, reading the bifurcation diagrams backward in ω (contact restriction factor) and p (case detection ratio), these findings imply that Q1 and Q2 attain disease-free states faster than Q3 and Q4 under the reduction of p and ω from $p_{\text{Ref}} \approx 0.4698$ and $\omega_{\text{Ref}} = 1$, correspondingly (see Figure 5 of Publication IV for the bifurcation diagrams; p_{Ref} and ω_{Ref} are the reference values for computing the average policy effect as can be seen in Table 1 of Publication IV). According to the network in Figure 7.2, Q2 receives a little 'injection' from Q1 but returns with a huge injection to Q1, whereas Q2 has no essential self-injection. Q1 also injects Q3 and Q4 at comparable rates, coupled with a minor self-injection. In the meantime, Q3 receives a pretty large self-injection but avoids injecting Q4. Overall, if all the injection rates (the non-zero entries of the contact matrix β) are simultaneously reduced, it is debatable that Q1 and Q2 lose endemicity faster than Q3 and Q4. On the one hand, there comes a point where the self-injection in Q1 and hence the injection into Q2 is negligible, rendering Q2 non-reproductive. However, the insignificant injection from Q1 is balanced by self-injection in Q3, which withstands both Q3 and Q4 in the endemic states.

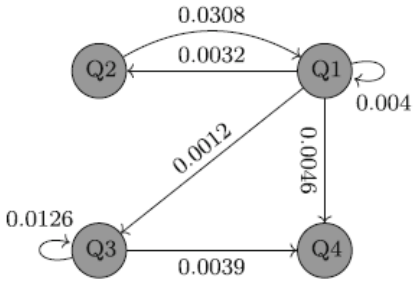


Figure 7.2: Network created with the contact matrix β. The arrow pointing from cluster Q_i to cluster Q_j implies that infected individuals in cluster Q_i cause infection in a susceptible individual from cluster Q_j upon contact at the associated infection rates.

However, it was also observed that disease-free equilibrium (DFE) may be found by lowering p and ω to values that are not too far away from p_{Ref} and ω_{Ref}, correspondingly. As a result, we argue that the government's initial measures were adequate during the observation period. Significant contact constraints, from the perspective of model transients, are both costly and not rewarding in terms of average policy effect (APE). This is demonstrated by APE's concave behaviour and the cost's convex behaviour against ω. As a

result, arbitrarily lowering p and ω makes little sense. Thus, scenarios where the optimal values of p and ω may minimise the cost under fixed magnitudes of APE were suggested. Even the optimal (p, ω)-values have a cost, as expected, since they move toward the third quadrant (defined based on the Moran scatter plot) by increasing APE values (see Figure 7b of Publication IV).

Finally, this research identifies certain areas in which improvements might be made. This include modifying some features in the original model to reflect more complexities and running a parameter identification study to discover possible parameter dependency and hence improve the model.

8 Discussion and conclusion

The main purpose of this research is to develop and present some modelling tools that can be used in answering some important questions in epidemiological studies. By definition, models are a simplified depiction of reality. Models are unable to account for every aspect of changes in the nature of the disease, and do not attempt to do so. Instead, they attempt to capture the key components. Because each disease is different, models must be selected based on the factors and variables in the epidemic. The complexity of models and the modelling approach vary based on a number of factors, including how much is known about disease epidemiology, the study's objective, and the amount and quality of data available. This work has presented many approaches to model the complexity of a system, ranging from simple to complex methods, as well as a blended approach.

Compartmental models have been employed in the formulation of control policies and outbreak responses given that they have proved to reflect the macro level dynamics of infectious disease epidemics. One of the important thresholds that can be easily learned from the compartmental model is if the disease will persist or die out in the population through the computation of $\mathscr{R}_0$. Despite their usefulness, a considerable disadvantage is that compartmental models are not sufficient for reproducing the real dynamics of some diseases and allow only a limited account of the complex process of disease dynamics. They make clearly artificial assumptions that seem to make them conceptually compelling, but may actually be inefficient. The ABM technique, unlike the continuous modelling paradigm, includes heterogeneous characteristics within the population of agents, which can be viewed as an endless number of compartments or states.

The ABMs have become an attractive alternative in epidemiological models in recent times because they allow the simulation of heterogeneous communities subjected to more realistic transmission scenarios and can incorporate complex and stochastic issues affecting diseases, which could have been treated as an infinite number of compartments in compartmental models. However, the computational burden, especially with increasing population size is a compelling disadvantage of the ABMs. Given the limitations of both kinds of models, it is possible to still capture the necessary heterogeneity while reducing the computing power required to run the model through the use of hybrid models. One of our works linked ABM results (which were simulated using various complex factors impacting malaria prevalence) to classical dynamic transmission models so as to enable public health predictions.

To answer the question of the nature of the relationship between the occurrence of a disease and the factors that are perceived to allow its spread, regression models come into play. In addition to the conventional way of employing regression models, which is to determine the range of available explanatory variables that can predict disease incidence, this research also evaluated the range of incidence that can be predicted by the available explanatory variable through the use of a clustering-integrated model. This approach does not only help to screen the incidence clusters where the available explanatory variables fail to predict, but also helps to improve fitting by providing more explanatory variables.

Furthermore, we study the spatial autocorrelation for infectious disease incidence data to understand how to curtail the disease. We adopt global Moran I index, the Geary' C

and Moran's scatter plot to evaluate the timely spatial pattern of disease incidence in a country and also to set the grouping. This study is particularly important for developing countries where prioritising high-risk locations or hotspots can help in the judicious utilisation of health care capacity. The novelty of the work is the combination of spatial autocorrelation analysis (for hierarchical prioritisation of non-pharmaceutical interventions (NPIs)) with a newly proposed metapopulation model that parameterises and reassesses non-pharmaceutical interventions.

The work also discusses the approach to quantify the uncertainty in model outputs given that models are imperfect simplifications of reality and the data used to calibrate them is noisy. In this study, the MCMC technique was used to generate a variety of parameter values that fit the data within a defined tolerance for measurement noise. This method was used to quantify the uncertainty of model outputs involving the notion of randomness as a source of uncertainty. Above all, the quality of the data that gets into a model determines its output quality. In general, the work models the importance of some determining elements that influence disease transmission in order to provide insight into how they might be considered when developing intervention strategies.

9 Appendix A: ODD protocol for the ABM model in Publication I

Here, the discussion of the ABM simulations following a recommended protocol 'ODD protocol' by Grimm et al. (2010) (see Section 3.2) is presented. The idea is to explain how the terms and concepts employed in the standard protocol applies to our ABM model, as adapted from the Appendix of Publication I.

9.1 Purpose and patterns

9.1.1 Purpose

The purpose of the model is to provide data reflecting the impact of various complex factors affecting malaria such as household size, LLIN coverage, and alterations in mosquito behaviour induced by malaria parasite, in the form that can be directly used by the continuous models of malaria. Thus, the ABM simulations are used as a 'computational laboratory' where data can be produced for regression analysis, so as to enable the calibration of the key parameters of classical malaria models. In considering these factors, the modelling of a single host in the hut is done, followed by the household level modelling, with multiple individuals sleeping under the same roof. The household model is then expanded to community-level scenarios, allowing simulations of heterogeneity in the mosquito-to-human contact rates due to partial net coverage or varied household sizes.

9.1.2 Pattern

The hut-level simulations are data driven, as they aim to reproduce the patterns of the data employed from Kitau et al. (2012). In the community-level case, since there is no data to calibrate the simulations, the literature values are employed and a sensitivity analysis is conducted based on these values to ascertain how the assumed parameter values impact the overall outcomes. See Table 9.2.

9.2 Entities, state variables and scales

9.2.1 Entities

The entities in the model include humans, mosquitoes, nets and chemicals. Humans are modelled as individual agents, attributed with the state of infection, the use of insecticidal nets and spatial position. Humans do not carry out any actions and their features are constant in time, with a snapshot of one night. This is because the main focus of the model is to control the mosquito population. Two female mosquito species were employed: *An. gambiae* and *An. Arabiensis*. Given that only *female* mosquitoes transmit the parasite during blood-feeding, and mating is outside the scope of this project, the ABM simulations do not include the *male* mosquitoes. The difference of the species is attributed by their host-seeking behaviours (anthropophilic or opportunistic preferences of mosquitoes)

when confronted with the insecticidal nets. The mosquitoes can either be infected or uninfected. The infected mosquitoes differ from the uninfected ones in terms of their biting habit. The number of infectious mosquitoes is constant for a single experiment. This is because it takes a period of 10 to 12 days for parasites to reach a stage whereby they are ready for transmission. However, the ABM simulation in this study is only for a night.
Insecticidal nets of 1.5m width, treated with four different chemicals: Carbonsulfan, Iconmax, Alphacypermethrin and Deltamethrin, are simulated. The difference between these chemicals is represented by their impact (in terms of contact irritancy, excito-repellency and poisoning) on each of the mosquito species under study. Considering that some of the bed nets used in rural communities are typically holed in practice, purposely holed nets are widely used in hut trials (see Okumu et al. (2013)). Therefore, broken nets are simulated in such a way that the likelihood of mosquito penetration is non-zero. For the hut-level case, the human agent is always covered with the net. However, at the household and community levels, the number of protected humans can vary from 20 to 100% and remains constant throughout the simulation. Furthermore, a hut barrier (walls) is simulated for each of the huts. In the hut-level experiment, the walls have window traps from-which mosquitoes can exit (see Kitau et al. (2012)). In the community-level experiment, a usual human dwelling is modelled.

9.2.2 State variables

For each of the mosquito agents, properties are individually assigned and updated within the simulation (see Table 9.1).

Table 9.1: Property list of each agent and the relevant model component.

Property	Model component	Type
Spatial position	Motion	Set of coordinates
Inside/Outside the hut	Motion	Binary
Inside/Outside the net	Motion	Binary
Trapped	Motion	Binary
CO_2 concentration	Motion	Float
Fed	Host-seeking	Binary
Time indoors	Host-seeking	Integer
Klinotaxis	Host-seeking	Binary
Dead	Death (Poisoning)	Binary
Accumulated dosage of chemical	Poisoning	Float

9.2.3 Scale

Hut scale The mosquitoes are initially represented in the simulations as a number of agents in a rectangular patch of 3m (which is a typical experimental hut-size (see WHO et al. (2006))) at uniformly random spatial positions.

Community scale In the community-level simulations, mosquitoes are randomly positioned inside the simulated transmission domain of 25,600 m^2 size with multiple households located at a distance not less than 40 m from each other so that there is no competitive attraction caused by vision (see Bidlingmayer and Hem (1980)). The hut-size for the household-case is 13 m.

Time scale The ABM simulations cover a period of one night (10 hours) plus a 24-hour additional delayed mortality. Each calculation simulates an experiment of 34 hours.

9.3 Process overview and scheduling

9.3.1 Processes

The model describes and calibrates the responses and behaviours of the mosquito based on four basic components: motion, host-seeking, poisoning and death (see Table 9.2), where each of the components has a number of related features. The movement and host-seeking behaviour of the mosquito is governed by an attraction model, based on the assumption that a mosquito estimates the direction of odour increase (the gradient) from the host via the klinotaxis mechanism (see Vickers (2000)). As the mosquito approaches the host, the likelihood of accepting steps away from the host reduces (see Figure 9.3b). The current spatial position of the mosquito is updated at every time step.
Apart from the physical material barrier posed by the insecticidal nets, they are also equipped with poisoning and repellent effects. A mosquito is said to be exposed to the poison upon contact with the net surface (see Jones et al. (2021)). Again, the explanation of detoxification is considered, in which the chemical concentration accumulated in the mosquito body is exponentially decaying with a rate which is dependent on both the chemical and mosquito species (see Nardini et al. (2012); Kerkut et al. (1985)). The total accumulated dosage of poison which depends on the number of contacts with the net and the detoxification rate, is updated for every mosquito at each time step, and determines their probability of death. The delayed mortality that is a result of the prolonged impact of poison in mosquitoes is also accounted for. If the mosquito is marked as dead, the mosquito is removed from the simulation such that no properties of the mosquitoes are updated again. The repulsion effect amplifies as the mosquito approaches the source of repellent (see Figure 9.4). This repulsion effect influences the mosquitoes decision of approaching the host and can induce early exit from the hut. If a mosquito exits, no other property of such mosquito is updated in the simulation except for their mortality status which is updated after 10 hours, and tracked for 24 hours. This is because the delayed lethal impact of the chemical is assumed to start after the 10 hours in the hut.
A mosquito is scored as fed if it penetrates through the net and its updated position is very close to the host (a minimal distance ϵ between a mosquito and the host is defined). In the hut-level case, a mosquito can take only one bite since there is only one human. However, in the household and community-level case, the mosquito can take several bites; in this case, up to 5 bites. In household and community-level simulations, the tendency of mosquitoes to switch to neighbouring individuals after spending a certain period of

time in unsuccessful attempts to feed on a protected human is also considered. Hence, in any case, the mosquito switches to a pure random walk, without any control of attractive odour, if the maximum number of bites is reached or if the maximum time a mosquito can spend on host-seeking is used up. Nevertheless, the barriers raised by the net and the repellent effect alongside the effect of chemical poisoning remains functional under this condition. For all mosquitoes that are inside the hut and are not dead or exited, the time spent indoors is updated at each time step. In the community-level case, if the mosquito consumes an insufficient amount of blood before exiting the household, the mosquito begins the host-seeking process from the outset, except that the abandoned household is not accounted for when the total concentration of the CO_2 is computed. It is also assumed that the host-seeking time count is reinitialised after entering a new household. See Figure 9.1 for a diagrammatic explanation of the above discussed processes.

Table 9.2: Modelled processes

Model component	Attributes	Definition
Host-seeking		
	· CO_2 concentration, Klinotaxis	Equation 9.2
	· Distance-dependent attraction	Equation 9.4
	· Host seeking time	
Motion		
	· Random walk, accept/reject steps	Equation 9.3
	· Excito-repellency	Equation 9.10
Poisoning		
	· Accumulation of the chemical dosage	Equation 9.6
	· Detoxification	Equation 9.7
Death		
	· Natural mortality	Equation 9.5
	· Insecticide-induced mortality	Equation 9.8
	· Delayed mortality	Equation 9.5 with model extension

9.3.2 Schedule

The update of the property list of mosquitoes takes place at the same time after each time step. One iteration step in the simulation corresponds to 2 seconds. The basic algorithm for the execution is given in Algorithm 1.

Algorithm 1 Model algorithm

1. Propose candidate position $\mathbf{x}^n$ by adding a stochastic increment to the previous position, i.e. compute $\mathbf{x}^n$ by Equation 9.1;
2. Account for natural mortality. Generate random number $u \sim U[0,1]$. Remove the agent if $u < \alpha^{\Delta t}$;
3. Account for insecticide-induced mortality. Generate random number $u \sim U[0,1]$. Remove the agent if $u < \alpha_p^{\Delta t}$;
4. Evaluate the CO_2 concentration $C(\mathbf{x}^n)$ at new position $\mathbf{x}^n$ as given in Equation 9.2;
5. Compute the scaling factor $\sigma_{acc}(\mathbf{x}^n)$ as given by Equation 9.4;
6. Recalculate the scaling factor, while considering the *excito-repellency*, conditioned on the amount of accumulated chemical by Equation 9.10;
7. Compute probability of acceptance by attraction, $\alpha_a(\mathbf{x}^n|\mathbf{x}^{n-1})$ for position $\mathbf{x}^n$ by Equation 9.3;
8. Compute the probability of rejection α_{rej} resulting from repellent $\alpha_r(\mathbf{x}^n|d_p, s)$ by Equation 9.9;
9. Generate random number $u \sim U[0,1]$, if $u < min\{1, \alpha_a(1-\alpha_r)\}$, mark position $\mathbf{x}^n$ as preliminarily accepted; otherwise, mark the position as rejected and remain at the old position $\mathbf{x}^n = \mathbf{x}^{n-1}$;
10. Account for the physical net barrier. If candidate step $\mathbf{x}^n$ is inside and old position $\mathbf{x}^{n-1}$ is outside of the net, and position $\mathbf{x}^n$ was preliminarily accepted, generate random number u. If $u < 1 - p_{net}$, accept the new position $\mathbf{x}^n$. Otherwise, select the closest point on the net $\mathbf{x}^{net}$ to $\mathbf{x}^{n-1}$ and assign new position $\mathbf{x}^n = \mathbf{x}^{net}$;
11. Account for the wall barrier. If candidate step $\mathbf{x}^n$ is outside and old the position $\mathbf{x}^{n-1}$ is inside of the hut and position $\mathbf{x}^n$ was preliminarily accepted, generate random number u. If $u < p_{hut}$, accept the new position $\mathbf{x}^n$. Otherwise, choose the closest point on the wall $\mathbf{x}^{wall}$ to $\mathbf{x}^{n-1}$ and assign a new position $\mathbf{x}^n = \mathbf{x}^{wall}$;
12. Update the total accumulated chemical dosage C_{tot} by Equation 9.7;
13. Account for *detoxification* of the total accumulated chemical dosage C_{tot} with the rate α;
14. Update the property list of the mosquito;
15. Move to step 1, $n \rightarrow n+1$

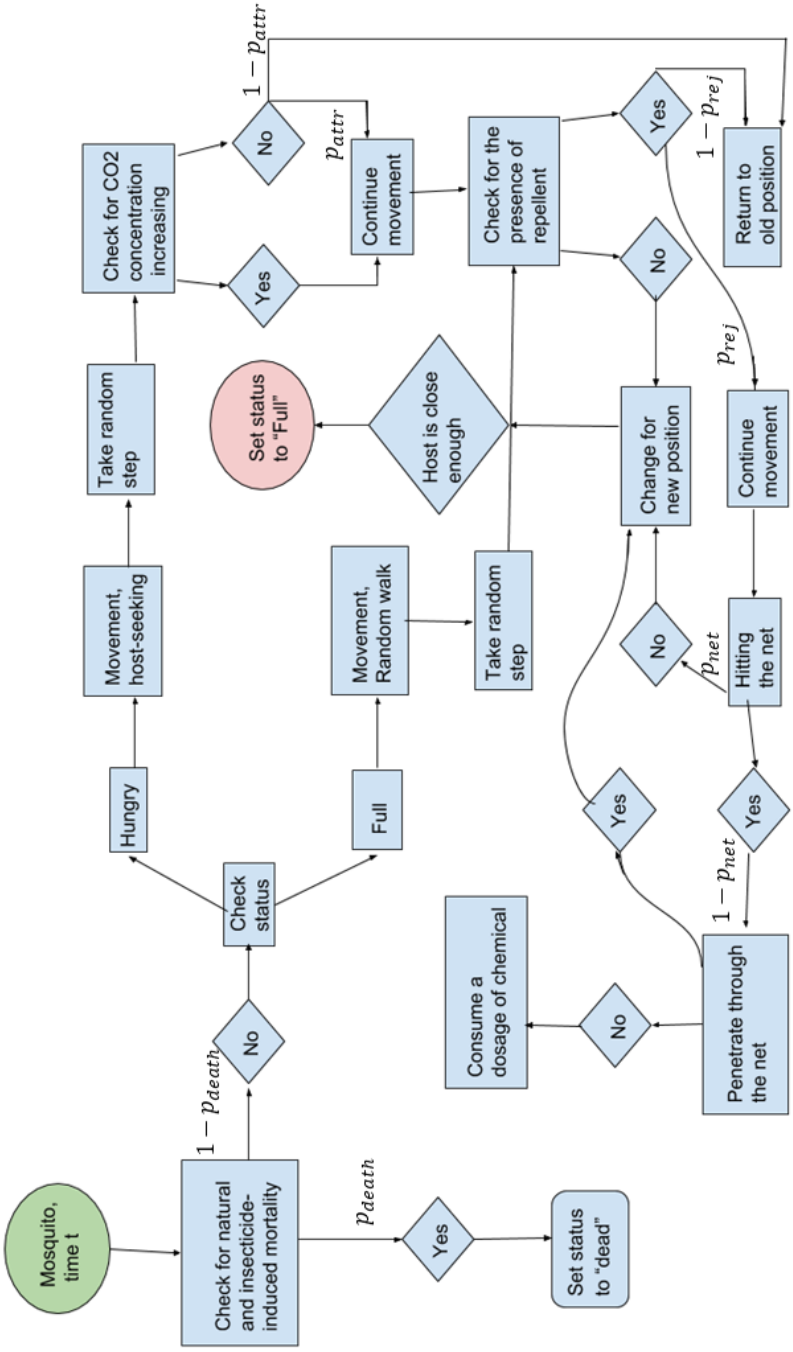


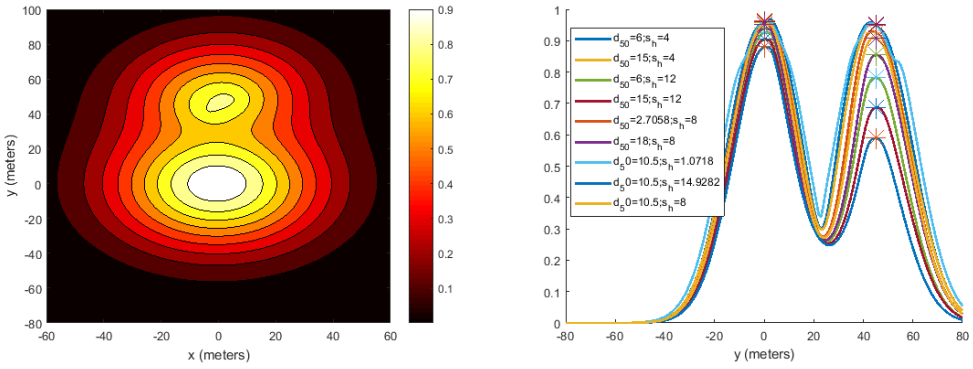
Figure 9.1: A decision tree showing the key features of the ABM algorithm of mosquito host-seeking actions in the presence of the LLINs introduced in Shcherbacheva et al. (2018), taken from Shcherbacheva (2019). Here, some of the choices are probabilistic, depending on the state of the agent. p_{death} denotes the probability of death, p_{net} stands for the probability of being blocked by the physical net barrier, p_{attr} denotes the probability of accepting the proposed step, p_{rej} stands for the probability of rejecting the proposed step due to the repellent effect.

9.4 Design concepts

9.4.1 Basic principles

Mosquito movement and attraction model The mosquito attraction model is based on the assumption that a mosquito estimates the direction of odour increase (the gradient) from the host via the klinotaxis mechanism Vickers (2000). During this plume-tracking activity, the mosquito samples the host odour at one location, changes location and then repeats the sampling, and uses its memory of the concentrations previously observed to select the next position (see Cardé (1996); Cummins et al. (2012)). In the present work, the space is continuous, and the movement of mosquitoes is guided based on Euclidean distances to the humans and households. In the absence of the sensory signals, the movement of mosquito constitutes pure random walk, which is typical for the ABMs that include animal navigation (see Tang and Bennett (2010)). Imitating this mechanism, the flight of mosquitoes is modelled as a discrete-time correlated random walk. However, when there are attraction effects, sufficiently close for sensing the host, the main features of the Metropolis algorithm are employed to simulate the random walk directionally biased by attraction (see Metropolis et al. (1953)). The Metropolis algorithm features an *accept/reject* movement. After a random candidate position is proposed by the Brownian motion, the probability of accepting the new position for a given agent is defined to favour candidate steps taken in the direction of increasing concentration of CO_2, i.e. towards the attraction source, (see Metropolis et al. (1953)). In addition, the acceptance probability is also influenced by the presence of treated nets and the barrier imposed by the walls in human dwellings. These effects are incorporated by a rejection function. The concentration of attractive odour and the area covered by the odour is modelled using the diffusion equation solution, taking into account only the diffusive spread of the odour. The effect of wind is ignored for simplicity. Thus, the region of high odour concentration may be assumed to be a specific location where the host is located and the maximum distance at which the mosquito is able to detect the host, is seen as the region of low concentration. This is consistent with the principle of the diffusion equation that describes the expel of the flow of certain quantities (intensity, temperature) over space (see Crank (1979)).Therefore, the Gaussian Kernel centred around the host's spatial location is used. Naturally, when other significant factors influencing the dispersion of mosquitoes are taken into account, the concentration may be defined in a different way, such as using advection-reaction-diffusion equations, which includes the flow of air and intermittent concentration plumes etc.(see Cummins et al. (2012)). The concentration that allows mosquitoes to sense humans in the household and community-level case is calculated similarly to the case of a single person, i.e. as a Gaussian, with the argument given by a weighted sum of the individual distances from the position of the mosquito to the location of each of the hosts. The total attracting concentration is based on the principle of the function *softmax*, which has been widely implemented in machine learning and neural networks, (see Bishop and Nasrabadi (2006), Montague (1999)). The weight is added to account for the fact that, depending on the mosquito species, the response of a mosquito to the cue emitted from households increases at a short distance of 5-15 m due to its attraction to visually conspicuous objects (see Bidlingmayer and Hem (1980); Hawkes et al. (2017); Van Breugel

et al. (2015)). The key emphasis here is on the nearest target concept, which basically implies that factors other than just CO2 alone often cause the mosquito to localise the search at a short distance, as stated in Bidlingmayer and Hem (1980); Hawkes et al. (2017); Van Breugel et al. (2015). Non-normalised weights are applied inversely proportional to the distance following this rationale. Note that the community-scale model's form of concentration is consistent with the evidence that larger agglomerates emit stronger odours, thus attracting more mosquitoes Cummins et al. (2012) (see the illustration in Figure 9.2).



(a) (b)

Figure 9.2: Softmax function in a special case of two households. The first household includes 6 individuals (located (0,0)) and the other household consists of 2 individuals (located at (0,45)) for different values of d_{50} and s (a) 2D plot, (b) 1D plot of the softmax function along the y axis.

The increased mosquito greediness, as a result of activation of the heat sensors at a short distance to the host, is accounted for by using a linearly distance-dependent scaling factor. The functional behaviour of the scaling factor results in such a movement that steps in the concentration plume taken towards the host are always accepted (see Figure 9.3a). The design of the algorithm basically resembles a well-known Simulated Annealing optimisation method, introduced in Kirkpatrick et al. (1983). The difference here is that the 'annealing temperature schedule' is replaced with the 'greediness scale', which is associated to the distance from the mosquito to the host. In addition, the scaling factor is further defined in such a way that it depends not only on the distance to the host but also on the repellent effect. This extension was done to fit the exit rates properly. The scaling factor is computed in the community-level case with the distance to the nearest hut, perceived as the nearest visible feature.

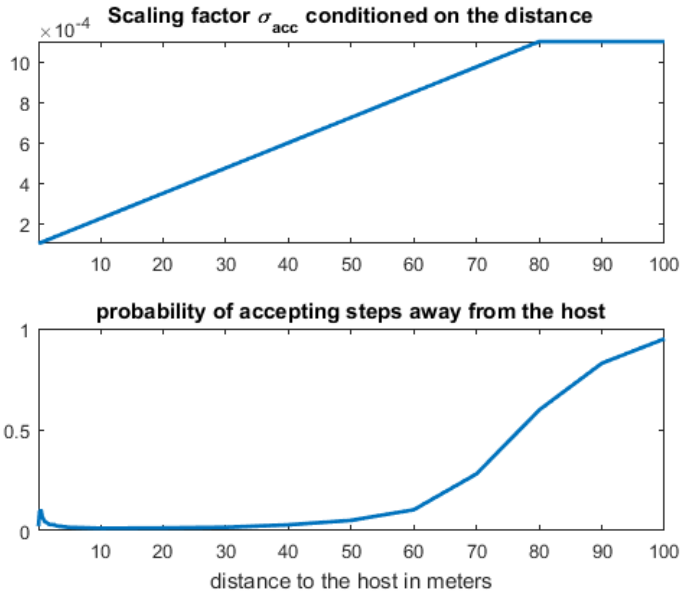


Figure 9.3: Average probability of accepting candidate steps taken away from the host; as a function of distance from the host.

Mosquito poisoning and mortality model In this work, both natural and insecticide-induced mortality are considered in the model. At the onset, when the mosquito has not yet taken the poisonous chemical, the death rate is reduced to the natural mortality. As the dosage of the chemical gradually increases in the mosquito, the chemical-induced death occurs from the lethal insecticide dosage. In continuous time, the natural mortality in a declining population is commonly modeled by means of an ordinary differential equation. Here, the continuous-time mortality rate is transformed into probability of death per unit time. This is achieved by discretisation in time leading to agent-based rules rather than the rates. The insecticidal induced mortality is modelled using the total accumulated dosage with effective poisoning impact given by a scaling coefficient that varies depending on the insecticide used for LLIN treatment. As a result, the total probability of death per unit change in time is simulated as the sum of natural and insecticide-induced mortality.

Repellent model The influence of spatial repellent is imitated by conducting the accept/reject method, with the rejection probability defined by logistic equation. The logistic function is used to describe certain kinds of growth rate that have an S-shaped behaviour. At first, this function grows exponentially, but eventually grow more slowly and levels off, due to certain restrictions. In order to model the repellent effect caused by the net, the function was modified so that the rejection probability at the candidate position attenuates as the distance to the host increases (see Figure 9.4).

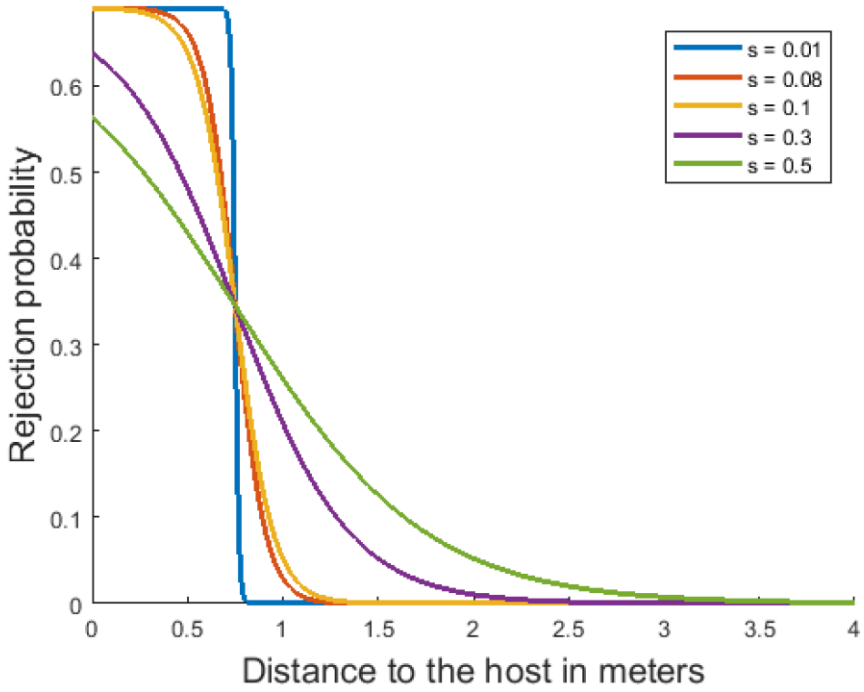


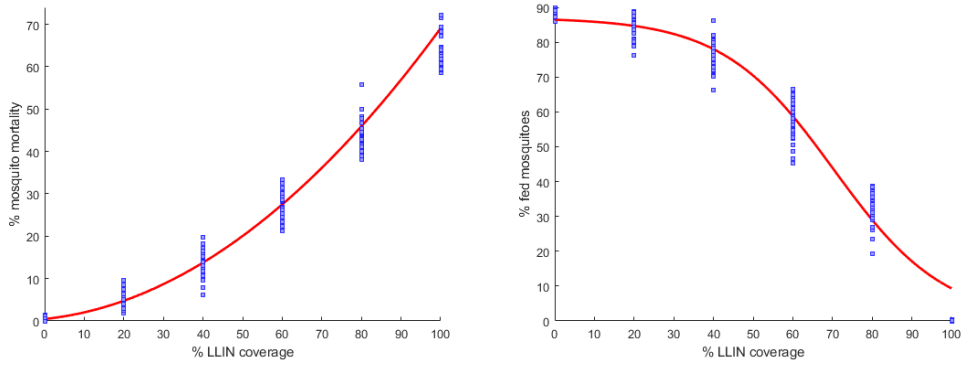
Figure 9.4: Probability of rejection associated with repellents for different values of the spatial range of repellents.

9.4.2 Emergence

The model behaviour and outputs emerge from the implicit structure of the model.

Hut level In the hut-level case, the impact of LLIN is calibrated by data from Kitau et al. (2012). Two different model parameterisation versions were selected to test various hypotheses explaining the different host-seeking behaviour of the species. Both model calibrations gave the same overall results for the impact of LLINs. Thus, the impact of LLINs is emerging by data and not by the specific hypothesis imposed in the model calibrations (see Shcherbacheva et al. (2018) for more details).

Community level At the community level, the uncertainty from the sampled parameters at the hut-level is taken into account and a sensitivity analysis is performed using a central composite design with respect to the assumed parameters. With moderate perturbations in the assumed parameter values, the sensitivity analysis indicates that the system's behaviour remains more or less the same (see Figure 9.5).



(a) (b)

Figure 9.5: Uncertainty from the sampled parameters at hut level together with the variability of the community-level assumed parameters for (a) mortality rate (b) biting rate, of *An. gambiae* when confronted with LLIN impregnated with an Alphacypermethrin treatment kit, fitted with respect to partial coverage of LLIN for the household size of 2 when assuming no behavioural alterations caused by the parasite.

9.4.3 Adaptation

In the mosquito host-seeking behaviour presented in this work, there is as attractive potential (CO_2 emitted by human) driving the mosquito movement. This makes the mosquito make more directional movement towards the host and can hardly accept steps away from the host. This potential is given as a solution of the diffusion equation with a point source specified as the Gaussian Kernel centred at a spatial location of the host. This behaviour does not change with time and the same set of rules applies regardless of the status of the agent. Furthermore, the increased mosquito greediness, as a result of activation of the heat sensors at a short distance to the host is accounted for by adding a linearly distance-dependent scaling factor.
Moreover, there is a repulsive force that is regarded as *contact irritancy* introduced by the LLINs impregnated with chemicals. The repulsive force can induce early exit from the hut. This effect is generated by a rejection probability of a new position, conditioned on the presence of chemicals. In general, different mosquito behaviours were observed when confronted with each of the chemical treatments under study.

9.4.4 Objectives

There is no individual success or objectives that agents work towards except the general interest to obtain a full blood-meal. They stick to an 'indirect objective seeking', in which they simply follow the rules that reproduce observed behaviour.

9.4.5 Learning

In the ABM for this study, the navigation capacities are described as the ability to orient in the odour plume emitted from the hosts known as klinotaxis, where the mosquito uses its memory of CO_2 concentration from the past to select the next direction of movement. This process is included to enable the mosquito to make more directional movement to find the host(s). The mosquitoes do not change their behaviour during the course of the simulations.

9.4.6 Prediction

The adaptive behaviour of mosquitoes is based on the implicit prediction that, taking steps leading away from the host is likely not to be accepted. This assumption is accurate in the sense that, if the new concentration (i.e. in the new mosquito position) is higher than the old concentration, the step is always accepted, otherwise, the step can be accepted with a certain probability. In addition, repellency and physical barrier play a role of prediction, as the candidate position is rejected upon being repelled or blocked physically. Mosquitoes do not intentionally make decisions. The 'decisions' are probabilistic and lead to the overall results in a statistical average sense.

9.4.7 Sensing

Motion capacities are captured by considering the mode of movement, switching from a pure random walk in the absence of sensory cues, to a directionally biased random walk, after entering the CO_2 plume. Mosquitoes are usually able to sense the human host only at a distance of less than 80 m (see Cardé (1996)). To account for a short-distance (less than 3 m) behaviour, where increased sensory information induces greater attraction to the host, a third mode of movement is involved. This effect is included by a concentration scaling factor which facilitates more directional movement towards the host. In the community level scenario, the CO_2 concentration sensed by the mosquito at a short distance (less than 15 m (see Bidlingmayer and Hem (1980); Hawkes et al. (2017))) is assumed to be the one emitted from the nearest household. The key emphasis here is on the nearest target concept, which basically implies that factors other than just CO2 alone often cause the mosquito to localise the search at a short distance, as stated in Bidlingmayer and Hem (1980); Hawkes et al. (2017); Van Breugel et al. (2015). Non-normalised weights are applied inversely proportional to the distance following this rationale, which aligns with the evidence that larger agglomerates emit stronger odours, thus attracting more mosquitoes (see Cummins et al. (2012)). The mechanism of sensing (in community-level case) is modelled with the softmax function and the reverse-logistic weights (see Figure 9.2). These sensing assumptions included in the simulations are typical for this modelling approach.

9.4.8 Interactions

The ABM presented here consists of non-interactive mosquito agents. However, there is direct interaction between mosquitoes and humans given that mosquitoes can sense and bite humans.

9.4.9 Stochasticity

The simulation results depend on random numbers, so the output of each experiment is stochastic. Initially, all the mosquito agents occupy randomly generated spatial locations in the simulation domain. In the community-level case, households are randomly located inside the spatial domain. These randomisations used for the initialisation of spatial positions is done at each successive repetition of the algorithm to average for stochasticity arising from difference in spatial arrangement and position.
The host-seeking process is given by a random walk with accept-reject stepping, with the acceptance probabilities being estimated to match the observed effects associated with mosquito responses to the host, in the presence of the LLIN (such as repulsion and early exit) and poisoning by insecticides. The candidate position is randomly proposed by using uniformly distributed random direction with respect to the previous position. Random numbers are generated from the uniform distribution to compare with the probabilities of accepting (by attraction) and rejecting (associated with repellent) a candidate position, accounting for dead mosquitoes and accounting for the barriers posed by the net and wall.
In the household-level model, mosquitoes are assumed to randomly choose one of the humans upon entering the household. The scenario is then reduced to the case of a single host in the hut. Moreover, the diversion to other humans which happens after a certain period of time spent in unsuccessful attempts to feed on the protected host was made by choosing another person at random among the other inhabitants of hut. Again, randomisation is employed for the multiple biting modelled in the household-level. The maximum number of successful feeding attempts can be up to 5, and this property is randomised and sampled separately for each of the mosquitoes.
In order to ensure statistical accuracy needed for the calibration of model parameters, the averaged model outputs obtained by multiple simulations are taken, using a sufficiently large swarm of mosquitoes in every case. It should also be noted that the data from Kitau et al. (2012) are given in percentages, and as such, the absolute number of mosquitoes does not influence the results. However, since the model is stochastic, it is necessary to average all the results over several repetitions. A combination of 6 repetitions and a swarm of 600 mosquitoes (for the hut-level) results in a relatively small variance considering the minimal CPU time. The number of repetitions in the community-level case is larger than in the hut-level experiment, to average for the stochasticity arising from the spatial arrangement of the households. Note that, in the community-level simulations, combinations of parameter values are randomly selected from the estimated posteriors at each successive iteration of the algorithm for uncertainty quantification.

9.4.10 Collectives

Collective effects are not covered in this model.

9.4.11 Observation

Field data is used to calibrate the hut-level model. At the end of the simulations, the proportion of fed and dead mosquitoes (which are of interest) are recorded, although the proportion of exited mosquitoes can also be recorded. These proportions are recorded separately for two cases: assuming no behavioural alterations and assuming alterations by parasite, separately for the two mosquito species and each of the chemical treatments considered in the study. Furthermore, in case of behavioural alterations, the contact rates are recorded separately for infectious and uninfected mosquitoes. Response surfaces are fitted to the relevant responses (contact and mortality rates) obtained from the ABM simulations, with respect to the household size and the coverage for outputs corresponding to each of the insecticidal treatments, respectively. The response surface is fitted for all the aforementioned cases. The coefficients from the fitted response surfaces can be used when incorporating the ODE-based model of malaria transmission, as they provide the values of the main parameters, which enables the extension of the ABM simulations carried out over a 'snapshot' period of one night to a continuous time interval. It was observed that as the coverage with LLINs increases, the death rates increase and the fed rates decrease. However, there is an insignificant dependence of the mortality rates on the household size.

9.5 Initialisation

In the hut-level situation, one human agent and with a swarm of 600 mosquitoes is used. At the household-level, one household (with several number of humans) and a swarm of 700 mosquitoes is employed. At the community level, households of different sizes ranging from 2 to 10 people are used. A constant number of 700 mosquitoes and about 20 individuals are utilised for each experimental run. The mosquitoes can either be infected or uninfected. The number of infectious mosquitoes is constant for a single experiment. This is because it takes a period of 10 to 12 days for parasites to reach a stage where they are ready for transmission, whereas the ABM simulation for this study is only for a night. Humans can either be protected or unprotected. The protection is marked with 0 (for unprotected humans) and 1 (for protected humans). The percentage of protected humans for each household remains constant in each of the simulations. Note that, for the hut-level case, the host is always protected. Again the status of the hosts and the households are marked as not-bitten in the initialisation. For each mosquito agent, there is an associated number of states that can be 0 or 1, like dead or alive. For each household, such state is assigned and updated, to track if the mosquito was host-seeking in that household recently. In the simulation model, the mosquitoes are presented as a number of agents in a two-dimensional rectangular domain, initially placed at uniformly generated random spatial locations. In the community-level simulations, mosquitoes are initially randomly positioned inside the experimental domain with multiple households located at a distance not less than 40 m from each other, such that there is no competitive attraction caused

by vision (see Bidlingmayer and Hem (1980)). Depending on the initial positions of the mosquitoes, the initial concentrations are assigned. The host-seeking time, the number of contacts with the net, and the accumulated dosage of chemicals are initially set to zero for all the mosquitoes and updated at each iteration. In the community-level case, it is assumed that upon entering a new household, the host-seeking time count is reinitialised. This was done to consider the habit of early exit after a certain time spent inside, as the so-called *exophily*.

9.6 Input data

The model does not employ input data to represent processes that vary with time.

9.7 Submodels

The equations for the modelled processes in Table 9.2 are given below:
A mosquito selects a new position $\mathbf{x}^n$ with the formula

$$\mathbf{x}^n = \mathbf{x}^{n-1} + \delta\mathbf{W}, \tag{9.1}$$

where the increment $\delta\mathbf{W}$ added to $\mathbf{x}^{n-1}$ is sampled in random direction, with a step size given by a normal distribution $N(\mathbf{x}_0, \sigma^2 I)$.

The concentration that enables a mosquito to sense the host at a distance $d(\mathbf{x}, \mathbf{x}^h)$ is expressed as

$$C(\mathbf{x}, \mathbf{x^h}) = \exp\left[-\frac{d^2(\mathbf{x}, \mathbf{x}^h)}{2\sigma_a^2}\right], \tag{9.2}$$

where $\mathbf{x}$ denotes the position of the mosquito. The standard deviation of the Gaussian σ_a determines a maximal distance at which the mosquito is able to sense the host.

The acceptance probability is defined as

$$\alpha_a(\mathbf{x^n}|\mathbf{x^{n-1}}) = \min\left(1, \frac{p(\mathbf{x^n})}{p(\mathbf{x^{n-1}})}\right), \tag{9.3}$$

where $p(\mathbf{x^n})/p(\mathbf{x^{n-1}})$ is the ratio of the attraction potential function $p(\mathbf{x})$ defined at each point $\mathbf{x}$, which depends on the concentration and other attraction factors.

The scaling factor for the attraction potential which depends on the distance to the host is modelled as

$$\sigma_{acc}(\mathbf{x}, \mathbf{x}^h) = \begin{cases} \sigma_{acc}^1 + \sigma_{acc}^2 d(\mathbf{x}, \mathbf{x}^h), & d(\mathbf{x}, \mathbf{x}^h) \leq 80 \\ \sigma_{acc}^{\max}, & d(\mathbf{x}, \mathbf{x}^h) > 80. \end{cases} \tag{9.4}$$

The above function increases from the minimum value of σ_{acc}^1 with a slope given by the parameter σ_{acc}^2 until it is replaced by a constant which suitably provides a purely random

movement outside the concentration plume (see Shcherbacheva et al. (2018)).

The natural death rate is parameterised as

$$\alpha^{\Delta t} = \min\left\{1, \mu\Delta t\right\}, \tag{9.5}$$

where $\Delta t = 2$ seconds is used for all simulations, and a value for μ taken from the literature (see Shcherbacheva et al. (2018) for more details).

The total accumulation of dosage of chemical is modelled as

$$C_{tot}(n+1) = \sum_{i=1}^{n+1} D_i = C_{tot}(n) + D_{n+1}, \tag{9.6}$$

where D_i is non-zero in case of hitting the net surface (i.e. equal to the unit dosage), and zero otherwise.

For the extended model with detoxification effect, the total accumulation of dosage of chemical is modelled as

$$C_{tot}(n+1) = C_{tot}(n) + D_{n+1} - \alpha C_{tot}(n)\Delta t. \tag{9.7}$$

The insecticide-induced mortality is parameterised as

$$\alpha_p^{\Delta t}(n) = \mu_p C_{tot}(n)\Delta t, \tag{9.8}$$

where the effective poisoning impact is obtained by a scaling coefficient μ_p which depends on the given insecticide used for LLIN treatment.

The repulsion probability is modelled as

$$C_{rej} = r\left[1 - 1/\Big(1 + \exp\big(-\big(d(\mathbf{x}, \mathbf{x}^h) - d_{50}\big)/s\big)\Big)\right], \tag{9.9}$$

where $d(\mathbf{x}, \mathbf{x}^h)$ denotes the distance from the mosquito to the protected human and r ranges from 0 to 1. The parameters d_{50} and s give the range of coverage and the spread of the chemical. The logistic function is adjusted such that the rejection probability at the candidate position $\mathbf{x}$ grows as the mosquito gets closer to the repellent source.

Excito-repellency is modelled as

$$\sigma_{acc}(\mathbf{x}, C_{tot}) = \sigma_{acc}(\mathbf{x}) + \mu_e \cdot C_{tot}, \tag{9.10}$$

where C_{tot} denotes the total dosage of chemical consumed by the mosquito (see Equation 9.6) and σ_{acc} is given by Equation 9.4.

References

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis*, 27(2), pp. 93–115.

Arino, J. and Van den Driessche, P. (2003). A multi-city epidemic model. *Mathematical Population Studies*, 10(3), pp. 175–193.

Aron, J.L. and May, R.M. (1982). The population dynamics of malaria. In: *The population dynamics of infectious diseases: theory and applications*, pp. 139–179. Springer.

Baudrot, V., et al. (2016). The adaptation of generalist predators' diet in a multi-prey context: Insights from new functional responses. *Ecology*, 97(7), pp. 1832–1841.

Bidlingmayer, W. and Hem, D. (1980). The range of visual attraction and the effect of competitive visual attractants upon mosquito (Diptera: Culicidae) flight. *Bulletin of Entomological Research*, 70(2), pp. 321–342.

Binder, B.J., Ross, J.V., and Simpson, M.J. (2012). A hybrid model for studying spatial aspects of infectious diseases. *The ANZIAM Journal*, 54(1-2), pp. 37–49.

Bishop, C.M. and Nasrabadi, N.M. (2006). *Pattern recognition and machine learning*, vol. 4, 4. Springer.

Bobashev, G.V., Goedecke, D.M., Yu, F., and Epstein, J.M. (2007). A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. In: *2007 winter simulation conference*, pp. 1532–1537.

Bobbitt, Z. (2021). *What is Moran's I? (Definition & Example)*. Technical report. https://www.statology.org/morans-i/.

Boyce, J.M. (2007). Environmental contamination makes an important contribution to hospital infection. *Journal of hospital infection*, 65, pp. 50–54.

Boyd, M.F. (1949). Epidemiology of malaria: factors related to the intermediate host. *Malariology: A comprehensive survey of all aspects of this group of diseases from a global standpoint: WB Saunders Company*.

Brauer, F., Castillo-Chavez, C., and Feng, Z. (2019). *Mathematical models in epidemiology*, vol. 32. Springer.

Browning, A.P., et al. (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173), p. 20200652.

Cameron, A.C. and Trivedi, P.K. (2010). *Microeconometrics using stata*, vol. 2. Stata press College Station, TX.

Cardé, R.T. (1996). Odour plumes and odour-mediated flight in insects. *Olfaction in mosquito-host interactions*, 200, pp. 54–70.

Citron, D.T., et al. (2021). Comparing metapopulation dynamics of infectious diseases under different models of human movement. *Proceedings of the National Academy of Sciences*, 118(18), p. e2007488118.

Colbourne, M. et al. (1966). Malaria in Africa. *Malaria in Africa.*

Crank, J. (1979). *The mathematics of diffusion*. Oxford university press.

Cummins, B., et al. (2012). A spatial model of mosquito host-seeking behavior. *PLoS computational biology*, 8(5), p. e1002500.

De Jong, P., Sprenger, C., and Van Veen, F. (1984). On extreme values of Moran's I and Geary's c. *Geographical Analysis*, 16(1), pp. 17–24.

Dicker, R.C., Coronado, F., Koo, D., and Parrish, R.G. (2006). Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics.

Diekmann, O., Heesterbeek, J.A.P., and Metz, J.A. (1990). On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28(4), pp. 365–382.

Diniz-Filho, J.A.F., Bini, L.M., and Hawkins, B.A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, 12(1), pp. 53–64.

Dowell, S.F. and Ho, M.S. (2004). Seasonality of infectious diseases and severe acute respiratory syndrome–what we don't know can hurt us. *The Lancet infectious diseases*, 4(11), pp. 704–708.

Dunham, J.B. (2005). An agent-based spatially explicit epidemiological model in MASON. *Journal of Artificial Societies and Social Simulation*, 9(1).

Filipe, J.A.N., et al. (2007). Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. *PLoS computational biology*, 3(12), p. e255.

Frees, E.W. et al. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

Geary, R.C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3), pp. 115–146.

Gilbert, N. (2019). *Agent-based models*, vol. 153. Sage Publications.

Goethert, H.K., Saviet, B., and Telford, S.R. (2009). Metapopulation structure for perpetuation of Francisella tularensis tularensis. *BMC microbiology*, 9(1), pp. 1–9.

Griffith, D.A. (2020). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models 1. In: *Practical handbook of spatial statistics*, pp. 65–82. CRC press.

Grimm, V., et al. (2010). The ODD protocol: a review and first update. *Ecological modelling*, 221(23), pp. 2760–2768.

Guţoiu, G.I. and Pandelea, S. (2016). The electoral geography of the 2016 Presidential Election in Portugal.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: efficient adaptive MCMC. *Statistics and computing*, 16(4), pp. 339–354.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, pp. 223–242.

Hanski, I. et al. (1999). *Metapopulation ecology*. Oxford University Press.

Hawkes, F.M., et al. (2017). Exploiting Anopheles responses to thermal, odour and visual stimuli to improve surveillance and control of malaria. *Scientific reports*, 7(1), pp. 1–9.

Hethcote, H.W., Stech, H.W., and van den Driessche, P. (1981). Periodicity and stability in epidemic models: a survey. In: *Differential equations and applications in ecology, epidemics, and population problems*, pp. 65–82. Elsevier.

Hunter, E., Mac Namee, B., and Kelleher, J. (2020). A hybrid agent-based and equation based model for the spread of infectious diseases. *Journal of Artificial Societies and Social Simulation*, 23(4).

Hunter, E., Mac Namee, B., and Kelleher, J.D. (2017). A taxonomy for agent-based models in human infectious disease epidemiology. *Journal of Artificial Societies and Social Simulation*, 20(3).

Hunter, E., Mac Namee, B., and Kelleher, J.D. (2018). A Comparison of Agent-Based Models and Equation Based Models for Infectious Disease Epidemiology. In: *AICS*, pp. 33–44.

Jewell, N.P. (2003). *Statistics for epidemiology*. chapman and hall/CRC.

Jones, J., Murray, G.P., and McCall, P.J. (2021). A minimal 3D model of mosquito flight behaviour around the human baited bed net. *Malaria Journal*, 20(1), pp. 1–16.

Kang, S.Y., McGree, J., and Mengersen, K. (2014). The choice of spatial scales and spatial smoothness priors for various spatial patterns. *Spatial and Spatio-temporal Epidemiology*, 10, pp. 11–26.

Keeling, M. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals. 837 Princeton University Press*. Princeton.

Kerkut, G.A., Gilbert, L.I., et al. (1985). *Comprehensive insect physiology, biochemistry and pharmacology*, vol. 9. Pergamon Oxford.

Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772), pp. 700–721.

Kilama, M., et al. (2014). Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae sl using three sampling methods in three sites in Uganda. *Malaria journal*, 13(1), pp. 1–13.

Kirkpatrick, S., Gelatt Jr, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *science*, 220(4598), pp. 671–680.

Kitau, J., et al. (2012). Species shifts in the Anopheles gambiae complex: do LLINs successfully control Anopheles arabiensis? *PloS one*, 7(3), p. e31481.

Kondo, K. (2018). Testing for global spatial autocorrelation in Stata.

Leeper, T.J. (2017). Interpreting regression results using average marginal effects with R's margins. *Reference manual*, 32.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *American Entomologist*, 15(3), pp. 237–240.

Li, M.Y. and Shuai, Z. (2009). Global stability of an epidemic model in a patchy environment. *Canadian Applied Mathematics Quarterly*, 17(1), pp. 175–187.

Li, P. and Vu, Q.D. (2015). A simple method for identifying parameter correlations in partially observed linear dynamic models. *BMC systems biology*, 9(1), pp. 1–14.

Linard, C., Ponçon, N., Fontenille, D., and Lambin, E.F. (2009). A multi-agent simulation to assess the risk of malaria re-emergence in southern France. *Ecological Modelling*, 220(2), pp. 160–174.

Mao, L. (2014). Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network-an agent-based simulation. *Applied Geography*, 50, pp. 31–39.

Metropolis, N., et al. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), pp. 1087–1092.

Miao, H., Xia, X., Perelson, A.S., and Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM review*, 53(1), pp. 3–39.

Molineaux, L., Gramiccia, G., Organization, W.H., et al. (1980). *The Garki project: research on the epidemiology and control of malaria in the Sudan savanna of West Africa*. World Health Organization.

Montague, P.R. (1999). Reinforcement learning: an introduction, by Sutton, RS and Barto, AG. *Trends in cognitive sciences*, 3(9), p. 360.

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

Nardini, L., et al. (2012). Detoxification enzymes associated with insecticide resistance in laboratory strains of Anopheles arabiensis of different geographic origin. *Parasites & Vectors*, 5(1), pp. 1–12.

Nguyen, M. (2022). *A Guide on Data Analysis*. Technical report. https://bookdown.org/mike/data_analysis/.

North, A.R. and Godfray, H.C.J. (2017). The dynamics of disease in a metapopulation: The role of dispersal range. *Journal of theoretical biology*, 418, pp. 57–65.

Okumu, F.O., et al. (2013). Comparative field evaluation of combinations of long-lasting insecticide treated nets and indoor residual spraying, relative to either method alone, for malaria prevention in an area where the main vector is Anopheles arabiensis. *Parasites & vectors*, 6(1), pp. 1–20.

Perasso, A. (2018). An introduction to the basic reproduction number in mathematical epidemiology. *ESAIM: Proceedings and Surveys*, 62, pp. 123–138.

Perez, L. and Dragicevic, S. (2009). An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1), pp. 1–17.

Rocha, F., Aguiar, M., Souza, M., and Stollenwerk, N. (2013). Time-scale separation and centre manifold analysis describing vector-borne disease dynamics. *International Journal of Computer Mathematics*, 90(10), pp. 2105–2125.

Ross, R. (1911). The prevention of malaria, 2nd edn London. *UK: John Murray*.

Ross, R. (1915). Some a priori pathometric equations. *British medical journal*, 1(2830), p. 546.

Rothman, K.J., Greenland, S., Lash, T.L., et al. (2008). *Modern epidemiology*, vol. 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN 0-13-103805-2.

Sattenspiel, L. and Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical biosciences*, 128(1-2), pp. 71–91.

Sawada, M. (2001). Global spatial autocorrelation indices-Moran's I, Geary's C and the general cross-product statistic. *Laboratory of Paleoclimatology and Climatology, Dept. Geography, University of Ottawa,(Mimeo)*, pp. 45–54.

Shaukat, A.M., Breman, J.G., and McKenzie, F.E. (2010). Using the entomological inoculation rate to assess the impact of vector control on malaria parasite transmission and elimination. *Malaria journal*, 9(1), pp. 1–12.

Shcherbacheva, A. (2019). Agent-based modelling for epidemiological applications.

Shcherbacheva, A., Haario, H., and Killeen, G.F. (2018). Modeling host-seeking behavior of African malaria vector mosquitoes in the presence of long-lasting insecticidal nets. *Mathematical biosciences*, 295, pp. 36–47.

Sherrard-Smith, E., et al. (2019). Mosquito feeding behavior and how it influences residual malaria transmission across Africa. *Proceedings of the National Academy of Sciences*, 116(30), pp. 15086–15095.

Sheytanova, T. (2015). *The accuracy of the Hausman Test in panel data: A Monte Carlo study*.

Shi, P., et al. (2020). Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Science of the total environment*, 728, p. 138890.

Silman, A.J., Macfarlane, G.J., and Macfarlane, T. (2018). *Epidemiological studies: a practical guide*. Oxford University Press.

Smith, T., et al. (2006). Relationship between the entomologic inoculation rate and the force of infection for Plasmodium falciparum malaria. *The American journal of tropical medicine and hygiene*, 75(2_suppl), pp. 11–18.

Sokal, R.R., Oden, N.L., and Thomson, B.A. (1998). Local spatial autocorrelation in a biological model. *Geographical Analysis*, 30(4), pp. 331–354.

Song, C. and Kulldorff, M. (2005). Tango's maximized excess events test with different weights. *International Journal of Health Geographics*, 4(1), pp. 1–7.

Strand, S., Cadwallader, C., and Firth, D. (2011). Using statistical regression methods in education research. *ESRC National Centre for Research Methods, University of Southampton*.

Suárez, E., Pérez, C.M., Rivera, R., and Martínez, M.N. (2017). *Applications of regression models in epidemiology*. John Wiley & Sons.

Sullivan, T.J. (2015). *Introduction to uncertainty quantification*, vol. 63. Springer.

Tang, W. and Bennett, D.A. (2010). Agent-based modeling of animal movement: a review. *Geography Compass*, 4(7), pp. 682–700.

Taylor, S.J., et al. (2014). A tutorial on cloud computing for agent-based modeling & simulation with repast. In: *Proceedings of the Winter Simulation Conference 2014*, pp. 192–206.

Van Breugel, F., Riffell, J., Fairhall, A., and Dickinson, M.H. (2015). Mosquitoes use vision to associate odor plumes with thermal targets. *Current Biology*, 25(16), pp. 2123–2129.

Vickers, N.J. (2000). Mechanisms of animal navigation in odor plumes. *The Biological Bulletin*, 198(2), pp. 203–212.

Wang, W. and Zhao, X.Q. (2004). An epidemic model in a patchy environment. *Mathematical biosciences*, 190(1), pp. 97–112.

West, B.T., Welch, K.B., and Galecki, A.T. (2006). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.

WHO et al. (2006). *Guidelines for testing mosquito adulticides for indoor residual spraying and treatment of mosquito nets*. Technical report. World Health Organization.

Wieland, F.G., et al. (2021). On structural and practical identifiability. *Current Opinion in Systems Biology*, 25, pp. 60–69.

Xia, X. and Moog, C.H. (2003). Identifiability of nonlinear systems with application to HIV/AIDS models. *IEEE transactions on automatic control*, 48(2), pp. 330–336.

Publication I

Amadi, M., Shcherbacheva, A., and Haario, H.
Agent-based modelling of complex factors impacting malaria prevalence

Reprinted with permission from
Malaria Journal
Vol. 20, pp. 1–15, 2021.

Amadi *et al. Malar J* (2021) 20:185
https://doi.org/10.1186/s12936-021-03721-2

Malaria Journal

RESEARCH Open Access

Agent-based modelling of complex factors impacting malaria prevalence

Miracle Amadi[1*], Anna Shcherbacheva[1,2] and Heikki Haario[1,3]

Abstract

Background: Increasingly complex models have been developed to characterize the transmission dynamics of malaria. The multiplicity of malaria transmission factors calls for a realistic modelling approach that incorporates various complex factors such as the effect of control measures, behavioural impacts of the parasites to the vector, or socio-economic variables. Indeed, the crucial impact of household size in eliminating malaria has been emphasized in previous studies. However, increasing complexity also increases the difficulty of calibrating model parameters. Moreover, despite the availability of much field data, a common pitfall in malaria transmission modelling is to obtain data that could be directly used for model calibration.

Methods: In this work, an approach that provides a way to combine in situ field data with the parameters of malaria transmission models is presented. This is achieved by agent-based stochastic simulations, initially calibrated with hut-level experimental data. The simulation results provide synthetic data for regression analysis that enable the calibration of key parameters of classical models, such as biting rates and vector mortality. In lieu of developing complex dynamical models, the approach is demonstrated using most classical malaria models, but with the model parameters calibrated to account for such complex factors. The performance of the approach is tested against a wide range of field data for Entomological Inoculation Rate (EIR) values.

Results: The overall transmission characteristics can be estimated by including various features that impact EIR and malaria incidence, for instance by reducing the mosquito–human contact rates and increasing the mortality through control measures or socio-economic factors.

Conclusion: Complex phenomena such as the impact of the coverage of the population with long-lasting insecticidal nets (LLINs), changes in behaviour of the infected vector and the impact of socio-economic factors can be included in continuous level modelling. Though the present work should be interpreted as a proof of concept, based on one set of field data only, certain interesting conclusions can already be drawn. While the present work focuses on malaria, the computational approach is generic, and can be applied to other cases where suitable in situ data is available.

Keywords: Computational biology, Socio-economic factors, Agent-based modelling, Prevention of reintroduction, Long-lasting insecticidal nets, Multiscale modelling

Background

Malaria is often regarded as a socio-economic disease associated with poverty and underdevelopment. The incidence of the disease tends to decline with economic development and associated improvement in domestic conditions, such as quality of housing and availability of medical aid [1, 2]. The elimination of malaria in, for

*Correspondence: miracle.amadi@lut.fi
[1] LUT School of Engineering Science, Lappeenranta University of Technology, Yliopistonkatu 34, Lappeenranta, Finland
Full list of author information is available at the end of the article

instance, Finland was preconditioned on a drop in household size [2–4]. However, malaria imported by visitors and migrants carries the risk of re-introducing malaria in areas that have suitable vectors and climatic conditions [5]. For instance, the proportion of imported malaria cases due to migrants in Europe has recently increased from 14 to 83% [6–9]. It is therefore topical to reconsider various factors controlling the spread of malaria.

Classical compartmental models contain a limited account of the complex processes of malaria transmission dynamics, and more detailed models tend to get overloaded with model parameters that are difficult to calibrate against real data [10]. Here, an approach to alleviate this dilemma is demonstrated by a combination of individual or agent-based modelling (ABM) strategy together with compartmental modelling. The ABM approach has become popular due to its enhanced realism, flexibility, explicitness and the advantages of spatial simulations with high resolution (see [11]). An agent-based modelling approach is employed in order to simulate the impact of factors such as intervention measures, household size, and the behavioural changes of the vector. The ABM results are then linked to basic dynamic transmission models in order to enable predictions on the level of public health [12–14].

The ABM modelling is done first for a single host in the hut, and then on a household level, with multiple individuals sleeping under the same roof. Subsequently, the household-level model is extended to community-level scenarios, enabling simulations of heterogeneity of mosquito-to-human contact rates due to partial coverage with nets or different household sizes. The crucial impact of socio-economic factors such as household size has been emphasized in [2–4]. The ABM simulations provide a 'computational laboratory' where data reflecting the impact of various complex factors can be produced. Upon repeated simulations, the ABM outputs can be used as synthetic data to produce regression models for the factors considered. Here, the focus is on household size, LLIN coverage, and alterations in mosquito behaviour induced by malaria parasite.

The agent-based model simulations are conducted over a 'snapshot' time period of one night. The results can be extended to continuous time by inserting the values fitted by the response surfaces as the key coefficients of classical compartmental models. Consequently, the impact of intervention measures or socioeconomic factors can be simulated over longer time periods, and to steady state. This allows for the estimation of the EIR [15] values in a wide variety of transmission scenarios.

The work-flow followed in the present study is summarized in the schematic illustration given in Fig. 1. The modelling process is iterative as there is back and forth movement from MCMC parameter identification to ABM of mosquito host-seeking behavior, such that the model fits the data well.

Other studies have also estimated key parameter statistics from data on experimental hut trials and subsequently employed them in dynamic transmission models to enable public predictions. In the work by Churcher et al. [12], key parameters of the continuous model [16] were estimated using statistical models (such as binomial and mixed effect models) and calibrated with hut-level experiment data. Sherrard-Smith et al. [17] systematically assessed experimental hut data to characterize different indoor residual spray (IRS) product efficacies in terms of mosquito mortality, blood-feeding inhibition and deterrence against *Anopheles* mosquitoes, when fitted with statistical models. The impacts of IRS assessed from experimental hut trials are extrapolated for public health predictions in areas with different levels of coverage and pyrethroid resistance using the mathematical model of malaria transmission from [16]. Using a slightly different approach, Okumu et al. [13] directly inputted values of relevant parameters from experimental hut trials into their transmission model to make public predictions. This model additionally considered animal hosts (cattle) and predicted community-level impact on

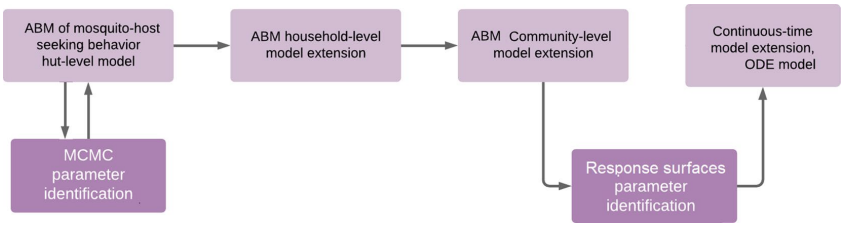


Fig. 1 Schematic representation of transition from the ABM of mosquito host-seeking behaviour to the continuous modelling. The procedure is conducted separately for each of the mosquito species and the chemicals under study

malaria transmission at high coverage (80%) using direct data from hut-level trials for various combinations of untreated nets or LLINs with IRS. Another related study [18] considered how coverage with ITNs (from 0 to 100%) influence the intensity of malaria transmission using an elaborated description of the classic feeding cycle model. The approach in this study differs from the above papers as it presents the model of mosquito host-seeking behaviour in a hut in terms of the mosquitoes' attraction to the host, host-seeking orientation, biting and death rate. The agent-based simulation of mosquito behavior at hut-level in the presence of different insecticides is then calibrated with field data from [19]. The ABM approach enables modelling of behavioral changes typical for infected mosquitoes at the household level and subsequent extension to community-level simulations using households of different sizes. Thus, upon simulating the ABM, the key ODE model parameters are created, unlike in [18] which is model-based and parameter values are mainly assumed. Additionally, the LLIN coverage and household size are elaborately considered to range from 0 to 100%, and 2 to 10 respectively. This approach enables integration of socio-economic factors and the study of malaria prevalence in a population at varied protection levels, while in [12, 13] a certain level of coverage is assumed.

The rest of the paper is organized as follows. In "Methods" section, the basic agent-based modelling approach at the hut level with a single host and its extension are presented. Next, the ABM extension to household and and subsequently to community level are described. Then, the next subsection discusses the regression applied to the outputs of community-level simulations. The extension of the response surface results to continuous time is given in "Results" section. Finally, the discussion is presented in the last section.

Methods

Basic ABM host-seeking model with a single host

A previous work [20] presented an ABM simulation approach for mosquito host-seeking behaviour on hut-level in the presence of LLIN, calibrated for one case of the treatment data from [19]. Here, the model is extended in several ways to make it capable of reproducing the data of other insecticidal treatments, and to enable the extension to continuous time modelling done in "Results" section. The basic modelling approaches utilized in [20] is briefly recalled and the modifications made in the present work is pinpointed. The model developed in this study consists of four basic components, where each of the components features a number of associated attributes, see the summary in Table 1. Additionally, the properties that are assigned individually for each of the mosquito agents and updated within the simulation (see Table 2) are listed. The model components and the property list of mosquitoes are described in detail in this subsection.

Motion and host-seeking

The mosquito attraction model is based on the assumption that a mosquito estimates the direction of odour increase (the gradient) by the mechanism of klinotaxis [21]. During this plume-tracking behaviour, the mosquito samples the host odour at one location, changes location and then repeats the sampling, and uses its memory of the concentrations previously encountered to choose the next position [22, 23]. Imitating this process, the flight of mosquitoes is modelled as a discrete-time correlated random walk. Suppose that a mosquito agent is at position $\mathbf{x}^{n-1}$ at time step $n-1$. A new position $\mathbf{x}^n$ is selected by:

$$\mathbf{x}^n = \mathbf{x}^{n-1} + \delta\mathbf{W}, \tag{1}$$

where the increment $\delta\mathbf{W}$ added to $\mathbf{x}^{n-1}$ is sampled in random direction, with a step size given by a normal distribution $N(\mathbf{x}_0, \sigma^2 I)$. In the experimental runs, the parameters $\mathbf{x_0}, \sigma$ were matched to imitate the real flight speed of a mosquito [20]. Mosquitoes are able to identify the host by making use of the olfactory cues that are given off by the host. As a primary stimuli, they move in response to the carbon dioxide (CO_2) exhaled by vertebrates. Additionally, at a short distance to the host, mosquitoes are able to discern by vision, olfaction and by using the heat sensors located around their mouthparts. In general, mosquitoes are unable to detect human prey from a distance greater than 80 m [22]. The concentration of attractive odour emitted from an individual host is modelled as a Gaussian kernel centered at a spatial position of the host $\mathbf{x}^h$:

$$C(\mathbf{x}, \mathbf{x^h}) = \exp\left[-\frac{d^2(\mathbf{x}, \mathbf{x}^h)}{2\sigma_a^2}\right], \tag{2}$$

where $\mathbf{x}$ denotes the position of the mosquito, and C stands for the concentration that enables a mosquito to sense the host at a distance $d(\mathbf{x}, \mathbf{x}^h)$. Note that the impact of wind is omitted for simplicity. The standard deviation of the Gaussian σ_a determines a maximal distance at which the mosquito is able to sense the host.

The mosquito flight is given by the above random walk in the absence of attraction effects towards the host. However, when there are attraction effects, the main features of the Metropolis algorithm is employed in order to simulate the random walk directionally biased by attraction [24]. Suppose that a mosquito takes a step from point $\mathbf{x}^{n-1}$ to a candidate point $\mathbf{x}^n$ with respective function values as p_{n-1} and p_n. Then a new point is accepted with probability:

$$\alpha_a(\mathbf{x^n}|\mathbf{x^{n-1}}) = \min\left(1, \frac{p(\mathbf{x^n})}{p(\mathbf{x^{n-1}})}\right), \quad (3)$$

where $p(\mathbf{x^n})/p(\mathbf{x^{n-1}})$ is the ratio of the attraction potential function $p(\mathbf{x})$ defined at each point $\mathbf{x}$, which depends on the concentration and other attraction factors. In order to parsimoniously account for other short-distance attraction factors, the attraction potential function is defined as:

$$p(\mathbf{x}) = \exp\left(C(\mathbf{x})/\sigma_{acc}\right) \quad (4)$$

with a scaling factor σ_{acc} that depends on the distance to the host. Outside the plume $p(\mathbf{x}) = 1$, so by Eq. 3 all steps are accepted, while closer to the host steps away from the host are increasingly rejected due to activation of the heat sensors. At a short distance to the host this is modelled by a linear scaling factor as:

$$\sigma_{acc}(\mathbf{x}, \mathbf{x}^h) = \begin{cases} \sigma^1_{acc} + \sigma^2_{acc} d(\mathbf{x}, \mathbf{x}^h), & d(\mathbf{x}, \mathbf{x}^h) \leq 80 \\ \sigma^{\max}_{acc}, & d(\mathbf{x}, \mathbf{x}^h) > 80. \end{cases} \quad (5)$$

The above function increases from the minimum value of σ^1_{acc} with a slope given by the parameter σ^2_{acc} until it is replaced by a constant which suitably provides a purely random movement outside the concentration plume [20].

Death, poisoning and repellency

The LLINs are assumed to be equipped with repellent and poisoning effects. In the absence of chemical treatment, the total probability of death reduces to the natural mortality rate. The continuous-time mortality rate μ can be transformed into a probability of death per unit time Δt by:

$$\alpha^{\Delta t} = \min\left\{1, \mu \Delta t\right\}, \quad (6)$$

where $\Delta t = 2$ s is used for all simulations, and a value for μ taken from the literature (see [20] for more details). This conforms with the 34-h natural mortality rates reported for *Anopheles gambiae* and *Anopheles arabiensis* as 10%, see [25] .

The poisoning effect is modelled with the assumption that at a time instance i, mosquito consumes a dosage of chemical D_i spread on the treated net upon contact to the net surface. Thus, the total accumulated dosage C_{tot} is computed as the number of contacts with the net:

$$C_{tot}(n+1) = \sum_{i=1}^{n+1} D_i = C_{tot}(n) + D_{n+1}, \quad (7)$$

where D_i is non-zero in case of hitting the net surface (i.e., equal to the unit dosage), and zero otherwise.

The insecticidal-induced increase in mortality is then modelled as:

$$\alpha_p^{\Delta t}(n) = \mu_p C_{tot}(n) \Delta t, \quad (8)$$

where the effective poisoning impact is obtained by a scaling coefficient μ_p which depends on the given insecticide used for LLIN treatment.

So the total probability of death per unit change in time Δt is modelled as the sum of natural and insecticide-induced mortality:

$$\alpha_{death} = \min\{1, \alpha^{\Delta t} + \alpha_p^{\Delta t}(n)\}. \quad (9)$$

Repellency is modelled with the logistic curve multiplied with the repulsion intensity parameter r:

$$C_{rej} = r\left[1 - 1/\left(1 + \exp\left(-\left(d(\mathbf{x}, \mathbf{x}^h) - d_{50}\right)/s\right)\right)\right], \quad (10)$$

where $d(\mathbf{x}, \mathbf{x}^h)$ denotes the distance from the mosquito to the protected human and r ranges from 0 to 1. The parameters d_{50} and s determine the range of coverage and the spread of the chemical. The logistic function is modified such that the rejection probability at the candidate position $\mathbf{x}$ amplifies as the mosquito approaches the source of repellent. Considering the properties of modern insecticidal treatments [26], the spatial range of the repellent s is taken to be small such that the impact is only within the vicinity of the net.

The repulsion by LLIN is computed in two stages. First, the accept/reject step is applied, where the probability of rejection is given by a logistic function describing the contact irritancy caused by the chemical, as given in Eq. 10. Next, the physical net barrier is taken into account, for which the probability of being blocked by the net is assigned as $p_{net} < 1$ such that there is a non-zero chance for penetration.

Model extensions

Motion and host-seeking: excito-repellency

The aim is to keep the host-seeking model as minimalistic as possible, by including only the indispensable factors listed in Table 1. It turns out, however, that the impact of different chemicals could not be fitted by the basic formulation given above. For instance, the model has to reproduce cases of higher exit and lower contact rates along with more than twice higher mortality rate for *An. gambiae* than *An. arabiensis*, following the data reported in [19]. Three new features necessary to characterize the impact of different chemicals on mosquitoes: metabolic detoxification [27, 28], delayed impact [19] and excito-repellency (or insecticide-induced exiting) [19, 29], are introduced. In order to account for insecticide-induced

exiting, a scaling factor which not only depends on distance but also on repellent effect is further obtained. The inclusion of both distance and repellent effect is essential in order to properly fit the exit rates, as it accounts for generally higher exit rate when confronted with the treated nets as compared to the control case with the untreated nets. Thus, an excito-repellency parameter [29], μ_e is introduced, which depends on the mosquito species and the insecticide utilized in treating a given LLIN, parameterized as:

$$\sigma_{acc}(\mathbf{x}, C_{tot}) = \sigma_{acc}(\mathbf{x}) + \mu_e \cdot C_{tot}, \tag{11}$$

where C_{tot} denotes the total dosage of chemical consumed by the mosquito (see Eq. 7).

The other two included features: metabolic detoxification (see Eq. 12) and delayed mortality are explained next.

Poisoning and death: detoxification and delayed death rate

Here, the scenarios in the datasets from [19], where *An. arabiensis* is revealed to have consistently higher (or equal) feeding rate than *An. gambiae* but considerably lower death rate, are accounted for. These scenarios are inconsistent with the mechanism of the model presented in [20]. The inconsistency is explained by the fact that it is not possible to have simultaneously high feeding rate and low mortality rate if both the probability of death and that of successful feeding is proportional only to the number of contacts with the net. A number of probable reasons can be offered to account for the conflicting situation. One explanation is that the rate of poisoning is different for the two species because it takes time for the poison to get from the salivary glands to the neural system of mosquito and this time delay is suspected to be different for the two mosquito species. However, a large dosage is equally lethal for both *An. gambiae* and *An. arabiensis* and mosquitoes do not acquire the lethal dosage upon a single contact with the net but rather a sub-lethal dosage [19]. So, the explanation of detoxification is followed such that the chemical concentration is exponentially decaying with a rate α which depends on the chemical and mosquito species [28, 30]. Hence, given the previous dosage of the chemical $C_{tot}(n)$ at the step n, the dosage at the next step $n+1$ is calculated by modifying Eq. 7 as:

$$C_{tot}(n+1) = C_{tot}(n) + D_{n+1} - \alpha C_{tot}(n)\Delta t. \tag{12}$$

Additionally, the delayed mortality that is a result of the prolonged impact of poison in mosquitoes is considered. Since poisoning effect is primarily associated with contact with the treated surface, some time is needed for the chemicals to penetrate and reach their target, which in turn depends on the physiological characteristics of the mosquito, such as the sensitivity of target proteins and the thickness of the cuticle [27]. Also, due to enzymatic detoxification, the knock-down time is prolonged. Owing to the high exit rates reported in [19], it was concluded that the mortality induced by the insecticides occurred only after a delay. Although the mosquitoes respond differently with different chemicals, the detailed modelling is spared and the enhanced probability of death is simply taken into account only after a 24-h time period as given by Eq. 6, with $\Delta t = 24 \cdot 1800$.

The improved model of the chemical-induced exiting and mortality introduced can be calibrated for all the different treatment kits data from [19] (see Additional file 1 for a summary of the datasets). The model is capable of reproducing, e.g., the experimentally recorded lower contact rates along with more than twice higher mortality rates for *An. gambiae* as compared to *An. arabiensis* [19]. The calibration is performed using Bayesian sampling methods (adaptive MCMC) in the same way as in [20], more the details are given in Additional file 1. The motive of the MCMC simulations is to find the posterior distibutions of model parameters, that is 'all' parameter combinations that reproduce the measured data, within the accuracy given by the estimated error bounds of the data. While most of the parameters are reasonably well identified, some of them are clearly correlated. For instance, as the chemically enhanced mortality rates are now explained by both detoxification and exito-repellency, the respective parameters are strongly correlated with μ_p, the earlier introduced death rate coefficient.

Household-scale simulations: household size effect and behavioral alterations

Here, the description of the household and community level modelling is presented, adding more details to the preliminary demonstration given in [20]. First, the ABM of mosquito host-seeking behaviour is extended to the household level with multiple individuals sleeping under the same roof. Next, the modelling is extended to community-level scenarios with several households located in the landscape of interest. See the illustration of the workflow in Fig. 1.

A significant correlation between malaria reduction and the decline in typical household size in malaria-endemic countries is discussed in [2–4]. It was concluded that the larger the number of people sleeping together in non-segregated quarters, the higher the probability of transmitting the infection to new uninfected humans [2]. In Finland, for instance, the probability of malaria disappearance increased when the average number of individuals in one household declined below the threshold of four people, even when no specific control measures were applied [2, 4].

Naturally, there are several other household-related factors that can influence the rate of transmission apart from the household size. Such factors include, e.g., household practices like livestock/poultry rearing, as well as the rate of hygiene maintenance in a given household [31]. For simplicity, these factors are omitted here. The interest of the ABM simulations is in the impact of LLINs. The mosquito density m which can be impacted by these omitted factors, is taken into account in the ODE model. Also, the situation is restricted to a given number of persons sleeping together in the same room, while the approach can be extended also to cases of many people sleeping in separated quarters. The aim here is to demonstrate how household-level factors can be included in ABM simulations, and how even most rudimentary considerations impact the modelling outcomes.

On entering a household, the mosquito is assumed to choose one of the hosts randomly. After this, the modelling reduces to the previous case of a single host in the hut. A few changes are needed, however. The tendency of mosquitoes to switch to neighbouring individuals after spending a certain time in unsuccessful attempts to feed on a protected host, should be considered. Thus, an additional parameter, t_{max}^{host}, the maximal time spent while attempting to feed on a protected host, is introduced [32, 33]. In the absence of more specific knowledge, the parameter is set to 10 min. In addition, same as in the hut-level experiment, mosquitoes are restricted to a maximum host-seeking time, t_{max} inside the household after which they switch to a random walk with no influence of the human bait. Another difference is easier exit from a usual household compared to that from the special design of experimental huts. A typical human dwelling [34] is mimicked by setting the probability of exit to constant value that produces about 90% exit rates per night in the absence of chemical treatment.

Infection with malaria parasites has been shown to alter the behaviour of mosquitoes, with varying effects that are based on the life stage of the parasite [35]. The underlying mechanisms that engender these behavioural alterations are not fully explored but mostly result from at least two manipulation processes. Firstly, the parasite increases the mosquito's motivation to continue a meal after interruption, thus increasing its probability of taking several bites. Secondly, the parasite impairs the vector's ability to obtain a full blood meal upon a single bite, inducing the vector to bite several times before it is fully engorged [36, 37]. These behavioural changes associated with infection seem likely to be an evolutionary mechanism that has been developed by malaria parasites, which enhances the spread of infection [38, 39]. A more profound understanding of the behavioural tendencies of parasite-infected mosquitoes alongside the stage-specific changes in their host-seeking behaviour could provide a potential target for genetic manipulation of mosquitoes, as a preventive measure for the elimination of malaria infection [40].

In the simulation, the impact of multiple biting typical for infected mosquitoes is accounted for. Both infected and uninfected mosquitoes are assumed to have the tendency of feeding on multiple hosts [41]. However the tendency of multiple feeding is higher for infected mosquitoes. Thus, the statistics from [36] is employed, which indicate that 10% of uninfected and 22% of infected mosquitoes obtain a blood meal on at least two hosts, while assuming that the maximal number of successful feeding attempts can be up to 5 for both, depending on the accessibility of the hosts. The dosage of blood sufficient for ovipositing is assumed to be achieved after the maximal number of successful feeding attempts is reached. Note that the hut-level data, with one person in the hut, does not contain information on the alterations in behaviour during the host-seeking, so at this point, the literature is relied on. On the other hand, the conjecture that humans infected by the parasite attract more mosquitoes [42] is not included in the simulations, since the hypothesized enhanced attractiveness has demonstrated insignificant impact on the outcome of the simulations (see [43]).

Note that in the simulations, the model parameter values are also re-sampled from the estimated parameter posteriors at each successive iteration of the algorithm to account for parameter uncertainty (see Additional file 1). The main model parameters are summarized in Table 3.

Community-scale simulations

Next, the modelling is extended to a community-scale experiment with the primary aim of quantifying the effect of household size and a partial population coverage with LLINs (see Fig. 1). Similar to the hut-level case, the movement of mosquitoes in the odour plume is governed by the mechanism of klinotaxis, but the concentration which enables the mosquitoes to sense the hosts is now computed as a function of a weighted sum of distances from all the individual hosts:

$$C_a^{tot}(\mathbf{x}) = C(W_n, \mathbf{x}, \mathbf{x}_n^h) = \exp\left[-\left(\frac{\sum_{n=1}^{N_h} W_n d(\mathbf{x}, \mathbf{x}_n^h)}{\sqrt{2}\sigma_a}\right)^2\right]. \quad (13)$$

Here N_h denotes the total number of individuals in the community, $d(\mathbf{x}, \mathbf{x}_n^h)$ stands for the distance from mosquito position $\mathbf{x}$ to the host location $\mathbf{x}_n^h$, and W_n is the weight attributed to the host n.

The total attracting concentration is modelled following the idea of the *softmax* function, which has been widely adopted in machine learning and neural networks (see [44, 45]). The weight W_n is introduced to account for the fact that a mosquito's response to the cue emitted from the households increases at a short distance of 5–15 m, depending on the mosquito species, due to their attraction to visually conspicuous objects [46–48]. Here, the main focus is placed on the nearest target concept, which practically means that at a short distance factors other than just the CO_2 alone also cause the mosquito to localize the search, as reported in [46–48]. Following this reasoning, the non-normalized weights $\hat{W}_n$ are introduced inversely proportional to the distance:

$$\hat{W}_n(\mathbf{x}, \mathbf{x}_n^h) = (1 - 1/\exp(-(d(\mathbf{x}, \mathbf{x}_n^h) - d_{50}^h)/s^h), \quad \mathbf{x}_n^h \in \mathbf{x}_n^h. \tag{14}$$

The value of d_{50}^h is set to be 10 m, to conform with the conjecture that at a distance of less than 10 m from the households, within which a mosquito is able to discern shapes, the concentration sensed by the mosquito is assumed to be that which is emitted from the closest household only. The second parameter $s^h > 0$ governs the spatial range of sensitivity that enhances at a short distance. Here, the value $s^h = 5$ m is used to account for the gradual boost of the mosquito's response to the cues. The weights $\hat{W}_n(\mathbf{x}, \mathbf{x}_n^h)$ are normalized by $W_n = \hat{W}_n / \sum_{j=1}^{N_h} \hat{W}_j$.

Note that the form of Eq. 13 is consistent with the evidence that larger agglomerates emit stronger odours, hence, attracting more mosquitoes [23] (see the illustration in Fig. 2).

Environmental factors such as wind and intermittency of the plume are omitted for simplicity. Initially in the simulations, mosquitoes are randomly placed inside the simulated transmission domain of 25,600 m^2 size with multiple households located at a distance not closer than 40 m from one another such that there is no competitive attraction induced by vision [46]. A constant number of 700 mosquitoes and around 20 individuals are used for each experimental run. To average the stochasticity due to spatial arrangement, households are randomly positioned at each successive repetition of the algorithm. Within a single run, all the households are of the same size. However, the household size varies between the runs. Seven repetitions are conducted for each of the runs to reduce the noise in the outputs. Figure 3 presents the randomly generated experimental layout.

The number of infectious mosquitoes is constant for a single experiment (since it takes a period of 10 to 12 days for parasites to reach a stage whereby they are ready for transmission). In the case when an insufficient amount of blood was consumed before the exit from the household,

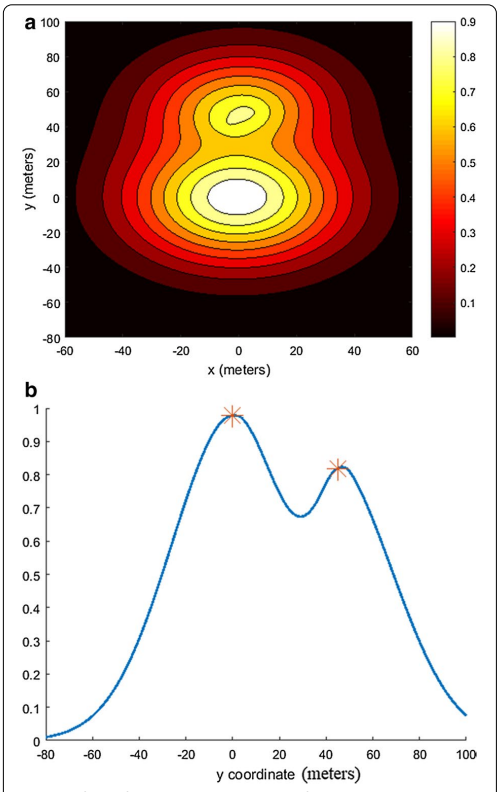


Fig. 2 Softmax function in a special case of two households. The first household includes six individuals (located (0,0)) and the other household consists of two individuals (located at (0,45)) for $d_{50} = 10, s = 5$ **a** 2D plot, **b** 1D plot along the y axis

Table 1 Model components

Model component	Attributes	Definition
Host-seeking	CO_2 concentration, Klinotaxis	Equation 2
	Distance-dependent attraction	Equation 5
	Host seeking time	
Motion	Random walk, accept/reject steps	Equation3
	Excito-repellency	Equation 11
Poisoning	Accumulation of the chemical dosage	Equation 7
	Detoxification	Equation 12
Death	Natural mortality	Equation 6
	Insecticide-induced mortality	Equation 8
	Delayed mortality	Equation 6 with model extension

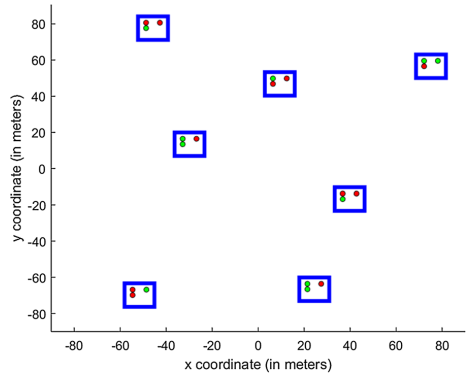


Fig. 3 Randomly generated experimental layout with household size of three individuals. Here blue rectangles denote the houses, green/red circles mark individuals protected/non-protected with the impregnated nets

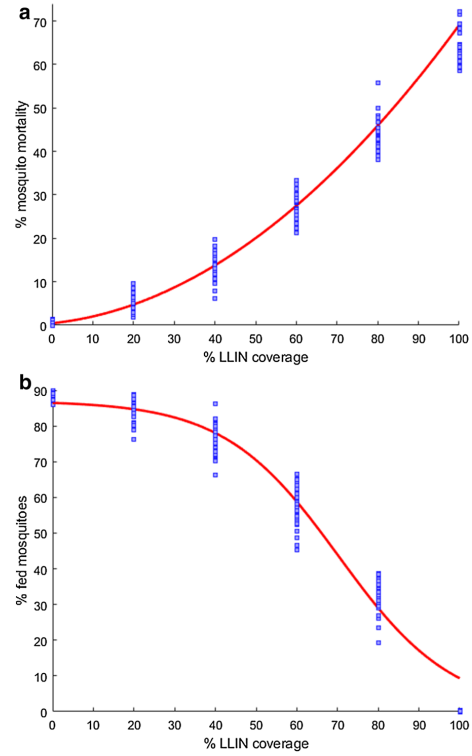


Fig. 4 Uncertainty from the sampled parameters at hut level together with the variability of the community-level assumed parameters for **a** mortality rate, **b** fed rate, of *An. gambiae* when confronted with LLIN impregnated with an Alphacypermethrin treatment kit, fitted with respect to partial coverage of LLIN for the household size of 2 when assuming no behavioural alterations caused by the parasite

the mosquito starts the process of host-seeking (from the outset) except that the abandoned household is not accounted for when computing the total concentration of the CO_2. Additionally, it is assumed that after entering a new household, the count of host-seeking time t_{max} is reinitialized.

Regression analysis of community-scale simulations

A final step of using the ABM results is to generate regression functions based on the main trends revealed by the ABM simulation results. The effects of the in situ behaviour, settlement patterns and parasite ecology are explored by fitting the response surfaces to the trends given by the simulations. That is, ABM is used as a 'computational laboratory' to produce data for response surfaces that capture the impact of the LLIN coverage and household size. The ABM simulations are inherently stochastic, due to the event generation by randomizing. In the community level, the uncertainty from sampled parameters at hut-level are included and a sensitivity analysis is conducted with respect to the assumed parameters using a central composite design. The assumed parameters were varied reasonably based on literature values as shown in Tables 4 and 5. The sensitivity analysis shows that the behavior of the system remains more or less the same with reasonable perturbations in the assumed parameter values. For illustration, the outputs with variability from both sampled parameters at hut level and the assumed parameters are presented, for contact and mortality rates of *An. gambiae* when confronted with LLIN treated with Aphacypermethrin chemical in Fig. 4. The outputs of the ABM simulations are averaged over 7 repetitions of the experiment. These number of repetitions was found to be sufficient by an extensive preliminary simulation. (see Additional file 1 for more ABM community-level simulation results).

A regression analysis is applied with respect to the household size and the coverage with LLINs, using the synthetic data. Given that one of the independent variables is discrete by definition, a uniform design is employed, considering household sizes of 2, 4, 6, 8 and 10, and with LLIN coverage varying from 0 to 100%. The regression is conducted in two cases: when assuming no behavioural alterations and when considering alterations caused by the parasite separately for *An.*

gambiae and *An. arabiensis* when confronted with each of the chemical treatments considered.

The ABM simulation data revealed a nonlinear, quadratic relationship between the mortality rate and LLIN coverage, but an insignificant dependence of mortality rates on the household size. Consequently, the mortality rates are fitted with second degree polynomial with respect to the coverage only, see Fig. 4 for an example. Comparing the impact of the chemicals, it can be seen from Fig. 5 that in case of *An. gambiae*, Carbosulphan is the most efficient, while the other treatments display similar performances. For *An. arabiensis* the highest impact is with IconMaxx, followed by Carbosulphan. Alphacypermethrin treatment induces the lowest mortality for *An. arabiensis* of all the studied chemicals. In the case of behavioural alterations the mortality rates are similar to Fig. 5, although slightly higher, which apparently results from more frequent exposure to insecticide due to a higher number of feeding attempts.

The contact rates showed a dependency on both the household size and LLIN coverage. Moreover, for both uninfected and infected mosquitoes, the respective contact rates $\tilde{a}$ and $\bar{a}$, displayed logistic behaviour with respect to the coverage x_2. A certain coverage threshold was required for the contact rate to start decreasing. Hence, the logistic functions is used:

$$\tilde{a}(x_1,x_2) = N_b * (1 - 1./(1 + \exp(-(x_2 - b_1 - b_2x_1)/b_3))) \quad (15)$$

$$\tilde{a}(x_1,x_2) = N_b * (1 - 1./(1 + \exp(-(x_2 - b_1 - b_2x_1)/b_3))), \quad (16)$$

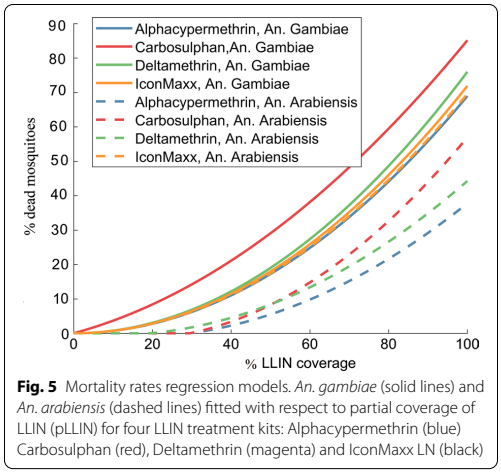


Fig. 5 Mortality rates regression models. *An. gambiae* (solid lines) and *An. arabiensis* (dashed lines) fitted with respect to partial coverage of LLIN (pLLIN) for four LLIN treatment kits: Alphacypermethrin (blue) Carbosulphan (red), Deltamethrin (magenta) and IconMaxx LN (black)

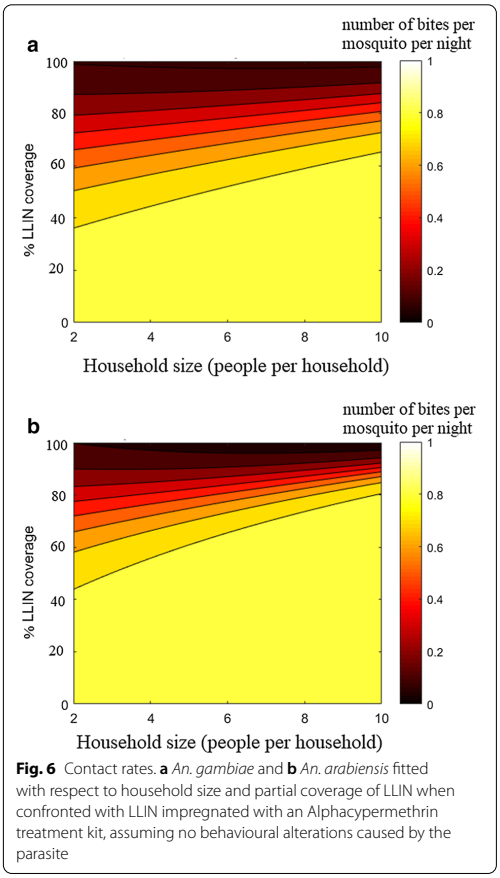


Fig. 6 Contact rates. **a** *An. gambiae* and **b** *An. arabiensis* fitted with respect to household size and partial coverage of LLIN when confronted with LLIN impregnated with an Alphacypermethrin treatment kit, assuming no behavioural alterations caused by the parasite

Table 2 Property list of each agent and the relevant model component

Property index	Property	Model component
1	Spatial position	Motion
2	Inside/outside the hut	Motion
3	Inside/outside the net	Motion
4	Trapped	Motion
5	CO_2 concentration	Motion
6	Fed	Host-seeking
7	Time indoors	Host-seeking
8	Klinotaxis	Host-seeking
9	Dead	Death (Poisoning)
10	Accumulated dosage of chemical	Poisoning

where x_1 and x_2 denote the household size and the fraction of LLIN coverage. The values of the parameters N_b, b_1, b_2, b_3 are obtained from the regression fits to the ABM data, separately for each chemical. Figure 4 gives an example for one of the chemicals.

The results suggest that Alphacypermethrin displayed the highest efficiency in reducing the contact rate when applied to *An. gambiae*, while all the other chemicals demonstrate similar reduction effects. This is consistent with the confidence intervals given in [19]. On the other hand, all the chemicals feature similar performance in reducing the contact rate, in the case of *An. arabiensis*, with slightly better efficiency attributed to IconMaxx LN. Moreover, unlike the other treatments, Alphacypermethrin displays substantially better performance in the reduction of contact rates for *An. gambiae* compared to *An. arabiensis*, as can be seen from Fig. 6. The other chemicals demonstrate similar protection against both mosquito species with slightly lower contact rates when applied to *An. gambiae*.

In both cases, the contact rate tends to increase with household size for the fixed rate of LLIN coverage. Thus, the regression analysis results provide convincing evidence that lower LLIN coverage is sufficient to achieve similar reduction in contact rates for smaller household sizes, see Figs. 6, 7. For more details of the regression analysis, see Additional file 1.

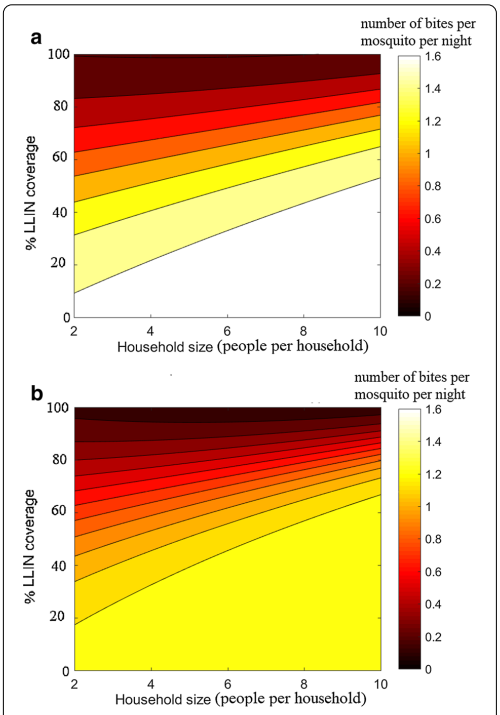


Fig. 7 Contact rates. **a** Infected and **b** uninfected *An. arabiensis* when confronted with LLIN impregnated with an Alphacypermethrin treatment kit, fitted with respect to household size and partial coverage of LLIN (pLLIN) when assuming behavioural alterations by parasite

Results

Extension to continuous time, EIR

Here, the ABM of mosquito host-seeking behaviour is linked to continuous-time compartmental modelling. This connects in situ mosquito behaviour to commonly measured quantifiers of malaria transmission, such as Entomological Inoculation Rate (EIR) and malaria incidence. A benchmark test for classical malaria models was conducted in [10], where five dynamic models of malaria transmission were tested on the basis of established performances. The most basic Ross malaria model was found to be capable of reproducing the EIR experimental data satisfactorily. Indeed, more complex models tend to suffer from poor identification of parameters and may produce results inferior to simple but more robust modelling. Following [10], the simple Ross model is considered, but utilized such that the complex factors (such as the LLIN coverage, household size or alterations of behaviour) are expressed via the ODE model parameters. That is, the regression functions from the previous section for the contact and mortality rates are substituted in place of the respective parameters in the Ross model:

$$\begin{aligned} di_h &= m\bar{a}bi_m(1-i_h) - i_h r \\ di_m &= \tilde{a}ci_h(1-i_m) - \mu i_m, \end{aligned} \tag{17}$$

where i_h and i_m denote the fractions of infected humans and mosquitoes, correspondingly, m stands for mosquito-to-human ratio, b and c are the probabilities of transmission during mosquito contact with the host, μ denotes the mosquito mortality rate, and r stands for recovery rate for the humans. The difference to the conventional Ross model is also that the contact and death rates $\bar{a}, \tilde{a}$ and μ are given by the response surfaces, fitted to various in-situ conditions. Indeed, two different contact rates, for infected $\bar{a}$ and uninfected $\tilde{a}$ mosquitoes are used in the case when alterations in mosquito behaviour is assumed. For the rest of the parameters, m, b, c, r, three sets of values were borrowed from [10], corresponding to low, medium and high transmission settings, see Table 4. The integration of the Ross model is done for household size comprising 2, 4, 6, 8 and 10 individuals while applying 20, 40, 60, 80 and 100% LLIN coverage for each household size considered.

The quantities of interest are the equilibrium fractions of infected mosquitoes and humans. Note that the units for mortality and contact rates are the same in both the ABM and Ross model, given as a fraction of mosquito population subject to mortality (feeding) per day. The contact rate is understood here as the average number of bites taken by the mosquito diurnally.

The mosquito-to-human ratio m is taken as a ratio of the number of humans to mosquitoes N_m/N_h, as given in [10]. Each value of m is combined with the three sets of the other parameters in Table 6, so nine pairs of equilibrium values of fractions of infectious humans i_h^* and infectious mosquitoes i_m^* were calculated. For each case, the response surfaces with respect to household size and LLIN protection can be now calculated.

The most direct approach for estimating the overall malaria transmission in a population is by computing the Entomological Inoculation Rate (EIR) [15]. EIR is commonly measured to quantify the intensity of an infected mosquito pool and its propensity to transmit malaria infection to human populace within a given time period. Conventionally, the EIR is measured per period of time: per night, monthly, seasonally or annually. The transmission patterns represented by the pair of EIR and Parasite Rate (PR) depend on a number of ecological, climatic and socioeconomic factors [49]. Here, the simulation results are compared to the experimental results reported in [50], where a trend curve together with the 95% confidence interval was created using data from 31 sites in Africa. At the time when the survey was published, the annual EIR varied from less than 1 to more than 1000 infective bites per person per year. The transmission patterns represented by the pair of EIR and Parasite Rate (PR) depend on a number of complex factors, such as ecology, climate and socioeconomic development [49]. By integration of the modified Ross model, the impact of partial population coverage with LLIN, alterations in mosquito behaviour and household size, on EIR and PR, can be quantified. These two factors are computed from equilibrium fractions of infectious mosquitoes i_m^* and humans i_h^*, i.e., by the steady state of the Ross model. As EIR is defined as the product mai_m^*, a direct computation gives:

$$EIR = m\bar{a}i_m^* = \frac{\bar{a}\tilde{a}bcm - \mu r}{\mu b + \tilde{a}bc},$$

where

$$i_m^* = \frac{\bar{a}\tilde{a}bcm - \mu r}{\mu m\bar{a}b + \bar{a}\tilde{a}bcm}.$$

As a result, three pairs of equilibrium EIR and malaria incidence i_h^* correspond to each of the original selections of parameters given in [10], see Fig. 8. In addition, the EIR and PR values for those LLIN and household values for which the regression models of $\bar{a}, \tilde{a}, \mu$ were calibrated, can now be computed. These values, as continuous functions of LLIN, are added in Fig. 8. Figure 9 gives an example of the response surface of the EIR values as a function of household size and LLIN coverage. Respective figures for all the chemicals are given in Additional file 1.

Note that all the results presented in this paper are based on data from [19]. Additional experimental data would improve the reliability of the results, especially for the behaviour of mosquitoes between the households. Given that similar data are available elsewhere, the approach allows general trends and response surfaces to be produced based on such data in an analogous way.

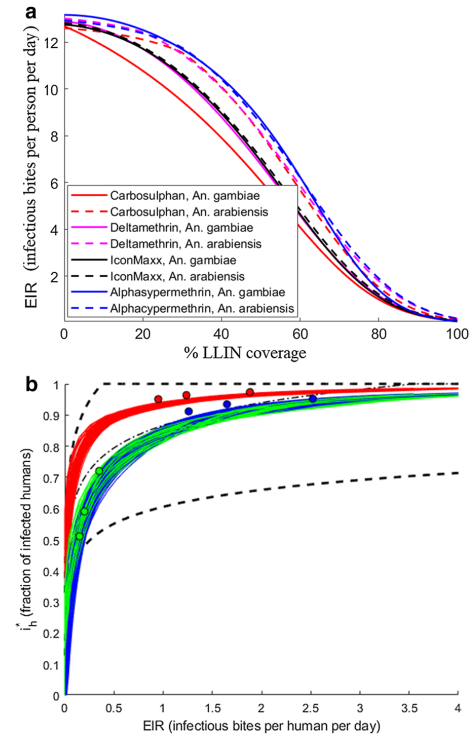


Fig. 8 Equilibrium EIR values, assuming behavioural alterations for household size 4. **a** Equilibrium EIR values as continuous functions of LLIN coverage. **b** Equilibrium values of malaria prevalence i_h^* versus EIR, in comparison to the mean trend (black dash-dotted line) and 95% confidence bounds (black dashed lines) as given in [50]. High (red circles), medium (blue circles) and low (green circles) transmission settings from [10], together with simulated values

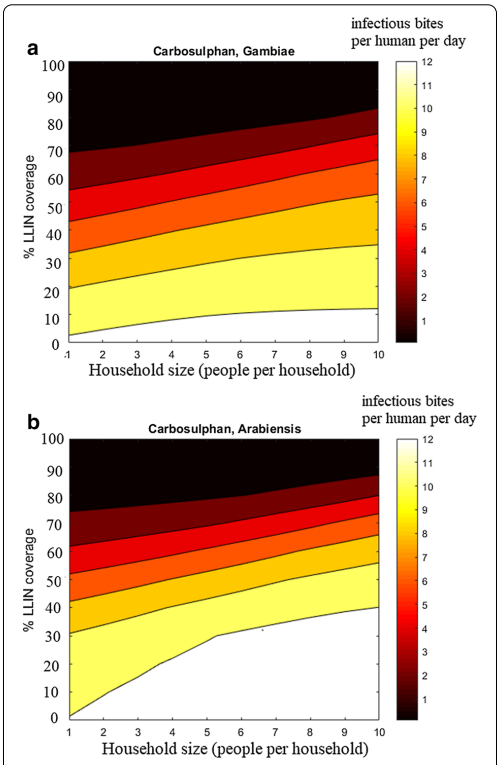


Fig. 9 Predicted equilibrium EIR conditioned on household size and the fraction of LLIN coverage, **a** *An. gambiae* **b** *An. arabiensis* when confronted with Carbosulphan

Discussion

Transmission of diseases depends on complex factors, medical, environmental or socio-economic, to mention a few. A serious issue of simulating such processes by traditional population-level compartmental models is the calibration; the models tend to get very complex, overloaded with unidentifiable parameters. The situation is typically made worse by the scarcity of real data needed for the calibration.

An approach to combine complex *in situ factors together with classical compartmental models, in the case of Malaria transmission, is presented*. The idea is to simulate the individual level processes by discrete ABM calculations under varying conditions for the factors considered. The resulting data is used as input for regression to quantify the impact of the factors as response functions. The key coefficients of a compartmental model can then be expressed by these functions.

Naturally, the underlying ABM model needs to be carefully calibrated. This is only possible if sufficient in situ data is available, and the ABM model is parsimonious enough. Such a model is developed and is calibrated using extensive MCMC (Markov chain Monte Carlo) simulations against a set of field data. The simulations can be extended to community level to study the impact of intervention levels and basic socio-economic factors. It appears that even if all the ABM model parameters are not well identified, the randomized simulations provide consistent trends with respect to the factors studied: the LLIN coverage, various chemicals, household size and behavioural changes of infected mosquitoes.

While the present work should be interpreted as a proof of concept, based on one set of field data only, certain interesting conclusions can already be drawn. A lower LLIN coverage is sufficient for smaller household

Table 3 Summary of the basic agent-based model parameters, [20]

Parameter symbols	Parameter
p_{net}	Probability of being blocked by the physical barrier created by the net
p_{hut}	Probability of exiting the hut
d_{50}	Range of repellent coverage
μ_p	Insecticide-induced death rate
r	Intensity of repulsion
t_{max}	Maximum host-seeking time (when confronted with the LLIN)
μ_e^G	Rate of increase of excito-repellency for *An. gambiae*
μ_e^A	Rate of increase of excito-repellency for *An. arabiensis*
α_G	Detoxification rate for *An. gambiae*
α_A	Detoxification rate for *An. arabiensis*
t_{max}^{host}	Maximal time (in minutes) spent
	Attempting to feed on protected host

Table 4 Sensitivity design table

Range	d_{50}	s_h	t_{host}^{max}	σ_a
Minimum	2.7058	1.0718	0.8756	40/3
Maximum	18.2942	14.9282	25.1244	80/3

Table 5 Sensitivity design table for the behavioral alteration

Range	Uninfected	Infected
Minimum	1.8934	5.8579
Maximum	23.1066	34.1421

Table 6 Summary of parameter selections and mosquito densities *m* from [10] used for integration of the Ross model

Parameters			
b, c, r	0.2, 0.5, 0.01	0.03, 0.275, 0.0035	0.4, 0.4, 0.05
Quantity	High transmission	Medium transmission	Low transmission
m	7.6	5.5	4.0

sizes in order to attain a certain reduction in the biting rate. The contact rates are higher when assuming behavioural alteration, but with high LLIN coverages the contact rates become virtually the same, i.e., the effect of alterations in mosquito behaviour due to the presence of the *Plasmodium* parasite becomes negligible. The difference between mosquito species is evident as well. The coverage required to achieve similar reduction in the number of infectious contacts is higher for *An. arabiensis* than *An. gambiae*, basically due to the lower death rate of *An. arabiensis*. The death rates of both species increase when considering the alterations in behaviour. An intuitive explanation is the more intensive exposure to insecticide for infectious mosquitoes, due to increased attempts to feed on multiple hosts during the night.

Different values of the Ross model parameters can result in the same EIR values, which prevents the identification of transmission factors based on EIR data alone. The agent-based model gives an approach which incorporates the in situ data with contact and mortality rates. So the overall transmission characteristics can be estimated by including various features that impact the EIR and malaria incidence, e.g., by reducing the mosquito–human contact rates and increasing the mortality through control measures or socio-economic factors. Additionally, local characteristics can be combined with spatially explicit model that accounts for heterogeneity in human and mosquito distribution, see [23].

The present study can naturally be extended in several ways. In addition to *An. gambiae* and *An. arabiensis*, other mosquito species can be be considered, as well as intervention methods other then LLINs. Here, constant values are assumed for the mosquito density *m*, although it actually is seasonally varying due to rainfall and temperature. Spatial aspects like the local disposition of mosquito-breeding sites can be included by calibrating the respective parameters to be site dependent. This way, the modelling can be extended to larger geographical areas. The mosquito–human contacts with an infected mosquito are assumed here to be equally infectious, whereas some people might have acquired partial immunity either by constant exposure to the parasite or by artificial means via vaccines [51]. Thus, the impact of naturally acquired immunity can be incorporated in the model by treating the hosts as a population of agents as well, and making the transmission parameter *b* dependent on the individual immunity level. Also, the present study is restricted to night-time in-house biting scenarios. The model can be improved to include outdoor biting scenarios [52]. All such extensions are technically feasible but require sufficient field data for a robust calibration of an underlying ABM model. Under this condition, agent-based models are capable of generalizing various effects from the in situ level to continuous modelling.

Conclusions

The common pitfall of obtaining data that could be directly used for model calibration in malaria transmission modelling, may be overcome by linking in situ field data with continuous malaria models. Thus, complex phenomena such as the impact of the coverage of the population with long-lasting insecticidal nets (LLINs), changes in behaviour of the infected vector and the impact of socio-economic factors can be included in continuous level modelling. The computational approach is generic, and can be applied to other cases where suitable in situ data is available.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12936-021-03721-2.

Additional file 1. From in situ to continous model. This file contains more detailed explanation of the data and likelihood, model calibrations, Regression with simulated ABM outputs, the extension to continuous times and a detailed description of the ABM using the ODD protocol.

Acknowledgements
We acknowledge the useful discussions with Gerry Killeen of Ifakara Health Institute, Tanzania.

Authors' contributions
HH conceived the project and supervised the work undertaken for this manuscript. MA, AS and HH contributed to the modelling approach. MA and AS participated in running the simulations. MA drafted the manuscript. MA, SA and HH edited the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Centre of Excellence of Inverse Modelling and Imaging (CoE), Academy of Finland, Decision number 312122.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] LUT School of Engineering Science, Lappeenranta University of Technology, Yliopistonkatu 34, Lappeenranta, Finland. [2] Finnish Geospatial Research Institute, Geodeetinrinne 2, 02431 Masala, Finland. [3] Finnish Meteorological Institute, Erik Palménin aukio 1, 00560 Helsinki, Finland.

Received: 21 August 2020 Accepted: 3 April 2021
Published online: 15 April 2021

References

1. De Silva PM, Marshall JM. Factors contributing to urban malaria transmission in sub-Saharan Africa: a systematic review. J Trop Med. 2012;2012:819563.
2. Huldén L, McKitrick R, Huldén L. Average household size and the eradication of malaria. J R Stat Soc Ser A Stat Soc. 2014;177:725–42.
3. Hulden L. Household size explains successful malaria eradication. Malar J. 2010;9:18.
4. Huldén L, Huldén L. The decline of malaria in Finland-the impact of the vector and social variables. Malar J. 2009;8:94.
5. Ejov M, Davidyants V, Zvantsov A. Regional framework for the prevention of malaria reintroduction and certification of malaria elimination 2014–2020. Copenhagen: World Health Organization Regional Office for Europe; 2014.
6. Jelinek T, Schulte C, Behrens R, Grobusch MP, Coulaud JP, Bisoffi Z, et al. Imported Falciparum malaria in Europe: sentinel surveillance data from the European network on surveillance of imported infectious diseases. Clin Infect Dis. 2002;34:572–6.
7. Develoux M, Le GL, Dautheville S, Belkadi G, Magne D, Lassel L, et al. Malaria among immigrants, experience of a Parisian hospital (2006–2010). Bull Soc Pathol Exot. 2012;105:95–102.
8. Ericsson CD, Hatz C, Leder K, Tong S, Weld L, Kain KC, et al. Illness in travelers visiting friends and relatives: a review of the GeoSentinel Surveillance Network. Clin Infect Dis. 2006;43:1185–93.
9. Mascarello M, Gobbi F, Angheben A, Concia E, Marocco S, Anselmi M, et al. Imported malaria in immigrants to Italy: a changing pattern observed in north eastern Italy. J Travel Med. 2009;16:317–21.
10. Wallace DI, Southworth BS, Shi X, Chipman JW, Githeko AK. A comparison of five malaria transmission models: benchmark tests and implications for disease control. Malar J. 2014;13:268.
11. Smith NR, Trauer JM, Gambhir M, Richards JS, Maude RJ, Keith JM, et al. Agent-based models of malaria transmission: a systematic review. Malar J. 2018;17:299.
12. Churcher TS, Lissenden N, Griffin JT, Worrall E, Ranson H. The impact of pyrethroid resistance on the efficacy and effectiveness of bednets for malaria control in Africa. Elife. 2016;5:e16090.
13. Okumu FO, Kiware SS, Moore SJ, Killeen GF. Mathematical evaluation of community level impact of combining bed nets and indoor residual spraying upon malaria transmission in areas where the main vectors are *Anopheles arabiensis* mosquitoes. Parasit Vectors. 2013;6:17.
14. Briët OJ, Penny MA, Hardy D, Awolola TS, Van Bortel W, Corbel V, et al. Effects of pyrethroid resistance on the cost effectiveness of a mass distribution of long-lasting insecticidal nets: a modelling study. Malar J. 2013;12:77.
15. Kilama M, Smith DL, Hutchinson R, Kigozi R, Yeka A, Lavoy G, et al. Estimating the annual entomological inoculation rate for *Plasmodium falciparum* transmitted by *Anopheles gambiae* sl using three sampling methods in three sites in Uganda. Malar J. 2014;13:111.
16. Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, Hinsley W, et al. Reducing *Plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. PLoS Med. 2010;7:e1000324.
17. Sherrard-Smith E, Griffin JT, Winskill P, Corbel V, Pennetier C, Djénontin A, et al. Systematic review of indoor residual spray efficacy and effectiveness against *Plasmodium falciparum* in Africa. Nat Commun. 2018;9:4982.
18. Le Menach A, Takala S, McKenzie FE, Perisse A, Harris A, Flahault A, et al. An elaborated feeding cycle model for reductions in vectorial capacity of night-biting mosquitoes by insecticide-treated nets. Malar J. 2007;6:10.
19. Kitau J, Oxborough RM, Tungu PK, Matowo J, Malima RC, Magesa SM, et al. Species shifts in the *Anopheles gambiae* complex: do LLINs successfully control *Anopheles arabiensis*? PLoS One. 2012;7:e31481.
20. Shcherbacheva A, Killeen GF, Haario H. Modelling host-seeking behaviour of African malaria vector mosquitoes in the presence of long-lasting insecticidal nets. Math Biosci. 2018;295:36–47.
21. Vickers NJ. Mechanisms of animal navigation in odour plumes. Biol Bull. 2000;198:203–12.
22. Cardé RT. Odour plumes and odour-mediated flight in insects. Ciba Found Symp. 1996;200:54–66. **(discussion 66-70)**.
23. Cummins B, Cortez R, Foppa IM, Walbeck J, Hyman JM. A spatial model of mosquito host-seeking behaviour. PLoS Comput Biol. 2012;8:e1002500.
24. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21:1087–92.
25. Clements AN, Paterson GD. The analysis of mortality and survival rates in wild populations of mosquitoes. J Appl Ecol. 1981;18:373–99.
26. World Health Organization. Guidelines for efficacy testing of spatial repellents. Geneva: World Health Organization; 2013.
27. Yahouédo GA, Chandre F, Rossignol M, Ginibre C, Balabanidou V, Mendez NG, et al. Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector *Anopheles gambiae*. Sci Rep. 2017;7:11091.
28. Nardini L, Christian RN, Coetzer N, Ranson H, Coetzee M, Koekemoer LL. Detoxification enzymes associated with insecticide resistance in laboratory strains of *Anopheles arabiensis* of different geographic origin. Parasit Vectors. 2012;5:113.
29. Okumu FO. Combining insecticide treated bed nets and indoor residual spraying for malaria vector control in Africa [Doctoral dissertation]. London School of Hygiene and Tropical Medicine. 2012.
30. Kerkut GA, Gilbert LI. Comprehensive insect physiology, biochemistry and pharmacology. Oxford: Pergamon; 1985.
31. Semakula HM, Song G, Zhang S, Achuu SP. Potential of household environmental resources and practices in eliminating residual malaria transmission: a case study of Tanzania, Burundi, Malawi and Liberia. Afr Health Sci. 2015;15:819–27.
32. Maia MF, Onyango SP, Thele M, Simfukwe ET, Turner EL, Moore SJ. Do topical repellents divert mosquitoes within a community? Health equity implications of topical repellents as a mosquito bite prevention tool. PLoS ONE. 2013;8:e84875.
33. Maia MF, Kreppel K, Mbeyela E, Roman D, Mayagaya V, Lobo NF, et al. A crossover study to evaluate the diversion of malaria vectors in a community with incomplete coverage of spatial repellents in the Kilombero Valley, Tanzania. Parasit Vectors. 2016;9:451.

34. World Health Organization. Guidelines for testing mosquito adulticides for indoor residual spraying and treatment of mosquito nets. Geneva: World Health Organization; 2006.
35. Cator LJ, Lynch PA, Read AF, Thomas MB. Do malaria parasites manipulate mosquitoes? Trends Parasitol. 2012;28:466–70.
36. Koella JC, Rieu L, Paul RE. Stage-specific manipulation of a mosquito's host-seeking behaviour by the malaria parasite *Plasmodium gallinaceum*. Behav Ecol. 2002;13:816–20.
37. Rossignol PA, Ribeiro JM, Spielman A. Increased intradermal probing time in sporozoite-infected mosquitoes. Am J Trop Med Hyg. 1984;33:17–20.
38. Schwartz A, Koella JC. Trade-offs, conflicts of interest and manipulation in Plasmodium-mosquito interactions. Trends Parasitol. 2001;17:189–94.
39. Moore J. Parasites and the behaviour of animals. New York: Oxford University Press; 2002.
40. Cator LJ, Pietri JE, Murdock CC, Ohm JR, Lewis EE, Read AF, et al. Immune response and insulin signalling alter mosquito feeding behaviour to enhance malaria transmission potential. Sci Rep. 2015;5:11947.
41. Vantaux A, de Sales Hien DF, Yaméogo B, Dabiré KR, Thomas F, Cohuet A, et al. Host-seeking behaviours of mosquitoes experimentally infected with sympatric field isolates of the human malaria parasite *Plasmodium falciparum*: no evidence for host manipulation. Front Ecol Evol. 2015;3:86.
42. Lacroix R, Mukabana WR, Gouagna LC, Koella JC. Malaria infection increases attractiveness of humans to mosquitoes. PLoS Biol. 2005;3:e298.
43. Shcherbacheva A, Haario H. The impact of household size on malaria reduction in relation with alterations in mosquito behavior by malaria parasite. J Multi-Valued Log Soft Comput. 2017;29:455–468.
44. Bishop CM. Pattern recognition and machine learning. Berlin: Springer; 2006.
45. Montague PR. Reinforcement Learning: An Introduction, by Sutton, RS and Barto, AG. Trends Cogn Sci. 1999;3:360.
46. Bidlingmayer WL, Hem DG. The range of visual attraction and the effect of competitive visual attractants upon mosquito (Diptera: Culicidae) flight. Bull Entomol Res. 1980;70:321–42.
47. Hawkes FM, Dabiré RK, Sawadogo SP, Torr SJ, Gibson G. Exploiting Anopheles responses to thermal, odour and visual stimuli to improve surveillance and control of malaria. Sci Rep. 2017;7:17283.
48. van Breugel F, Riffell J, Fairhall A, Dickinson MH. Mosquitoes use vision to associate odour plumes with thermal targets. Curr Biol. 2015;25:2123–9.
49. Kelly-Hope LA, McKenzie FE. The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan Africa. Malar J. 2009;8:19.
50. Beier JC, Killeen GF, Githure JI. Entomologic inoculation rates and *Plasmodium falciparum* malaria prevalence in Africa. Am J Trop Med Hyg. 1999;61:109–13.
51. Filipe JA, Riley EM, Drakeley CJ, Sutherland CJ, Ghani AC. Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. PLoS Comput Biol. 2007;3:e255.
52. Sherrard-Smith E, Skarp JE, Beale AD, Fornadel C, Norris LC, Moore SJ, Mihreteab S, Charlwood JD, Bhatt S, Winskill P, Griffin JT. Mosquito feeding behavior and how it influences residual malaria transmission across Africa. Proc Natl Acad Sci. 2019;116:15086–95.

Publisher's Note

Publication II

Wijaya, K.P., Aldila, D., Erandi, K.K.W., Fakhruddin, M., Amadi, M., and Ganegoda, N.
Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia

Reprinted with permission from
Stochastic Environmental Research and Risk Assessment
Vol. 35, pp. 437–456, 2021.

Stochastic Environmental Research and Risk Assessment (2021) 35:437–456
https://doi.org/10.1007/s00477-020-01887-w

ORIGINAL PAPER

Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia

Karunia Putra Wijaya[1] · Dipo Aldila[2] · K. K. W. Hashita Erandi[3] · Muhammad Fakhruddin[4,5] · Miracle Amadi[6] · Naleen Ganegoda[7]

Accepted: 23 September 2020 / Published online: 8 October 2020

Abstract
Medical statistics collected by WHO indicates that dengue fever is still ravaging developing regions with climates befitting mosquito breeding amidst moderate-to-weak health systems. This work initiates a study over 2009–2017 panel data of dengue incidences and meteorological factors in Jakarta, Indonesia to bear particular understanding. Using a panel random-effect model joined by the pooled estimator, we show positively significant relationships between the incidence level and meteorological factors. We ideate a clustering strategy to decompose the meteorological datasets into several more datasets such that more explanatory variables are present and the zero-inflated problem from the incidence data can be handled properly. The resulting new model gives good agreement with the incidence data accompanied by a high coefficient of determination and normal zero-mean error in the prediction window. A risk measure is characterized from a one-step vector autoregression model relying solely on the incidence data and a threshold incidence level separating the low-risk and high-risk regime. Its magnitude greater than unity and the weak stochastic convergence to the endemic equilibrium mark a persistent cyclicality of the disease in all the five districts in Jakarta. Moreover, all districts are shown to co-vary profoundly positively in terms of epidemics occurrence, both generally and timely. We also show that the peak of incidences propagates almost periodically every year on the districts with the most to the least recurrence: Central, South, West, East, and North Jakarta.

Keywords Dengue · Clustering-integrated multiple panel regression · Risk measure · Spatial correlation · Outbreak propagation

✉ Dipo Aldila
aldiladipo@sci.ui.ac.id

[1] Mathematical Institute, University of Koblenz, 56070 Koblenz, Germany

[2] Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia

[3] Department of Mathematics, University of Colombo, Colombo 00700, Sri Lanka

[4] Department of Mathematics, Bandung Institute of Technology, Bandung 13240, Indonesia

[5] Department of Mathematics, Indonesia Defense University, Bogor 16810, Indonesia

[6] Department of Mathematics and Physics, Lappeenranta University of Technology, 53851 Lappeenranta, Finland

[7] Department of Mathematics, University of Sri Jayewardenepura, Nugegoda 10250, Sri Lanka

1 Introduction

Until today, dengue fever and its debilitating variants remain to be notable health burdens in Indonesia. Factors influencing the widespread of dengue diseases in big cities like Jakarta are manyfold. Environmental support for mosquito breading (Arcari et al. 2007; Morin et al. 2013), compromised healthy lifestyle (Haryanto 2016), the presence of all four dengue serotypes (Kusriastuti and Sutomo 2005), migration and populousness (Suwandono et al. 2006), close proximity to shores (Haryanto 2016; Wijaya et al. 2019), lack of indoor visibility in typical households (Spiegel et al. 2005), lack of community participation (Haryanto 2016), and pro and contra on dengue vaccines (Deng et al. 2020; Wilder-Smith 2020) are conspicuous factors, just to name. High population density also helps increase the incidence risk given the nature of *Aedes* mosquitoes as *multi–biters*, i.e., that a single vector can bite multiple humans (Jansen and Beebe 2010; Harrington et al. 2014). Difficulties in suppressing dengue incidences set a tremendous challenge for the government. Besides uncontrolled urbanization flows, lack of perseverance, motivation, and community participation in anti-dengue actions gives another problem. Among the enhancers of dengue widespread, meteorological factors have been a special subject of discussions due to their learnable patterns. In tropical countries, epidemics mainly occur in the rainy season and reach the highest peak given some time lags after meteorological factors favor mosquitoes to breed (Wang et al. 2019; Kakarla et al. 2019). Indonesia's Ministry of Health even combines climate and early warning system as marked fulcrums of intervention against dengue diseases (Haryanto 2016). Furthermore, finding a meticulous early warning system that integrates meteorological factors has been an active research area.

A number of studies have been conducted to study the effects of meteorological factors on the dengue incidence level in the context of multiple regression models. Lowe et al. (2011) use a generalized linear model (GLM) and the generalized linear mixed model (GLMM) counterpart pronouncing highly endemic areas to model large panel data of monthly incidence levels in Brazil, associated meteorological factors (rainfall, temperature, and El Niño), and a series of relevant factors reflecting zones and their interaction. Rainfall and temperature, each of lags 1 and 2 months, were shown to significantly increase, whereas El Niño 3.4 index decreases the incidence level. Several previous studies have also proposed different forecasting techniques and decision-support systems for Indonesia. Ramadona et al. (2015) assess the relation between seasonal incidence level and meteorological factors (rainfall, temperature, humidity) in Yogyakarta, Indonesia during 2001–2013 using a GLM in which humidity data were first adjusted according to rainfall and temperature variation, i.e., due to multicollinearity. For all meteorological factors, all the data of time lags 0 to 3 months were used as showing positive correlations with the incidence level. An autoregression model was done separately, however returning weaker predictability as compared to the model involving also lagged meteorological factors. Astuti et al. (2019) apply a GLM to assess the short-term effect of climate on the incidence level and the results showed that temperature of lag 4 months, rainfall of lag 2 months, and humidity of lag 0 month were associated with high incidence levels in children. Spatial analyses were conducted by using Moran's autocorrelation coefficient and local indicator of spatial association (LISA) to explore geographical clustering in the incidences and to identify high-risk villages. As a result, they identified the total of 38 high-risk villages in Cirebon, Indonesia for dengue. Chien and Yu (2014) combine a quasi-Poisson distributed lag nonlinear model (DLNM) with a spatial function adapting spatial autocorrelations and describing spatial heterogeneity to identify the relation between weekly minimum temperature, maximum 24-h rainfall, and dengue incidence level in southern Taiwan during 1998–2011 in the panel sense. With a characterized measure Relative Risk (RR), they were able to show that the increase of both meteorological factors rose RR while certain levels of maximum 24-h rainfall determine the lasting time of the associated RR. Another study for southern Taiwan during the highest epidemic year 2007 was done by Yu et al. (2011) via a log-link Poisson model incorporating panel incidence level, seasonal lag meteorological factors, El Niño Southern Oscillation Index (SOI), Breteau index, and panel population sizes. They found that most included meteorological factors are positively correlated with the incidence level while, surprisingly, most lag SOI (associated with warmer winter) and Breteau index (mosquito-positive containers) are negatively correlated with the incidence level. A set of multiple linear regression models integrating the least absolute shrinkage and selection operator (LASSO) method was proposed by Shi et al. (2016) to assess the relation between weekly incidence level, mean temperature, maximum hourly temperature, number of hours for high temperature ($> 27.8^{\circ}$C), humidity, and breeding percentage (BP) describing the ratio between *Ae. aegypti* and total breeding sites in Singapore during 2001–2012. Having shown good fitting and strong predictability via the mean absolute percentage error, the method entails caution for possible lack of interpretation due to complexity in selecting different lags among explanatory variables. Bouzid et al. (2014) attempt to project dengue incidences in 27 European Union members by assessing the incidences in Mexico during 1985–2007 and their relation with

minimal and maximal temperature, rainfall, humidity, GDP, percentage of the urban population, and population density. A generalized adaptive model (GAM), which is a GLM involving transformed explanatory variables, was used for the fitting. Under short-term, medium-term, and long-term projection, areas around Mediterranean and Adriatic coasts and in northern Italy were forecasted to exhibit higher risk of infection. Similar related seasonal models with the same attempt to relate dengue incidence level with meteorological factors in Sri Lanka, India, and Guangzhou, China, can be found in Withanage et al. (2018), Ramachandran et al. (2016), and Xu et al. (2017), respectively.

In comparison to the studies mentioned earlier, our base contributions lie in the following aspects. We develop a clustering-integrated multiple regression model for aggregated weekly panel data of incidence level and meteorological factors (rainfall, average temperature, humidity) for the instance of Jakarta, Indonesia. A particular motivation of the clustering is the zero-inflated problem in the incidence data and limited accessible supporting data in Indonesia, while at the same time, a model with proper fitting and strong predictability is highly required to help decision-makers build up an early warning system. As the incidence data are clustered into low, medium, and high level, the associated meteorological factors are classified following the places of the incidence levels. After sequential tests, we opt for a random-effect model optimized using the pooled estimator. As varying the clustering barriers returns different modeling results, we then attempt to find the optimal barriers in the sense that the mean squared error is minimized. The fact that constant coefficients in the model are used for different districts prompts a unified model where the coefficients represent the marginal effects of the meteorological factors to the incidence level in the general sense, i.e., district-independent. Unlike previous studies, we are only using the lags for the meteorological factors associated with the maximal Spearman's correlation coefficients for computational efficiency. A framework for estimating risk is proposed via a one-step vector autoregression model involving only incidence data. The resulting risk measure can be used to study both disease persistence and which districts correspond to the low and high risk. As far as spatial competition is concerned, we perform global and local spatial autocorrelation analysis using Moran's and Geary's autocorrelation coefficient, respectively. The results show how all the districts either simultaneously raise the incidence level or exhibit significant discrepancy without given clear evidence, which district should be of the priority at a specific time point. For this reason, we determine how the highest peak of incidences from all the districts propagates over time through spatial periodicity investigation.

2 Preliminaries

2.1 Study area and data

The city of Jakarta is divided into six districts, among which the central part serves as the administrative regency. The districts are North Jakarta (approximated area 142km^2, share 21.48%), East Jakarta (187 km^2, 28.3%), South Jakarta (146 km^2, 22.1%), West Jakarta (126 km^2, 19.1%), the Thousand Islands (12 km^2, 1.82%), and Central Jakarta (48 km^2, 7.2%). To the north stretches the coast along 35 km, where 13 rivers and two canals disembogue. Due to a tiny share and almost-zero incidence level, we skip the Thousand Islands from our investigation. The means of the population shares from 2009 to 2017 (see Table 1) impart that North, East, South, West, and Central Jakarta share 16.99%, 27.89%, 21.63%, 23.89%, and 9.36% of the total population in the city, respectively. From the preceding calculation, we can extract further information regarding the population-to-area ratio, where the quotient returns 0.79, 0.99, 0.98, 1.25, and 1.3 for the respective districts. The numbers clearly indicate that North, East, and South Jakarta are *relatively* underpopulated, whereas West and Central Jakarta are *relatively* overpopulated. Jakarta has the highest population density in Indonesia with the total population of 10,374,235 inhabitants (661 km^2 area) as of

Table 1 The total population of the five districts of Jakarta. Source: *Jakarta Dalam Angka* (Jakarta in Numbers) X+1, where X denotes the corresponding year the data were recorded

	2009	2010	2011	2012	2013	2014	2015	2016	2017
North	1471663	1645659	1716345	1715564	1711036	1729444	1747315	1764614	1781316
East	2448653	2693896	2926732	2801784	2791072	2817994	2843816	2868910	2892783
South	2159638	2062232	2135571	2148261	2141941	2164070	2185711	2206732	2226830
West	2221243	2281945	2260341	2395130	2396585	2430410	2463560	2496002	2528065
Central	902216	902973	1123670	908829	906601	910381	914182	917754	921344

2017, which is far above the second-largest city, Surabaya, with the corresponding population of 3,057,766 inhabitants (327 km^2 area). The weather condition of the city generally is warm and humid with a maximum temperature of 35.4°C during the day, and a minimum temperature of 21.5°C at night. The average rainfall throughout the year during the period 2009–2017 is 6.48 mm, and the air humidity falls within 47.5–99.5% on a daily basis.

2.2 Incidence data and scaling

All datasets undertaken in this study cover the time window January 6, 2009—September 25, 2017, defined on a daily basis. The data of dengue hospitalized cases were collected from Jakarta Health Office, while the meteorological data were supplied by the Indonesian Agency for Meteorology, Climatology, and Geophysics. The dengue data are compiled records of dengue fever inpatients from 154 hospitals in Jakarta, irrespective of variants of illness. Note that 154 represents the total number of public and private hospitals that accommodate dengue patients and participate in dengue surveillance. Unlike meteorological data that are taken from unique regional weather stations and are ranging in comparable orders of magnitude, the incidence data suffer from bias due to the unmentioned population sizes. A region might appear to present large outliers in epidemics while serving as the most populated region. Another region might also show uniformity in the incidence levels with the others, however serving as the least populated region. To avoid further correction due to such bias, we utilize the density data. The required population data are taken from the annual book *Jakarta Dalam Angka* (Jakarta in Numbers) X+1 released by Statistics Indonesia, where X and X+1 denote the corresponding year in which the data were collected and the publication year, respectively. The data are presented in Table 1.

2.3 Reasons for data selection

Dengue virus is estimated to expand in the propagation that is marked by high rainfall, temperature, and humidity. An increment in the air temperature alters water temperature at the breeding place of mosquitoes that, in turn, hastens the egg hatching (Christophers 1960; Byttebier et al. 2014). Additionally, such increments also shorten the extrinsic incubation period of the virus in mosquitoes (Ogden and Lindsay 2016). A previous study on cross-correlation between monthly dengue incidence, rainfall, and temperature in eight provinces in Indonesia reveals that rainfall steers the geographical distribution while temperature intensifies the epidemics (Arcari et al. 2007). Other empirical studies confirm, however, that only under a proper combination of temperature and humidity can *Aedes aegypti* perform two-fold survival and produce evermore eggs (Juliano et al. 2002; de Almeida Costa et al. 2010). In fact, a certain range of humidity values leads to the accelerated larval growth and may decrease the proportion of the population surviving extrinsic incubation period (Christophers 1960). Therefore, the role of humidity in the context of dengue widespread is inasmuch as that of temperature. Rainfall is a weather element whose every datum is measured using a rain gauge so that the amount can be known in millimeters. Stagnant water in abandoned bottles and cans can become natural mosquito breeders since *Aedes* mosquitoes favor to breed in standing water that is not in direct contact with the ground (Jansen and Beebe 2010).

2.4 Data transformation

A few aspects shall be mentioned here regarding data preprocessing. The first problem appears as the incidence data contain lattice values, each of which has a relatively high recurrence. The resulting histogram merely shows a geometric distribution, which is of discrete type. We mitigate such incompatibility with the meteorological data by accumulating all the daily datasets into weekly datasets. The yielded cardinality is $454\ \text{weeks} + 6\ \text{days} \approx 455\ \text{weeks}$. The resulting datasets yet show another incompatibility in the context of linear regression. We will see that the incidence and rainfall dataset exhibit exponential distributions, whereas temperature and humidity dataset exhibit normal distributions. The eligibility of linear regression requires all datasets to be identically and independently distributed (*i.i.d.*). For wider compatibility with dummy or coded variables, we normalize the incidence and rainfall dataset. The transformation idea follows from the so-called *Gaussianization* (Chen and Gopinath 2000). For a random seasonal variable x of an exponential distribution, the first transformation into the cumulative distribution function (CDF) of the uniform distribution can be done via $\Theta(x) = 1 - \exp(-x/\bar{x})$, where $\bar{x}$ denotes the mean of x. Then, the second transformation $\Phi^{-1}(\Theta)$, where Φ denotes the CDF of the standard normal distribution, gives an approximation to the normal distribution. The preceding formula contains a flaw as the inverse of the error function returns $-\infty$ at 0. Consequently, overwhelming zero values obstruct the normalization process, which is the case in our incidence and rainfall dataset. Here, we ideate an approximation to the inverse of the error function, $\Psi : [0, 1] \to \mathbb{R}$. The function Ψ should behave similarly as Φ^{-1} and is flexible in terms of boundary conditions $\Psi(0)$, $\Psi(1)$, a mid-point M, and a stiffness

measure κ. We come across, but not limited to, the following formula

$$\Psi = \kappa \cdot \tan(\eta\Theta + \nu) + M \tag{1}$$

where

$$\nu = \tan^{-1}\left(\frac{\Psi(0) - M}{\kappa}\right), \quad \eta = \tan^{-1}\left(\frac{\Psi(1) - M}{\kappa}\right) - \nu.$$

Since both Θ and Ψ are monotonically increasing, so is the composite form $\Psi \circ \Theta$. For both incidence and rainfall dataset, we use $\Psi(0) = 0$, $\Psi(1) = 1$, $M = 0.5$, and $\kappa = 0.1$. The corresponding transformation results can be seen in Fig. 1.

2.5 Calculation of true marginal effects

A critical viewpoint to be considered after data transformation is how the result of linear regression can be interpreted. For a simple seasonal model $y = \beta x + \varepsilon$, the marginal effect of x to y is defined as $\mathrm{d}y/\mathrm{d}x = \beta \approx \Delta y/\Delta x$. In this case, for any current value of the explanatory variable x, the increase of it by $(\Delta x/x) \times 100\%$ produces the increase of the corresponding response variable at y by $(\Delta y/y) \times 100\% \approx (\beta \Delta x/y) \times 100\%$.

In the sequel, the data are to be approached by (panel) linear regression models. When the pooled estimator is used, the computation of the coefficients is based on rewriting the panel model into a seasonal model by piling up district datasets into single-district datasets. In a first representation of the seasonal model, the interrelationship among datasets can be written as $d = \beta_0 + \beta_r r + \beta_t t + \beta_h h + \varepsilon$, where $d = (d_i), r = (r_i), t = (t_i), h = (h_i)$ with $i = 1, \cdots, IJ$ denote the concatenated dengue incidence, rainfall, temperature, and humidity dataset, respectively. Here, I and J denote the number of time points and that of districts, respectively. Under Gaussianization of the incidence and rainfall dataset, the model changes to

$$\Psi \circ \Theta(d) = \beta_0 + \beta_r \Psi \circ \Theta(r) + \beta_t t + \beta_h h + \varepsilon.$$

The marginal effects thereupon do not represent the true marginal effects. In case of the marginal effect of rainfall to dengue incidence, we obtain the true marginal effect β_r^{true} by the following identity

$$\beta_r = \frac{\mathrm{d}\{\Psi \circ \Theta(d_i)\}}{\mathrm{d}\{\Psi \circ \Theta(r_i)\}} = \frac{\Psi'\Theta'|_{d_i}\beta_r^{\text{true}}}{\Psi'\Theta'|_{r_i}}, \tag{2}$$

where the apostrophes in Ψ' and Θ' denote the first derivative with respect to the argument, and d_i, r_i any taken values from the datasets. A caveat here is, that β_r^{true} as in (2) returns different values when using d_i and r_i for different i's.

Two notions mitigate the problems of having variable true marginal effects (Cameron and Trivedi 2010). The first notion is the so-called *marginal effect at means*. The basic idea is that the true marginal effect should be evaluated at the means of the dengue incidence, $\bar{d}$, and rainfall dataset, $\bar{r}$. As in (2), the marginal effect at means is given by

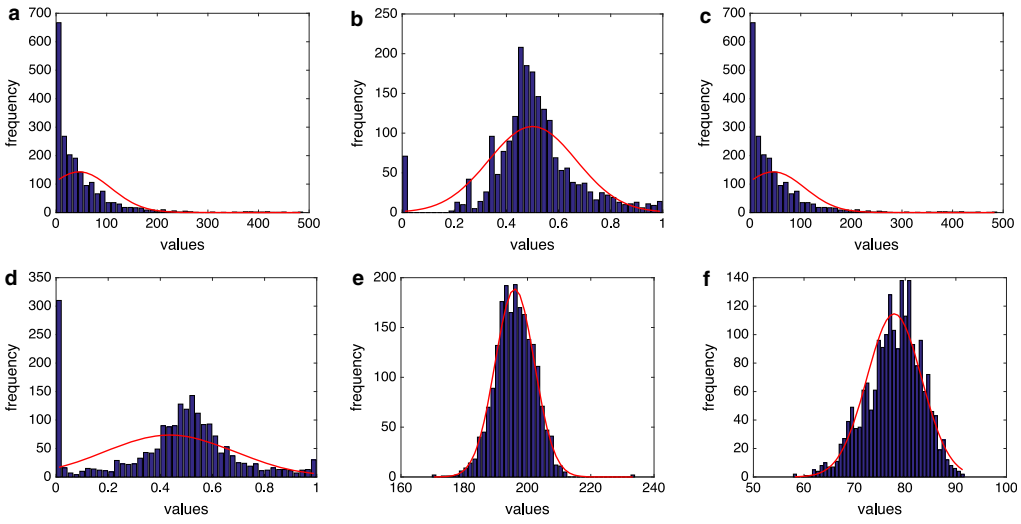


Fig. 1 Distribution of the weekly incidence, rainfall, temperature, and humidity dataset in **a**, **c**, **e**, **f**, respectively. Those of the transformed incidence and rainfall datasets are shown in **b** and **d**, respectively

$\beta_r^{\text{true}} = \beta_r \Psi'\Theta'|_{\bar{r}} / \Psi'\Theta'|_{\bar{d}}$. The second notion is the *average marginal effect*. As the name tells, it calculates $\Psi'\Theta'$–terms using d_i and r_i for all i and uses the mean of all calculations $\overline{\Psi'\Theta'|_d} := (1/IJ)\sum_{i=1}^{IJ} \Psi'\Theta'|_{d_i}$ and $\overline{\Psi'\Theta'|_r} := (1/IJ)\sum_{i=1}^{IJ} \Psi'\Theta'|_{r_i}$ to replace the numerator and denominator in (2), respectively. Due to the absence of such individuals $\bar{d}$ and $\bar{r}$, the average marginal effect provides better results than the marginal effect at means in terms of interpretability. Therefore, this study uses the average marginal effects $\beta_r^{\text{true}} = \beta_r \overline{\Psi'\Theta'|_r} \Big/ \overline{\Psi'\Theta'|_d}$, $\beta_t^{\text{true}} = \beta_t \Big/ \overline{\Psi'\Theta'|_d}$, and $\beta_h^{\text{true}} = \beta_h \Big/ \overline{\Psi'\Theta'|_d}$.

Unless stated otherwise, in what follows we will only use the transformed data in order not to envisage possible confusion as well as keep I, J to denote the number of time points and that of districts, respectively.

2.6 Clustering

The aims of clustering in our study are not only to increase the number of explanatory variables but also to drop meteorological data that correspond to the zero incidence levels. The clustering idea is inspired by our previous work on seasonal models in Fakhruddin et al. (2019). Let $D = (d_{ij}), R = (r_{ij}), T = (t_{ij}), H = (h_{ij}) \in \mathbb{R}^{I\times J}$ (where $i = 1, \cdots, I$, $j = 1, \cdots, J$) represent the panel incidence, rainfall, temperature, and humidity dataset, respectively. The data clustering in this study centers on the dengue incidence dataset. We restrict our investigation to the use of three clusters with a lower b_l and an upper barrier b_u. An incidence case level d_{ij} is classified as low if $d_{ij} < b_l$, medium if $b_l \le d_{ij} < b_u$, or high if $d_{ij} \ge b_u$. For brevity, we collect the barriers into a vector $b := (b_l, b_u)$. We can set an auxiliary function $\delta_l(d_{ij}, b)$, which is 1 in case d_{ij} belongs to the low cluster, i.e., $d_{ij} < b_l$, or otherwise 0. Constructions of $\delta_m(d_{ij}, b)$ and $\delta_u(d_{ij}, b)$ can be done similarly. Now the matrices

$$E^l := \begin{pmatrix} \delta_l(d_{11}, b) & \cdots & \delta_l(d_{1J}, b) \\ \vdots & \ddots & \vdots \\ \delta_l(d_{I1}, b) & \cdots & \delta_l(d_{IJ}, b) \end{pmatrix} \tag{3}$$

and E^m, E^u constructed as for E^l by replacing δ_l with δ_m, δ_u, respectively, denote the *clustering-specific effects*. These are binary matrices serving as dummy variables. Observe that $E^l + E^m + E^u = E$, the one–matrix. Let $P \odot Q = (p_{ij}q_{ij})$ denote the Hadamard product between two matrices P, Q of equal size, which is the entrywise product. Then, the matrices

$$R^l := R \odot E^l, \quad R^m := R \odot E^m, \quad R^u := R \odot E^u \tag{4}$$

denote the clustered rainfall datasets manifested from the clustering of the incidence dataset, where it holds $R^l + R^m + R^u = R$. The clustering in the temperature and humidity dataset can likewise be done as in the rainfall dataset.

2.7 Cross-correlation

We analyze the time lags between dengue incidence and meteorological factors before they are put into the modeling. The utilized tool is Spearman's correlation coefficient. We use Spearman's in favor of Pearson's correlation coefficient for its robustness against monotonic transformations over the datasets. Based on 455 data points from the first week of January 2009 until the third week of September 2017, we fixed the last 403 dengue incidence data and moved the 403-week windows of meteorological parameters back in time, for all districts. This way, we tested the correlations for the full one year's shift. During every window's fixation, we calculate the correlation coefficients and later find the maximum. The time lags corresponding to the maximal correlation coefficients can be seen in Table 2.

We found that all the incidence–rainfall, incidence–temperature and incidence–humidity correlations change in sign, i.e., indicating wavering sine-like curves. Especially for the incidence–temperature and incidence–humidity correlations, the maximum absolute value of the negative part is relatively slightly smaller than that of the positive part. The incidence–temperature correlations mostly start with negative coefficients, overtaken by positive coefficients as the lag increases. Meanwhile, the incidence–rainfall and incidence–humidity correlations behave the other way around. Such indicators of instability may be related to the fact that only proper combinations of temperature and humidity can sustain *Aedes* mosquitoes' lives, whereas the rest can kill evermore (Juliano et al. 2002; de

Table 2 Time lags and maximal correlation coefficients found after Spearman's cross-correlation between dengue incidence and meteorological factors in the five districts in Jakarta

	Rainfall	Temperature	Humidity
North	11w, 0.2616	44w, 0.2288	9w, 0.3446
East	11w, 0.4308	22w, 0.2684	9w, 0.3149
South	7w, 0.3867	34w, 0.4089	6w, 0.4524
West	6w, 0.4037	21w, 0.2520	10w, 0.5086
Central	8w, 0.4565	29w, 0.2780	7w, 0.5083

The unit w stands for week

Almeida Costa et al. 2010). Now, let $\tau_j^r, \tau_j^t, \tau_j^h$ denote the time lags of rainfall, temperature, and humidity with respect to dengue incidence on the jth district, whose values can be taken from Table 2. As the maximum of all the time lags is 44w, we use the last 412 data points for dengue incidence, meanwhile the meteorological parameters are shifted back according to the time lags. Eventually, we change to $I = 412$ while $J = 5$ remains.

3 Random-effect model

Let us recall the one–matrix E and define new matrices based on the cross-correlations

$$R := (r_{i-\tau_j^r, j}), \quad T := (t_{i-\tau_j^t, j}), \quad H := (h_{i-\tau_j^h, j}).$$

Alluded from the panel regression is the use of individual-specific effects $\beta_1, \cdots, \beta_{J-1}$, which only differ by districts. Leaving out an individual-specific effect for one district is intentionally avoiding linear dependence with the pre-determined constant β_0. In order to fit into a matrix equation, we use the following operation for the individual-specific effects

$$\beta^d \otimes \mathbb{1} := (\beta_1 \ \cdots \ \beta_{J-1}) \otimes \left.\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right\} I.$$

The notation $P \otimes Q = (Pq_{ij})$ denotes the usual Kronecker product between two matrices P and Q. A simple form of linear regression model only reveals the direct interrelationship among the datasets

$$D = \beta_0 E + \beta^d \otimes \mathbb{1} + \beta_R R + \beta_T T + \beta_H H + \varepsilon. \tag{5}$$

The variable ε in the model (5) denotes the *idiosyncratic error*.

The eligibility of the linear model (5) relies on several assumptions, according to which certain diagnostics have to be done. First, the model must be linear with respect to all the coefficients β's, which is evident, and $\mathbb{E}(\varepsilon_{ij}) = 0$ and $\mathbb{E}(\beta_j) = 0$ for all i, j. The latest assumption is what we instantly impose in this study without any test, even though the standard t-test advocates all the individual-specific effects being significantly different from zero. Second, all the involved variables are assumed to be *i.i.d.* which we endeavor to see in Sect. 2.4. Additionally, we assume that all the individual-specific effects also serve as normal random variables without requiring any test. Third, the idiosyncratic error cannot be correlated with the explanatory variables as well as individual-specific effects, i.e., exogeneity. In connection with the first assumption, this means that $\mathbb{E}(\varepsilon_{ij}) = 0$ independently from the other predictor variables. Fourth, the individual errors $\varepsilon_j := (\varepsilon_{1j}, \cdots, \varepsilon_{Ij})^\top$ are uncorrelated, i.e., no spatial correlation in the individual errors, and there is no serial correlation on ε_{ij}, i.e., $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{kj}) = 0$ independently from all the predictor variables when $i \neq k$. Violation to the latter especially leads to smaller standard errors of the coefficients and a higher coefficient of determination. We will use the Wooldridge test for serial correlation. A p-Value smaller than α rejects the null hypothesis that no serial correlation is found from the error.

The appearance of individual-specific effects β^d automatically designates the model to be either a fixed-effect or a random-effect model. Durbin–Wu–Hausman test (Davidson and MacKinnon 1993) is performed to see if a random-effect model is chosen in favor of a fixed-effect model, i.e., the null hypothesis. The test is in line with the aforementioned third assumption on the eligibility of the linear model that the errors are uncorrelated with all the predictor variables. Additional to the eligibility of a random-effect model is the absence of multicollinearity. The higher the multicollinearity, the higher the standard deviations of the coefficients, and therefore the wider the confidence intervals and the smaller the t-statistics. An overt sign of multicollinearity is that the suspected explanatory variables have the same correlation to the response variable and standard deviation, yet their significance levels are radically different (Farrar and Glauber 1967; Willis and Perlack 1978). To see multicollinearity, we calculate the *Inverse of Variance Inflation Factor* (1/VIF). Any value of it measures one minus the coefficient of determination resulting from a linear regression among all explanatory variables where the one under measurement serves as the response variable. Consequently, a 1/VIF smaller than, say 0.1, indicates a multicollinearity problem (Mansfield and Helms 1982). After all, multicollinearity only scopes out redundancy or possible overfitting and does not influence the validity of the significance tests.

A purposive measure to see the validity of the significance tests is error normality. One can say that the standard deviations of the coefficients cannot be reliable in case the error is not normal. This means that the confidence intervals should have been too wide or too narrow. Nonetheless, even if the distribution is grossly non-normal, the significance tests will still provide good approximations for the standard deviations. We used Shapiro–Wilk W–test for normality (null hypothesis) of any random variable (Chen 1971; Hanusz et al. 2016). Note that error normality is independent of whether the computed optimal coefficients are biased or unbiased. Finally, we check for heteroscedasticity, a specific feature showing that the variance of the error is dependent of the explanatory variables, otherwise homoscedasticity. One obvious sign of

Table 3 Results from model (5). AME, StD, Adj R^2, W, B–P, D–W–H, Wo stand for average marginal effect, standard deviation, Adjusted R^2, p-Value of Shapiro–Wilk W–test for normality, p-Value of Breusch–Pagan test for heteroscedasticity, p-Value of Durbin–Wu–Hausman test for a random-effect vs. a fixed-effect model, and Wooldridge test for the serial correlation, respectively

	Val	AME	StD	p-Value	F p-Value	1/VIF	R^2	Adj R^2	W	B–P	D–W–H	Wo
β_0	− 1.3564		.1181	0	0		.3002	.2978	0	0	1	0.0510
β_1	− .1211		.0097	0		.6218						
β_2	− .0813		.0106	0		.5254						
β_3	− .0327		.0105	.002		.5349						
β_4	− .0025		.0106	.813		.5292						
β_R	.1092	7.2914×10^{-9}	.0143	0		.7491						
β_T	.0052	2.8191×10^{-7}	.0005	0		.7451						
β_H	.0094	4.5642×10^{-7}	.0007	0		.6613						

heteroscedasticity is, that when a certain explanatory variable tends to step up larger values, the error simultaneously exhibits larger jumps. When the pooled estimator is preferred over the random-effect estimator, we can use Breusch–Pagan test for heteroscedasticity (Gujarati and Porter 2008). The p-Value is gained using a chi-squared statistic, of which smaller than α rejects the null hypothesis that the error is homoscedastic.

From the datasets, we are given the number of time points $I = 412$ and of districts or individuals $J = 5$. The panel data with such a characteristic are categorized as of a small sample size. This study compares the usual pooled estimator and random-effect estimator. Despite the evidence that the random-effect estimator is only consistent and asymptotically normally distributed under a large sample size (Schmidheiny 2019), we perform such comparison using Breusch–Pagan Lagrange Multiplier test, where the null hypothesis states that the variance across districts is negligible (no panel effect). The pooled estimator, on the other hand, is unbiased in small sample sizes given clearance on the eligibility of the linear model and random-effect model. The only caveat against pooled estimator for panel regression is its inefficiency, i.e., the variances of the coefficients are not minimal and incorrect due to error normality (Schmidheiny 2019). Keeping us warned on this fact, the pooled estimator is retained against the random-effect estimator for the sake of keeping further computations the simplest. At last, the usual variable significance t-test is done to determine whether every single coefficient involved in a regression model is significantly different from zero. Along with the single tests, the F-test's result is also presented to see if the overall coefficients of the predictor variables are significantly different from zero, i.e., that the current model provides a better fit to the data than a model that contains no explanatory variables, except the constant. In some cases, p-Values returning from t-tests can be large, whereas the accompanying p-Value from F-test is sufficiently small. This is another indicator of multicollinearity (Johnston 1972).

We fixed $\alpha = 0.01$ overall the statistical tests in our study. According to Table 3, Durbin–Wu–Hausman test strongly recommends the random-effect in favor of the fixed-effect model. Moreover, Breusch–Pagan Lagrange Multiplier test gives p-Value 1, suggesting no difference in the pooled and random-effect estimator. All the explanatory variables are significant, except for the fourth individual-specific effect. The F-test, however, indicates that all the variables are significant, which also exhibit no multicollinearity according to 1/VIF values. The model produces a non-normal, heteroscedastic error evincing no serial correlation. The average marginal effects reveal that the increases in rainfall, temperature, humidity by 10% at the means of the datasets $(\bar{D}, \bar{R}, \bar{T}, \bar{H}) = (7.4689 \times 10^{-6}, 45.7152, 195.97, 77.8258)$ results in that in the incidence by 0.45%, 73.97%, 47.56%, respectively.

4 Clustering-integrated model

4.1 General expression

So far, we have collected new explanatory variables based on the clustering, namely the clustering-specific effects E^l, E^m, E^u and clustered meteorological factors $R^l, R^m, R^u, T^l, T^m, T^u, H^l, H^m, H^u$. One trait of infectious diseases is the impact of past incidences on the number of current incidences. In this regard, autoregressive terms are also added to represent the sequential connection between past and current incidences. Our analysis reveals that the autocorrelation coefficients conceal relatively decreasing amplitudes. Focusing on positivity, we thus utilize for the sake of less computation effort, but not limited to,

incidence cases in the past four weeks. In the matrix form, these cases can be collected into $D_{-1} := (d_{i-1,j}), \cdots, D_{-4} := (d_{i-4,j})$. Let $k \in \{l, m, u\}$ and $s \in \{1, \cdots, 4\}$. Then, the general form of the model integrating clustering is presented as follows

$$D = \beta_0 E + \sum_k \beta_E^k E^k + \beta^d \otimes \mathbb{1} + \sum_k \beta_R^k R^k + \sum_k \beta_T^k T^k + \sum_k \beta_H^k H^k + \sum_s \beta_{-s} D_{-s} + \varepsilon. \tag{6}$$

Note that some of the explanatory variables in (6) are still subject to dropping due to possible multicollinearity (overfitting). We also note that all the clustering-specific effects and clustered meteorological factors—therefore the mean squared error—are subject to changes as we perturb the clustering barriers b_l, b_u. In the sequel, we will look for the optimal barriers b_l, b_u that give the minimum of the mean squared error.

4.2 Optimal clustering barriers

The model (6) can be arranged into a seasonal model taking the form

$$Y = X(b)\theta + \varepsilon, \quad b = (b_l, b_u),$$

with the pooled estimator $\theta = [X(b)^\top X(b)]^{-1} X(b)^\top Y$. Let $y_{\min} := \min Y$ and $y_{\max} := \max Y$. The dependency $X = X(b)$ returns from the fact that the clustered meteorological factors are dependent on the clustering barriers, see (3) and (4). Finding such optimal barriers requires to solve the following optimization problem

$$\begin{aligned} \min_b \quad & f(b) := \frac{1}{2} \left\| X(b)[X(b)^\top X(b)]^{-1} X(b)^\top Y - Y \right\|^2 \\ \text{s.t.} \quad & y_{\min} \le b_l \le b_u \le y_{\max}. \end{aligned} \tag{7}$$

The fact that Heaviside functions $\delta_l(d_{ij}, b) = 1 - \mathcal{H}(d_{ij} - b_l)$, $\delta_m(d_{ij}, b) = \mathcal{H}(d_{ij} - b_l) \cdot (1 - \mathcal{H}(d_{ij} - b_u))$, and $\delta_u(d_{ij}, b) = \mathcal{H}(d_{ij} - b_u)$ from (3) take part in the objective function f in (7) gives us a non-smooth constrained optimization problem.

The first treatment on the problem (7) is to transform the constrained into an unconstrained type. The basic idea is to augment the constraint functions into the original objective function using a penalty function. The new objective function becomes sufficiently large as the argument verges on the boundary of the feasible domain. We use a penalty function that not only gives such a feature but also saturates the objective function to a certain large value in case the argument is actually outside the feasible domain. The message here is that staying outside the feasible domain is allowed, but one can never find an optimal solution there. This strategy would also hinder the objective function from obtaining arbitrary large values around the boundary, possibly infinity, like in the usual penalty log function. Let $g \ge 0$ be a constraint function. The penalty function with the preceding specifications is modeled as

$$\mathcal{P}(g; \gamma, \upsilon) := \frac{1}{\upsilon} \left(\frac{\pi}{2} + \tan^{-1} \left(\frac{\gamma - g}{\upsilon} \right) \right).$$

The penalty function $\mathcal{P}$ resembles a Z-shaped curve bounded below by zero, where a downward jump happens around the mid-point γ. The parameter υ, on the one hand, stretches the curve to obtain large values when $g < \gamma$. This means that, by passing a small value $\gamma \simeq 0$, $\mathcal{P}$ becomes dominantly large as $g \simeq 0$ and saturates to a large constant as $g < 0$. On the other hand, υ also controls the stiffness of the jump. When υ is sufficiently small, the curve of $\mathcal{P}$ stretches and exhibits tremendous stiffness where the slope at γ tends to $-\infty$. Including the penalty function, the objective function transforms into

$$f(b) \mapsto f(b) + \sum_{i=1}^{3} \mathcal{P}(g_i(b); \gamma, \upsilon) \tag{8}$$

where $g_1 = b_l - y_{\min}$, $g_2 = b_u - b_l$, $g_3 = y_{\max} - b_u$.

This study undertook tests for the aforementioned parameters in which the optimization results are compared with the approximate optima using brute-force computations. We obtain a similar result as υ and γ are varied within $[10^{-8}, 10^{-1}]$ and $[10^{-3}, 10]$, respectively.

4.3 PSO-solver

Let us rewrite $f(b) = \frac{1}{2} \langle V(b), V(b) \rangle^2$ from (7). The derivative is given by $f'(b) = \langle V(b), V'(b) \rangle$ where

$$\begin{aligned} V &= X[X^\top X]^{-1} X^\top Y - Y, \\ V' &= \Big\{ X'[X^\top X]^{-1} X^\top + X[X^\top X]^{-1} X'^\top \\ &\quad - X[X^\top X]^{-1} [X'^\top X + X^\top X'][X^\top X]^{-1} X^\top \Big\} Y. \end{aligned}$$

Computing V' requires to relax the Heaviside functions in $X(b)$, e.g. $\mathcal{H}(d_{ij} - b_l)$, using their smooth approximation

$$\frac{1}{1 + e^{-2\zeta(d_{ij} - b_l)}} \xrightarrow{\zeta \to \infty} \mathcal{H}(d_{ij} - b_l).$$

With this approximation, the problem (7) becomes a smooth problem, to be handled by any gradient-based method. The advantage of using gradient-based methods is that only one "player" is used on every iteration. A disadvantage appears as the relaxation parameter ζ is set to be extremely large, in which case the derivative term V' discloses infinities. Also, imminent from such methods is the calculation of V' itself, which requires a number of matrix multiplications and inverse. Metaheuristics, such as

Particle Swarm Optimizer (PSO), do not require any information regarding the gradient of the objective function. PSO uses several "players" that create clusters shrinking to an optimal solution at times. The stochastic nature pronounced by random numbers in the algorithms helps increase the probability of finding the global optima, at least from the empirical point of view (Trelea 2003). Chances will be higher with a higher number of players. Even though there has been immense development of PSO algorithms recently (Rana et al. 2011; Wang et al. 2018), this study uses the following variant for its simplicity

$$\begin{aligned} b_{t,i} &= b_{t-1,i} + c_0 \cdot v_{t,i}, \\ v_{t+1,i} &= \omega \cdot \text{rand} \cdot v_{t,i} + c_1 \cdot \text{rand} \cdot (p_{t,i} - b_{t,i}) \\ &\quad + c_2 \cdot \text{rand} \cdot (q_t - b_{t,i}). \end{aligned}$$

Here rand denotes a random number in [0, 1].

The algorithm features the update of players' position b (in this case, the barriers) and velocity v. There, t, i denote the pointers for the iteration and player, respectively. Let us in addition introduce a measure for the individual best optimum $y_{t,i}$, which together with $p_{t,i}$, are given initial conditions. Let f_g denote the minimum of the objective function evaluations over all players and the entire history. As $f(b_{t,i}) < f_g$, we update $p_{t,i} = b_{t,i}$ and $y_{t,i} = f(b_{t,i})$. In the algorithm, the entity $q_t := p_{t,i^*}$ where $i^* = \arg\min_i y_{t,i}$ denotes the best player among all players on the current iteration. The *self confidence* c_1 determines how confident every player in performing local search as if the first equation signifies $p_{t,i}$ when c_1 is dominant among other parameters. The *player's confidence* c_2 measures every player's confidence against the global best position found on every iteration, making all players attracted to a single best position when $c_1 \ll c_2$. The case $c_1 \gg c_2$ means that players are trapped in their local best positions. The *inertia* ω emphasizes the influence of the historical velocities in determining the current velocities, while the *constriction factor* c_0 determines how far players can jump around between iterations. This study engineers 100 players, which indicate a relatively large swarm. For smooth objective functions, empirical studies suggest to use between 5 to 30 players (van den Bergh and Engelbrecht 2001; Brits et al. 2002). Previous studies (van den Bergh 2002; Brits et al. 2002; van den Bergh and Engelbrecht 2006) recommend to have a tradeoff between disproportionate wandering ($c_1 \gg c_2$) and premature finding of optima ($c_1 \ll c_2$), therefore we attested $c_1 = c_2 = 1$. Additionally, we use $\omega = 1$ and $c_0 = 0.3$.

In connection with the previous section, we show in Fig. 2 under excessive brute-force computations how the relaxation of the objective function $f(b)$ progresses towards its true evaluation as the relaxation parameter ζ increases. Along with this result, the PSO iterations are shown to not only approach the approximated global minimum from the brute-force computation but also locates the closest point to the true one.

4.4 Diagnostic results

We test the model (6) giving the trade-off between a large coefficient of determination, multicollinearity-free, and significance. We thus drop a variable when it is insignificant in order to preserve the coefficient of determination, then check the multicollinearity. Under this process, the following variables are dropped in the order of reading: H^l, β_3, β_4, E^m, E^l, E^u, R^u, T^u, T^l, H^m. The final model is given as

$$\begin{aligned} D = {} & \beta_0 E + (\beta_1, \beta_2) \otimes \mathbb{1} + \beta_R^l R^l \\ & + \beta_R^m R^m + \beta_T^m T^m + \beta_H^u H^u + \sum_s \beta_{-s} D_{-s} + \varepsilon, \end{aligned} \tag{9}$$

with the corresponding diagnostics presented in Table 4. This last format suggests that only rainfall corresponding to the low and medium, temperature to the medium, humidity to the high dengue incidence levels as well as the lag autoregressive terms determine the entire incidence level.

The least average marginal effect for β_R^l also suggests to drop the corresponding variable, from which case the clustering strategy successfully clears up the zero-inflated problem from the incidence data. Durbin–Wu–Hausman test strongly suggests using the random-effect model, whereas, Breusch–Pagan Lagrange Multiplier test gives p-Value 1, suggesting to opt for the pooled estimator. All the predictor variables are found to be significant, with a note that the second individual-specific effect is only significant at $\alpha = 0.1$ level. The F-test indicates the significance of the overall predictor variables. No multicollinearity is found. The produced coefficient of determination is relatively high, with the warning of overestimation due to the presence of serial correlation in the error. Moreover, the error is also found to be heteroscedastic with improved normality as compared to the simple model (5). Figure 3 depicts the computation results for the barriers and for the optimized model (9).

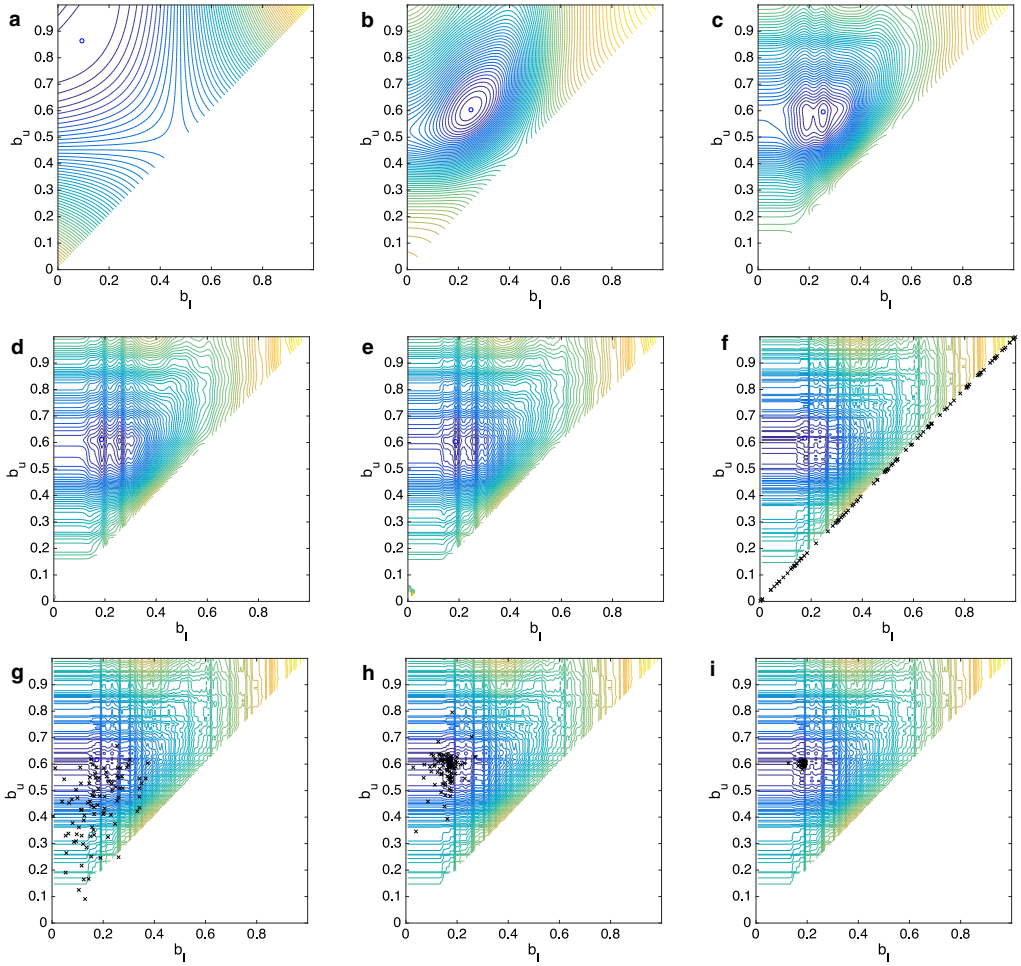


Fig. 2 Excessive brute-force computation of the objective function using **a** $\zeta = 2$, **b** $\zeta = 10$, **c** $\zeta = 30$, **d** $\zeta = 70$, **e** $\zeta = 100$ where the small blue circle encodes the approximated global minimum. Realization of PSO algorithm on **f** initial condition, **g** 5th-iteration, **h** 10th-iteration and **i** 20th-iteration

5 Prediction strategy

Here we pay our attention to a prediction algorithm. Making a prediction for the model involving clustered variables is not as instant as that for the simple model. Suppose that, with the help of a meteorological agency, the meteorological factors are predicted for several weeks in the future. To determine the predicted incidences, we need to cluster those factors. However, clustering cannot be defined unless we also know the incidences in the future. We are just encountering an indefiniteness.

Let Δ be the number of incidences on the first week in the prediction horizon. Using one row from the matrix equation (9), we see that the clustering-specific effects, as well as the clustered meteorological parameters, are dependent not only on the optimal barriers but also on Δ. In

Table 4 Results from model (9). AME, StD, Adj R^2, W, B–P, D–W–H, Wo stand for average marginal effect, standard deviation, Adjusted R^2, p-Value of Shapiro–Wilk W–test for normality, p-Value of Breusch–Pagan test for heteroscedasticity, p-Value of Durbin–Wu–Hausman test for a random-effect vs. a fixed-effect model, and Wooldridge test for the serial correlation, respectively

	Val	AME	StD	p-Value	F p-Value	1/VIF	R^2	Adj R^2	W	B–P	D–W–H	Wo
β_0	− .0463		.0097	0	0		.8496	.8489	0	0	1	.0031
β_1	− .0235		.0039	0		.8318						
β_2	− .0068		.0038	.076		.8861						
β_R^l	− .0909	2.5585×10^{-10}	.0299	.002		.5263						
β_R^m	.0349	7.9554×10^{-9}	.0066	0		.6135						
β_T^m	.0012	1.6785×10^{-8}	.0000	0		.1589						
β_H^u	.0058	1.6216×10^{-7}	.0001	0		.1496						
β_{-1}	.2383	.2724	.0149	0		.3278						
β_{-2}	.1245	.1962	.0158	0		.2944						
β_{-3}	.0518	.0147	.0156	.001		.2982						
β_{-4}	.0726	.0831	.0145	0		.3447						

other words, the "correct" Δ should satisfy one row of the matrix equation given the optimal coefficients. For short, this row equation reads as

$$\Delta = \mathcal{S}(\Delta). \tag{10}$$

For every fixed district index j, eq. (10) unfolds the scalar equation

$$\begin{aligned}\Delta_j = \beta_0 + \beta_j + \beta_R^l \delta_l(\Delta_j, b) r_j + \beta_R^m \delta_m(\Delta_j, b) r_j \\ + \beta_T^m \delta_m(\Delta_j, b) t_j + \beta_H^u \delta_u(\Delta_j, b) h_j + \sum_s \beta_{-s} \Delta_{-s,j}\end{aligned} \tag{11}$$

where $\Delta_{-s,j}$ represent all the incidences in the past 4 weeks and $b = (b_l, b_u)$ denote the optimal barriers. For $j \geq 3$, we have $\beta_j = 0$.

Algorithm 1 Fixed point iteration for prediction.

Require: error tolerance χ, $\Delta_{\text{pred}} := \emptyset$, and Δ_{-1}
for $i = 1, \cdots, n$ **do**
  set $\Delta^{(0)} := \Delta_{-1}$
  set $\delta := \chi + 1$
  set $k := 1$
  while $\delta \geq \chi$ **do**
    $\Delta^{(k)} := \mathcal{S}(\Delta^{(k-1)})$
    set $\delta := \|\Delta^{(k)} - \Delta^{(k-1)}\|$
    set $k := k + 1$
  end while
  set $\Delta_{\text{pred}} := \Delta_{\text{pred}} \cup \{\Delta^{(k-1)}\}$
  set $\Delta_{-1} := \Delta^{(k-1)}$
end for
Ensure: Δ_{pred}

Our key method to solve (10) is using the standard fixed point iteration. Algorithm 1 walks us through the basic idea. The procedure for a prediction test entails the division of the time window emanated from the original data into a testing and a prediction window. The testing window consists of 380 time points and the prediction window of 32 time points. Basically, we impose $n = 32$ in the algorithm. The optimal coefficients of the model (9) are calculated using the barriers produced by PSO only for the corresponding data in the testing domain. The resulting coefficients are then used to evaluate the prediction in `step 6` in Algorithm 1. We note that the definite values of the Heaviside functions in $\delta_l, \delta_m, \delta_u$ help designate the right-hand side of (11) having zero derivatives almost everywhere except at the barriers. The piecewise-affinity and continuous differentiability implying Lipschitz continuity suggest that the iterative algorithm converges as long as Δ_j never traces one of the barriers during iterations. One can appoint a sufficiently small Lipschitz constant for showing the contractivity of the right-hand side function. Violating the condition on not passing through a barrier might still lead to convergence, but is not guaranteed theoretically. The entire prediction result is then presented in Fig. 4. The t-test has been done to show that the error is normal with the mean zero at $\alpha = 0.1$ level.

6 Next generation operator and risk mapping

This section is devoted to the estimation of a number determining risk, relying solely on the incidence data. Therefore, the forthcoming discussions use the original density data without any transformation. The underlying idea is taken from the next generation operator in population dynamics. Accompanied by the significance of the autoregressive terms (cf. Table 4), we depart from the

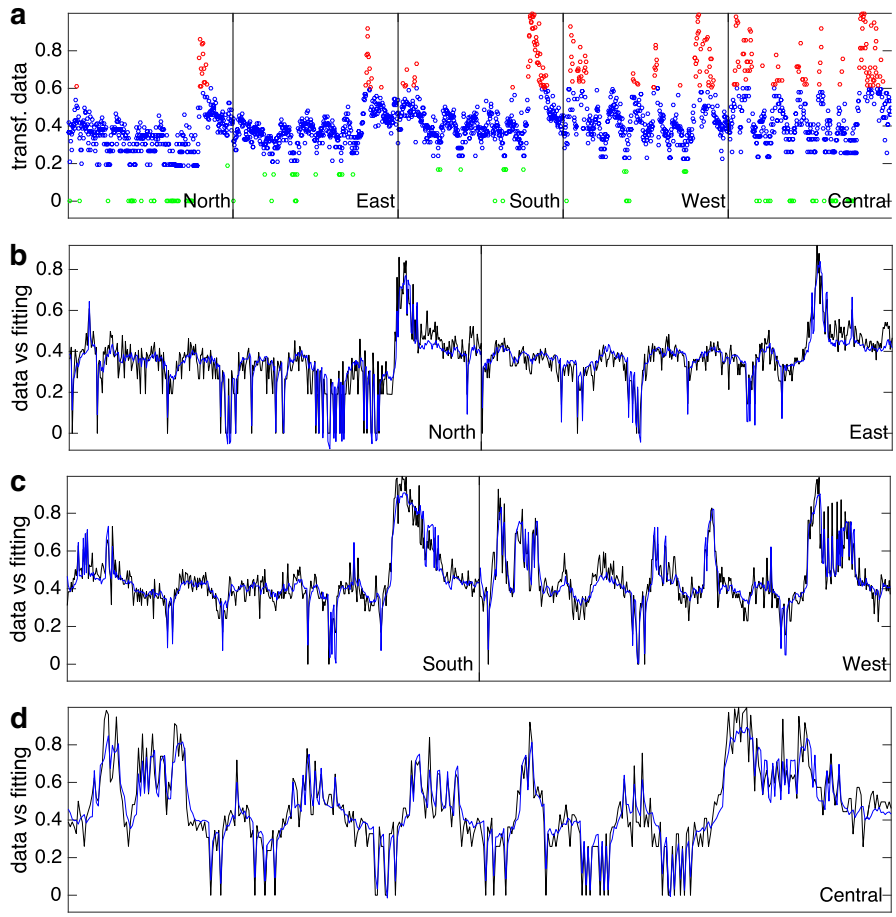


Fig. 3 **a** Depicts the data and clustering with the barriers found from the PSO algorithm. The low, medium, and high incidence levels are colored green, blue, and red, respectively. **b–d** Comparison between the model (9) using the coefficients' values in Table 4 (in blue) with the real data (in black)

assumption that the incidence cases in a certain district in the current week are attributed to those in the other districts in the past week. Of course, one can also consider the past two-week or three-week cases for more sophistication. Short incubation-to-illness period (Ogden and Lindsay 2016), short adult mosquito lifespan and mosquito traveling distance (Christophers 1960) are among several reasons why we focus on such "freshly" infected cases. Therefore, all cases happening at a certain region in the past week introduce marginal effects to the cases at the other regions in the current week. Recall our incidence panel data $D = (d_{ij})$ and let $D_i = (d_{i1} \cdots d_{iJ})^\top$ denote the incidence cases collected from all districts at time t_i. The basic model is then governed by the linear difference equation

$$D_{i+1} = \mu + GD_i, \quad \mu \geq 0, \tag{12}$$

where every element of G, g_{ij}, denotes the marginal effect of the incidence cases at district j to those at district i. The vector μ denotes a certain baseline determining to which level the D–sequence converges, when it does. In the present context, the matrix G serves as the *next generation operator*, since it determines the "next generation" of incidence cases by knowing the present cases. If d_{ij} were the number of individuals of an age class j in the population

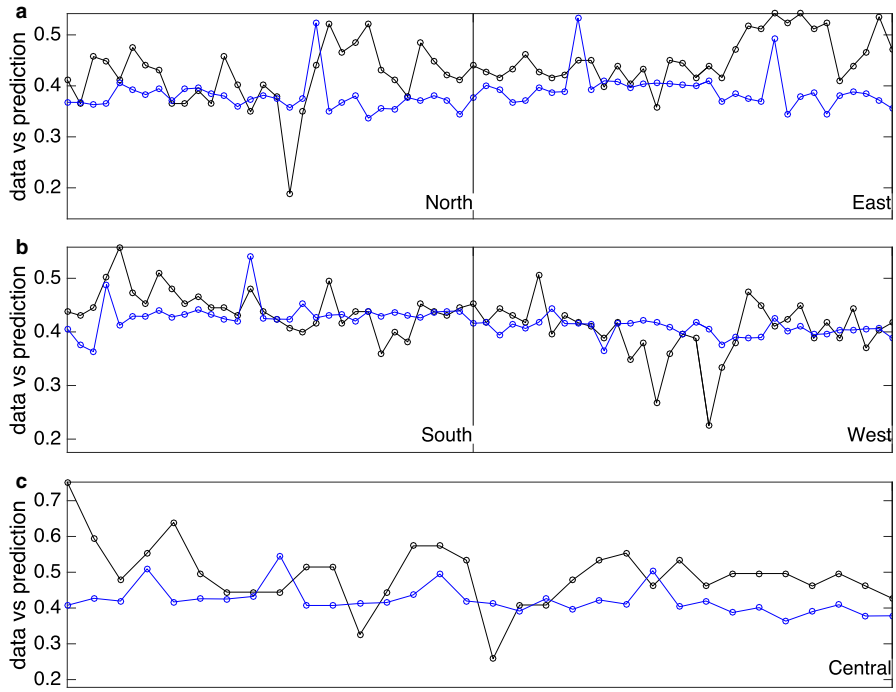


Fig. 4 Prediction result (in blue) based on Algorithm 1 in comparison with the real data (in black) on the 32-week prediction window

demography, then the matrix G is called *Leslie matrix* (Caswell 2000).

Due to stochastic nature of the data, the deterministic model defined above contains some seasonal error. Technically speaking, the error returns from the exclusion of unobservable entities, which may also be meteorological factors. Moreover, districts that experience large fluctuations in the incidences would need to be distinguished from others. In the present aim, we would like to model the uncertainty taking into account its state-dependence. Only if a proper noise in the model is found based on data can the risk mapping be more realistic. Our model including uncertainty is then governed by

$$D_{i+1} = \mu + GD_i + H(D_i)\xi_{i+1}, \tag{13}$$

where (ξ_i) are multivariate $\mathcal{N}(0,1)$–distributed random variables. Our next step is determining the state-dependent covariance function $H(D_i)$ via the representation of the stochastic difference equation (13) in a stochastic differential equation. We first reform the equation (13) into

$$D_{i+1} = D_i + \mu + [G - \mathrm{id}]D_i + H(D_i)\xi_{i+1}.$$

Observe that this equation is the Euler–Maruyama approximation (Kloeden and Platen 1992) for the stochastic differential equation

$$\mathrm{d}D = (\mu + [G - \mathrm{id}]D)\,\mathrm{d}t + H(D)\,\mathrm{d}W \tag{14}$$

using the step size of 1 week. All elements μ_i, $(g_{ij} - \delta_{ij})d_j$ (with $i,j = 1,\cdots,J$ and δ_{ij} denoting the Kronecker delta) represent "events", indexed by k where $k = 1,\cdots,J(J+1)$. The frequencies of these events are given by

$$\lambda_k := \begin{pmatrix} 0 \\ \vdots \\ \mathrm{sign}(\mu_i) \\ \vdots \\ 0 \end{pmatrix} \longleftarrow i$$

for $k = 1,\cdots,J$ and

$$\lambda_k := \begin{pmatrix} 0 \\ \vdots \\ \text{sign}(g_{ij} - \delta_{ij}) \\ \vdots \\ 0 \end{pmatrix} \longleftarrow i$$

for $k = J+1, \cdots, J(J+1)$. The probabilities of these events happening after a certain time step Δt are given by $P_k = |\mu_i|\Delta t$ for $k = 1, \cdots, J$ and $P_k = |g_{ij} - \delta_{ij}|d_j \Delta t$ for $k = J+1, \cdots, J(J+1)$. Using the idea in Allen and Burgin (2000), McCormack and Allen (2006), the incidence increase ΔD is assumed to follow the Gaussian process

$$\Delta D = \mathbb{E}(\Delta D) + \sqrt{\text{Var}(\Delta D)}\frac{\Delta W}{\sqrt{\Delta t}} \tag{15}$$

where W denotes the multidimensional Wiener process such that $\Delta W/\sqrt{\Delta t} \sim \mathcal{N}(0,1)$. We have $\mathbb{E}(\Delta D) = \sum_{k=1}^{J(J+1)} \lambda_k P_k = (\mu + [G - \text{id}]D)\Delta t$, which is our original *drift* in (14). Meanwhile,

$$\begin{aligned} \text{Var}(\Delta D) &= \mathbb{E}(\Delta D \Delta D^\top) - \mathbb{E}^2(\Delta D) \approx \mathbb{E}(\Delta D \Delta D^\top) \\ &= \sum_{k=1}^{J(J+1)} P_k \lambda_k \lambda_k^\top \end{aligned}$$

for a sufficiently small Δt. Since this formula is linear in Δt, then the factor Δt in the *noise* of (15) entirely disappears. We can now take $\Delta t \to 0$ in (15) to get (14), where then

$$H^2(D) = \sum_{k=1}^{J(J+1)} \begin{pmatrix} 0 & \cdots & & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & \cdots & P_k & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & & \cdots & 0 \end{pmatrix}.$$

Apparently, H^2 is just a linear function and has the format of a diagonal matrix. Surely is taking the square root of H^2 straightforward. Now with the function H being formulated, we are still worrying if this noise would lead the solution of the original model (13) jumps over larger values than the error emanated from the optimization. For this reason, we introduce a *damping factor* σ to control the volatility emanated from the state-dependent noise $H(D)$ and correct the model (13) into

$$D_{i+1} = \mu + GD_i + \sigma H(D_i)\xi_{i+1}. \tag{16}$$

The value of σ is to be estimated by how far the error from the deterministic model delineates that from the datasets. Under the model (16), we know that the endemic equilibrium $D^* := \mathbb{E}(D_i) = [\text{id} - G]^{-1}\mu$ is stochastically stable (i.e. stable in probability) providing that the maximal real part of the eigenvalues of the Jacobian $[G - \text{id}]$ is negative. The latter happens if all the eigenvalues of G lie in the open unit disc in $\mathbb{C}$, which is the case if the spectral radius $\rho(G) < 1$. Moreover, the variance $\text{Var}(D_i) = [\text{id} - G^2]^{-1}\sigma^2 H^2(D_{i-1})\Sigma$, where $\Sigma = \text{Var}(\xi_i)$, appears to be state-dependent.

Suppose that G_+, G_- are nonnegative matrices containing absolute entries of G that are nonnegative and negative, respectively. Otherwise, their elements are zero. We have $G_+ - G_- = G$. If $\|G_-\|$ is sufficiently small, then the endemic equilibrium enjoys the approximation

$$\begin{aligned} D^* = \ & [\text{id} - G_+ + G_-]^{-1}\mu \approx [\text{id} - G_+]^{-1}\mu - [\text{id} - G_+]^{-1} \\ & G_-[\text{id} - G_+]^{-1}\mu. \end{aligned}$$

When non-zero entries of G_- are solely off-diagonal, then a contradiction can be done to show that $\rho(G) < 1$ implies $\rho(G_+) < 1$. Therefore, we obtain $D^* \approx [\text{id} - G_+]^{-1}\mu$ in the first-order approximation, which is always nonnegative due to the property of M–matrix $[\text{id} - G_+]$, see (Berman and Plemmons 1994). On optimality, D^* approximates the means of the datasets since otherwise the solution of (16) deviates away from the data. This implies that when $\rho(G_+)$ is close enough to 1 such that $[\text{id} - G_+]^{-1}$ is ready to blow up, μ adjusts to be small enough. Moreover, $D^* = (d_1^*, \cdots, d_J^*) \to \mu$ when $\rho(G) \to 0$.

Let d_{sep} denote a barrier separating the *low-risk regime* $(d_j^* < d_{\text{sep}})$ and *high-risk regime* $(d_j^* > d_{\text{sep}})$. In the low-risk regime, one would say that the disease is such insignificant that the surveillance and treatments can be kept in the current magnitudes. Based on the biological existence of the endemic equilibrium, we estimate a risk measure as $\mathcal{R}^j = A_0 + A_1\left(d_j^* - d_{\text{sep}}\right)$. Two conditions are assigned in order to determine the unobservable parameters A_0, A_1. Here, we assume that $\mathcal{R}^j = 1$ when $d_j^* = d_{\text{sep}}$, implying $A_0 = 1$. We also assume that under the low-risk regime, $\rho(G)$ dominates the value of $\mathcal{R}^j$ and eventually $\mathcal{R}^j = \rho(G)$ when $d_j^* = 0$. We acquire $A_1 = (1 - \rho(G))/d_{\text{sep}}$ and finally

$$\mathcal{R}^j := \rho(G) + (1 - \rho(G))\frac{d_j^*}{d_{\text{sep}}}. \tag{17}$$

The formula in (17) assures the nonnegativity of $\mathcal{R}^j$, which has a certain biological implication. Observe that under the low risk regime $(d_j^* < d_{\text{sep}})$, it always holds $\mathcal{R}^j < 1$ independent of $\rho(G)$. The value $\mathcal{R}^j$ gets even smaller as $\rho(G) \ll 1$. Only when d_j^* is small enough does $\rho(G)$ determine $\mathcal{R}^j$. Under the high risk regime $(d_j^* > d_{\text{sep}})$,

$\mathcal{R}^j > 1$ but $\mathcal{R}^j$ gets larger as $\rho(G) \ll 1$. Note that when $\rho(G) \ll 1$, the noise emanated from (16) exhibits small jumps, meaning that the smaller $\rho(G)$, the more attractive the endemic equilibrium is or the higher the sureness that the solution trajectory converges to the endemic equilibrium. As the regional risk measure gives individual persistence of the disease, we define the global risk measure as their mean

$$\mathcal{R} := \sqrt[J]{\mathcal{R}^1 \cdots \mathcal{R}^J}. \tag{18}$$

The geometric mean is used in the formula in order to avoid the severe effect of outliers. The facts that the datasets have positive means, and apparently, $\rho(G) \gg 0$ avoid any single $\mathcal{R}^j$ being close to zero. In this study, the risk measure (18) clearly relies on the definition of the threshold parameter d_{sep}, under and above which every district or region can be categorized as of low and high risk. This allows one to share the responsibility in defining the transmission severity from a modeler to the decision-maker.

For a test case, we define $d_{\text{sep}} := (\max_j d_j^*)/2 \approx 5.49 \times 10^{-6}$, roughly approximating half of the maximal mean of the incidence datasets or 55 per 100,000 inhabitants. This yields the regional risk measures .9798, .9914, 1.0556, 1.0660, 1.0986 and global risk measure $\mathcal{R} \approx 1.0373$. With the chosen d_{sep}, North and East Jakarta are essentially of low risk, whereas South, West, and Central Jakarta are essentially of high risk. Additionally, attributed to $\rho(G) \approx .9063$ is the weak stochastic stability of all the equilibrium states. This advocates the situation that all the districts remain to be of considerable non-vanishing risk.

7 Spatial autocorrelation

Spatial autocorrelation, in the present context, performs a comparison of the incidence levels of a certain district with those of its neighbors. The underlying concept thus relies on the definition of *close neighbors*, i.e., a set of neighboring districts in which commuting between neighbors is possible. In the presence of spatial autocorrelation, the incidence levels are either positively or negatively linked to each other. A positive linkage indicates similar incidence levels appearing among neighbors from a certain timestamp, which is called *spatial mutualism*. When dissimilar incidence levels are instead observed, a negative linkage or *spatial competition* happens.

7.1 Global autocorrelation coefficient

The widely-used idea for computing the spatial autocorrelation coefficient is due to Moran (1950). In our case, we consider $J = 5$ close neighbours. Let d_{ij} ($i = 1, \cdots, I$ and $j = 1, \cdots, J$) be the incidence cases of the close neighbours during the discrete time points $t_1, \cdots, t_I$. By writing

$$\bar{d} = \frac{\sum_{i,j} d_{ij}}{IJ} \quad \text{and} \quad z_{ij} = d_{ij} - \bar{d} \quad (\text{i.e. } \sum_{i,j} z_{ij} = 0),$$

Moran's autocorrelation coefficient is defined as

$$r := \left(\frac{IJ}{2IJ - I - J}\right) \frac{\sum_i \sum_{j=1}^{J-1} z_{ij} z_{i,j+1} + \sum_{i=1}^{I-1} \sum_j z_{ij} z_{i+1,j}}{\sum_{i,j} z_{ij}^2}.$$

The mean and standard deviation of this statistic are presented in the original paper (Moran 1950), allowing to compute p-Value after normalization thus perform a statistical inference under sufficiently large datasets. The coefficient is outlined by three numbers: -1 (perfectly negative autocorrelation), 0 (no autocorrelation) and 1 (perfectly positive autocorrelation). Since the coefficient is parameter-free, a straightforward computation returns $r = 0.7530$. This suggests that the close neighbors co-vary positively on the overall observation time (p-Value $= 0.623$). Practically, we are suggested not to only worry on some but all districts during epidemics.

7.2 Time-dependent autocorrelation coefficient

To break down the autocorrelation measurement on the weekly level, a time-dependent coefficient is used instead. Let us freeze time and consider J close neighbours with incidence levels d_j ($j = 1, \cdots, J$). Denote by w_{jk} ($j, k = 1, \cdots, J$) the *spatial connectivity measure* between jth and kth district. Write $J\bar{d} = \sum_j d_j$ and $z_j = d_j - \bar{d}$. The so-called Moran's time-dependent autocorrelation coefficient (Cliff and Ord 1981; Anselin 1995) is presented as

$$r_{\text{M}} := \left(\frac{J}{\sum_{j,k} w_{jk}}\right) \frac{\sum_j \sum_k w_{jk} z_j z_k}{\sum_j z_j^2}.$$

A drawback of Moran's autocorrelation coefficient is its instability for general problems, as seen shortly. Datasets representing the same title can have different values in case there are several parties who do the same measurement/estimation (e.g. local and international bodies), or can lose certainty during data collection and transfer. We can thus generate noise over the incidence data Δd, defined

uniformly for all d_j, i.e. $d_j \mapsto d_j + \Delta d$. An example could be $\Delta d = \sigma_d Z$ for some σ_d where $Z \sim \mathcal{N}(0,1)$. Due to such noise, Moran's correlation coefficient suffers from the following change

$$r_{\mathrm{M}} \mapsto \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} r_{\mathrm{M}} + \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} \left(\frac{J}{\sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk}(z_j + z_k)}{\sum_j z_j^2} \Delta d + \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} \left(\frac{J}{\sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk}}{\sum_j z_j^2} \Delta d^2.$$

It is clear that the perturbed coefficient depends on not only the incidences but also weighting measures, cf. (Upton and Fingleton 1985).

The more stable coefficient was introduced by Geary (Geary 1954; Anselin 1995), which is called Geary's autocorrelation coefficient

$$r_{\mathrm{G}} := \left(\frac{J-1}{2 \sum_{j,k} w_{jk}} \right) \frac{\sum_j \sum_k w_{jk} (d_j - d_k)^2}{\sum_j z_j^2}.$$

The less sensitivity is told by the behaviour of the coefficient with respect to perturbation over the data. Introducing noise as before, Geary's autocorrelation coefficient takes the following change

$$r_{\mathrm{G}} \mapsto \frac{\sum_j z_j^2}{\sum_j (z_j + \Delta d)^2} r_{\mathrm{G}}.$$

One can see that the perturbed Geary's r_{G} only takes the first term of Moran's r_{M}, which solely depends on the incidences. According to Cliff and Ord (1981), $\mathbb{E}(r_{\mathrm{G}}) = 1$ represents no spatial autocorrelation, while $r_{\mathrm{G}} < 1$ and $r_{\mathrm{G}} > 1$ point to a positive and a negative spatial autocorrelation among the close neighbours. A high similarity is indicated by the situation where $d_j \simeq d_k$ for any j, k and therefore, r_{G} close to 0.

To compute the time-dependent coefficient, we require to define the connectivity measures w_{jk}. A first insight was gained from the 2014 commuter data in Jakarta, representing the average daily percentages of commuters from all the regions (Irawan and Pragesari 2014). However, the same percentages from two different districts of discrepant populations inevitably hide different proportions. To avoid bias, we instead use the proportions, each of which represents the percentage multiplied by the total population of the district where the commuters depart. The inability to access similar data from different years leads us to constant

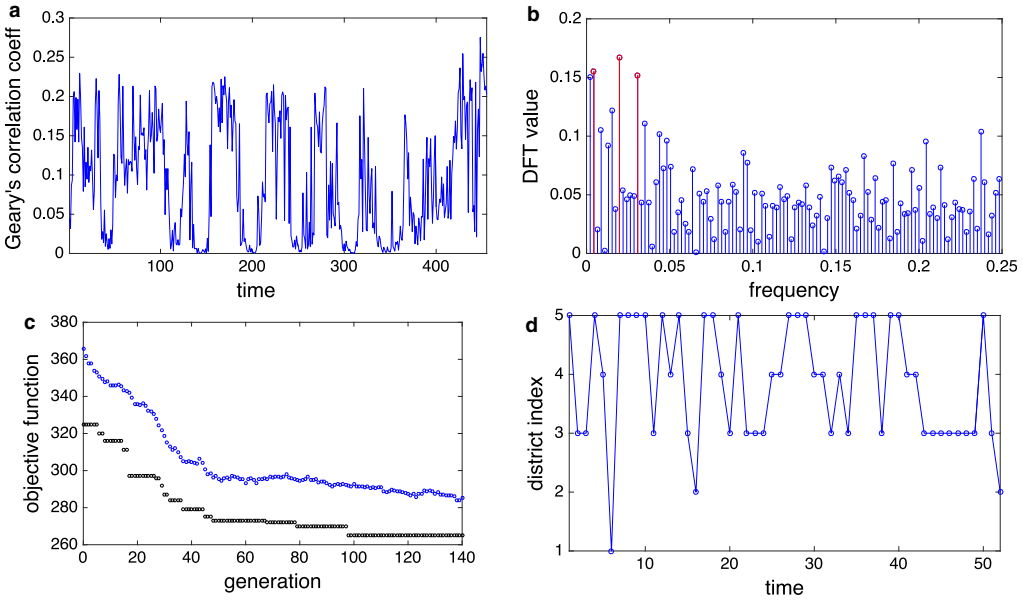


Fig. 5 **a** Presents the computation result of Geary's autocorrelation coefficient. **b** Displays the Fourier amplitudes related to the frequencies of the index data, with the leading frequencies (in red) correspond to the period of 52 weeks, 227.5 weeks, and 32.5 weeks, respectively. **c** Depicts the computed best (in black) and average objective function value from (19) using a genetic algorithm. **d** The optimal periodic solution from (19) where index 1 to 5 represent North, East, South, West, and Central Jakarta, respectively

percentages but time-dependent w_{jk} due to varying populations, see Table 1. The computation result of timely r_G can be seen in Fig. 5a. The mean value $\bar{r}_G \approx .0934$ reveals that all the close neighbors co-vary significantly positively from time to time, amplifying the global autocorrelation.

8 Outbreak propagation

This section asks how far the trajectory for the highest peak of incidences can be learned. In case the highest peak shows spatial periodicity, it would help decision-maker see which districts the highest peaks locate on a certain prediction horizon. As a result, the targeting for the intervention measures can be more accurate. Technically, we aim to locate the maximum of the incidence level for every week that correspond to which district giving the maximum. Then, data containing district indices related to such highest epidemics can be presented as a time series $\text{idx} = (\text{idx}_i)$. To see the periodicity, we calculate the Fourier spectrum of the created seasonal data using Fast Fourier Transform. The result is given in Fig. 5b. One can see that the leading low frequencies are not clearly distinguished from the high frequencies, meaning that the index data are weakly periodic. For the sake of bearing a rule of thumb or guidance for control interventions, revealing a certain spatial periodicity for the peak of outbreaks can still be important. Based on the Fourier analysis, the computed leading frequency corresponds to the period of 52 weeks. The period suggests that the incidence peaks resemble annual cases. Now we can generate a periodic function of period 52 weeks that is as close as possible to the index data. The corresponding optimization problem reads as

$$\begin{aligned} \min_{x\in\{1,\cdots,J\}^I} \quad & \sum_{i\in I} 1-\delta(x_i,\text{idx}_i) \\ \text{s.t.} \quad & x_i = x_{i+52},\ i=1,\cdots,I-52, \end{aligned}$$

where $\delta(x,y)$ denotes the Kronecker delta function. This is a constrained integer optimization problem with the cardinality of the solution space $J^I = 5^{412}$. A first step to treat the problem is redesigning it to integrate the constraint into the search space, which returns

$$\begin{aligned} \min_{x\in\{1,\cdots,J\}^{52}} \quad & \sum_{j=0}^{\lfloor I/52\rfloor-1}\sum_{i=1}^{52} 1-\delta(x_i,\text{idx}_{52j+i}) \\ & + \sum_{i=1}^{I-\lfloor I/52\rfloor 52} 1-\delta(x_i,\text{idx}_i). \end{aligned} \tag{19}$$

Now the cardinality reduces to $J^{52} = 5^{52}$. This study employs a genetic algorithm to solve (19) engineering 20 players (individuals) in every generation. The best objective function value and the mean of all the 20 values for all generations can be seen in Fig. 5c. On the 97th-iteration onwards, the best objective function value turns to be constant at 265. This means that only 18 out of 52 data points per 52 weeks can be fitted by the optimal periodic solution x as depicted in Fig. 5d. This figure is in line with the finding on the risk measure, that North and East Jakarta are considered of low risk as the peaks only occur once and twice per year, respectively.

9 Concluding remarks

We derive a panel random-effect model describing a direct relationship between lagged meteorological factors (rainfall, temperature, humidity) and dengue incidence level. We show that rainfall has the time lag ranging between 6w to 11w, temperature between 21w to 44w, humidity between 6w to 10w in the past leading to the current magnitudes of incidence levels for all the five districts in Jakarta. Based on the model, all the meteorological factors are shown to influence the incidence level significantly positively. The fact that a small coefficient of determination leads to biased error in the prediction window (not zero-mean normal) and that the incidence data suffer from the zero-inflated problem, we initiate a clustering to the incidence data. The timely meteorological factors that correspond to a certain cluster are classified as to cause the incidence level stays in the cluster. Under three clusters, this strategy triples the number of explanatory variables, poised to achieve better fittings. The optimal barriers minimizing the mean squared error are sought with the aid of a Particle Swarm Optimizer applied to the corresponding constrained, non-smooth optimization problem. The new clustering-integrated model with the optimal barriers returns a relatively high coefficient of determination. As the classification on the meteorological factors is based on the incidence level, making a prediction with predicted meteorological factors is not straightforward. A fixed point iteration is used in this study to deal with such indefiniteness. The prediction result is shown to underlie unbiased error for the forthcoming 32 weeks at $\alpha = 0.1$ level. We then propose a certain formula for estimating the (regional) risk measure relying only on the incidence data. The handling equation is a vector autoregression model with state-dependent noise. The noise is determined via connecting the model with the stochastic differential equation counterpart. The endemic equilibrium is shown to be stochastically stable, with the regional risk measure either smaller than 1 (low-risk) or greater than 1 (high-risk). Under a test case for the threshold incidence level, we show that North and East Jakarta are relatively of low risk while South, West, and Central Jakarta are relatively of high risk. To see the spatial connection among the districts, we calculate the

global and local spatial autocorrelation coefficient. Our computation shows that all the districts co-vary highly positively in terms of epidemics occurrence, both generally and timely. However, the local government might still need to know which district requires the most attention per week. We then compute the highest peak of the outbreaks from all the districts and study its behavior over time. Using Fourier analysis, we see that the highest peak resembles annual periodicity in the propagation among all the districts. The recurrence of the peak combined with the risk measures suggest that the following districts require the most to the least attention during epidemics and inter-epidemics: Central, South, West, East, and North Jakarta.

We point out several limitations of this study. First, the population data in Table 1 show an increasing pattern in time due to imports, with an anomaly in the year 2011 as the numbers suddenly jump high, especially in Central Jakarta. We failed to find the reason why. Second, our study was not compared with the usual treatment of dropping time points corresponding to zero incidence levels. We based our reasoning for doing this on having limited data points on the testing window, which will return weak prediction power. Third, since the final model only incorporates meteorological factors and autoregressive terms, the prediction accuracy highly depends on those entities. Moreover, we observe that the prediction outcome only affords unbiased error, and no essential information regarding absolute error is required since in our case, the explanatory variables are covariance-stationary, i.e., the mean and variance of each time series do not change over time. To maximize the prediction power, another prediction of incidence levels from medical experts combined with predicted meteorological factors can be included as a certain regulator for the updated epidemic alerts. Fourth, both the simple and clustering-integrated model return heteroscedastic error. Consequently, the error is steered by some of the explanatory variables, and the prediction can be harmed by possible irregularities in the predicted values of those variables. It is the nature of the disease to interplay with various confounding factors other than climate, for example, the socio-economic elements inspired by several studies mentioned in the introduction part. A future study using such additional instrumental variables is thus needed to clear up the heteroscedasticity issue. Fifth, the stochastic stability of the endemic equilibrium from the vector autoregression model is independent of the regional risk measures, but the latter depend on the threshold incidence level of d_{sep}. Such a threshold d_{sep} should carefully be estimated to minimize false alarms that would inflict the epidemic response capabilities of certain districts.

Acknowledgements DA has financially been supported by Universitas Indonesia through PUTI KI Q2 research grant scheme 2020 [No. NKB-775/UN2.RST/HKP.05.00/2020] and MF by the Indonesia Ministry of Research and Technology through PMDSU Program [No. 1511/E4.4/2015].

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest existing in the submission of the manuscript.

References

Allen LJS, Burgin AM (2000) Comparison of deterministic and stochastic SIS and SIRS models in discrete time. Math Biosci 163:1–33

Anselin L (1995) Local Indicators of Spatial Association - LISA. Geographic Anal 27(2):93–115

Arcari P, Tapper N, Pfueller S (2007) Regional variability in relationships between climate and dengue/DHF in Indonesia. Singap J Trop Geogr 28:251–272

Astuti EP, Dhewantara PW, Prasetyowati H, Ipa M, Herawati C, Hendrayana K (2019) Paediatric dengue infection in Cirebon, Indonesia: a temporal and spatial analysis of notified dengue incidence to inform surveillance. Parasit Vect 12(1):186

Berman A, Plemmons RJ (1994) Nonnegative Matrices in the Mathematical Sciences. Society for Industrial and Applied Mathematics, Philadelphia

Bouzid M, Colón-González FJ, Lung T, Lake IR, Hunter PR (2014) Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever. BMC Public Health 14:781–12

Brits R, Engelbrecht A, van den Bergh F (2002) A niching particle swarm optimizer. In: Wang L (ed) Proceedings of the Fourth Asia-Pacific conference on simulated evolution and learning, Nanyang Technological University, School of Electrical & Electronic Engineering, Singapore, pp 692–696

Byttebier B, Majo MSD, Fischer S (2014) Hatching response of Aedes aegypti (Diptera: Culicidae) eggs at low temperatures: Effects of hatching media and storage conditions. J Med Entomol 51(1):97–103

Cameron AC, Trivedi PK (2010) Microeconometrics Using Stata, Revised edn. Stata Press, College Station

Caswell H (2000) Matrix population models: construction, analysis, and interpretation, 2nd edn. Sinauer Associates Inc., Sunderland

Chen EH (1971) The power of the Shapiro-Wilk W test for normality in samples from contaminated normal distributions. J Am Stat Assoc 66(336):760–762

Chen SS, Gopinath RA (2000) Gaussianization. In: Leen TK, Dietterich T, Tresp V (eds) NIPS'00: Proceedings of the 13th international conference on neural information processing systems. MIT Press, Cambridge, pp 402–408

Chien LC, Yu HL (2014) Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. Environ Int 73:46–56

Christophers SR (1960) Aedes aegypti: the Yellow fever Mosquito: its life, bionomics, and structure. Cambridge University Press, London

Cliff AD, Ord JK (1981) Spatial Processes. Pion, London

Davidson R, MacKinnon JG (1993) Estimation and inference in econometrics. Oxford University Press, Oxford

de Almeida Costa EAP, de Mendonca Santos EM, Correia JC, de Albuquerque CMR (2010) Impact of small variations in temperature and humidity on the reproductive activity and

survival of Aedes aegypti (Diptera, Culicidae). Revista Brasileira de Entomologia 54(3):488–493
Deng SQ, Yang X, Wei Y, Chen JT, Wang XJ, Peng HJ (2020) A review on dengue vaccine development. Vaccines 8(1):63–13
Fakhruddin M, Putra PS, Wijaya KP, Sopaheluwakan A, Satyaningsih R, Komalasari KE, Mamenun Sumiati, Indratno SW, Nuraini N, Götz T, Soewono E (2019) Assessing the interplay between dengue incidence and weather in Jakarta via a clustering integrated multiple regression model. Ecol Comple 39:100768–8
Farrar DE, Glauber RR (1967) Multicollinearity in regression analysis: the problem revisited. Rev Econ Stat 49(1):92–107
Geary RC (1954) The contiguity ratio and statistical mapping. Incorpor Stat 5(3):115–146
Gujarati DN, Porter DC (2008) Basic econometrics, 5th edn. MCGraw Hill Book Co., New York
Hanusz Z, Tarasinska J, Zielinski W (2016) Shapiro-Wilk test with known mean. REVSTAT-Stat J 14(1):89–100
Harrington LC, Fleisher A, Ruiz-Moreno D, Vermeylen F, Wa CV, Poulson RL, Edman JD, Clark JM, Jones JW, Kitthawee S et al (2014) Heterogeneous feeding patterns of the dengue vector, Aedes aegypti, on individual human hosts in rural Thailand. PLoS Negl Trop Dis 8(8):e3048-16
Haryanto B (2016) Health adaptation scenario and dengue fever vulnerability assessment in Indonesia. Advances in Asian Human-Environmental Research. In: Akhtar R (ed) Climate Change and Human Health Scenario in South and Southeast Asia. Springer, Cham, pp 221–236
Irawan N, Pragesari NN (2014) Statistik Komuter Jabodetabek: Hasil Survei Komuter Jabodetabek 2014. Statistics Indonesia, Jakarta
Jansen CC, Beebe NW (2010) The dengue vector Aedes aegypti: what comes next. Microbes Infect 12(4):272–279
Johnston J (1972) Econometric methods, 2nd edn. McGraw Hill Higher Education, Pennsylvania
Juliano SA, O'Meara GF, Morrill JR, Cutwa MM (2002) Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes. Oecologia 130:458–469
Kakarla SG, Caminade C, Mutheneni SR, Morse AP, Upadhyayula SM, Kadiri MR, Kumaraswamy S (2019) Lag effect of climatic variables on dengue burden in India. Epidemiol Infect 147:e170(1)–10
Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Springer, Berlin
Kusriastuti R, Sutomo S (2005) Evolution of dengue prevention and control programme in Indonesia. Dengue Bull 29:1–7
Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CA, Carvalho MS, 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
Mansfield ER, Helms BP (1982) Detecting multicollinearity. Am Stat 36(3a):158–160
McCormack RK, Allen LJS (2006) Multi-patch deterministic and stochastic models for wildlife diseases. J Biol Dyn 1:63–85
Moran PA (1950) Notes on continuous stochastic phenomena. Biometrika 37(1/2):17–23
Morin CW, Comrie AC, Ernst K (2013) Climate and dengue transmission: evidence and implications. Environ Health Perspect 121(11–12):1264–1272
Ogden NH, Lindsay LR (2016) Effects of climate and climate change on vectors and vector-borne diseases: ticks are different. Trends Parasitol 32(8):646–656
Ramachandran VG, Roy P, Das S, Mogha NS, Bansal AK (2016) Empirical model for estimating dengue incidence using temperature, rainfall, and relative humidity: a 19-year retrospective analysis in East Delhi. Epidemiol Health 38:e2016052–8
Ramadona AL, Lazuardi L, Hii YL, Holmner A, Kusnanto H, Rocklöv J (2015) Prediction of dengue outbreaks based on disease surveillance and meteorological data. PLoS ONE 11(3):e0152688–18
Rana S, Jasola S, Kumar R (2011) A review on particle swarm optimization algorithms and their applications to data clustering. Artif Intell Rev 35:211–222
Schmidheiny K (2019) Panel data: fixed and random effects. Unversität Basel, Basel
Shi Y, Liu X, Kok SY, Rajarethinam J, Liang S, Yap G, Chong CS, Lee KS, Tan SS, Chin CKY, Lo A, Kong W, Ng LC, Cook AR (2016) Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in Singapore. Environ Health Perspect 124(9):1369–1375
Spiegel J, Bennett S, Hattersley L, Hayden MH, Kittayapong P, Nalim S, Wang DNC, Zielinski-Gutiérrez E, Gubler D (2005) Barriers and bridges to prevention and control of dengue: the need for a social-ecological approach. EcoHealth 2(4):273–290
Suwandono A, Kosasih H, Kusriastuti R, Harun S, Ma'roef C, Wuryadi S, Herianto B, Yuwono D, Porter KR, Beckett CG et al (2006) Four dengue virus serotypes found circulating during an outbreak of dengue fever and dengue haemorrhagic fever in Jakarta, Indonesia, during 2004. Trans R Soc Trop Med Hyg 100(9):855–862
Trelea IC (2003) The particle swarm optimization algorithm: convergence analysis and parameter selection. Inf Process Lett 85:317–325
Upton GJG, Fingleton B (1985) Spatial data analysis by example, volume 1: point pattern and quantitative data, 1st edn. Wiley Series in Probability and Statistics: Applied Probability and Statistics Section (Book 182), Wiley
van den Bergh F (2002) An Analysis of Particle Swarm Optimizers. PhD thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa
van den Bergh F, Engelbrecht A (2006) A study of particle swarm optimization particle trajectories. Inf Sci 176:937–971
van den Bergh F, Engelbrecht AP (2001) Effects of swarm size on cooperative particle swarm optimisers. In: Spector L, Goodman ED, Wu A (eds) GECCO'01: Proceedings of the 3rd annual conference on genetic and evolutionary computation, Morgan Kaufmann Publishers Inc., San Francisco, CA, pp 892–899
Wang D, Tan D, Liu L (2018) Particle swarm optimization algorithm: an overview. Soft Comput 22(2):387–408
Wang X, Tang S, Wu J, Xiao Y, Cheke RA (2019) A combination of climatic conditions determines major within season dengue outbreaks in Guangdong Province, China. Parasit Vect 12(45):1–10
Wijaya KP, Aldila D, Schäfer LE (2019) Learning the seasonality of disease incidences from empirical data. Ecol Comple 38:83–97
Wilder-Smith A (2020) Dengue vaccine development: status and future. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 63(1):40–44
Willis CE, Perlack RD (1978) Multicollinearity: effects, symptoms, and remedies. J Northeastern Agric Econ Council 7(1):55–61
Withanage GP, Viswakula SD, Gunawardena YINS, Hapugoda MD (2018) A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. Parasit Vect 11:262–310
Xu L, Stige LC, Chan KS, Zhoud J, Yang J, Sang S, Wang M, Yang Z, Yan Z, Jiang T, Lu L, Yue Y, Liu X, Lin H, Xu J, Liu Q, Stenseth NC (2017) Climate variation drives dengue dynamics. Proc Nat Acad Sci USA 114(1):113–118
Yu HL, Yang SJ, Yen HJ, Christakos G (2011) A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan. Stoch Environ Res Risk Assess 25:485–494

Publication III

Ganegoda, N., Wijaya, K.P., Amadi, M., Erandi, K.K.W., and Aldila, D.
Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany

Reprinted with permission from
Scientific reports
Vol. 11, pp. 1–16, 2021.

scientific reports

Check for updates

OPEN

Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany

Naleen Chaminda Ganegoda[1], Karunia Putra Wijaya[2], Miracle Amadi[3], K. K. W. Hasitha Erandi[4] & Dipo Aldila[5]✉

COVID-19 pandemic continues to obstruct social lives and the world economy other than questioning the healthcare capacity of many countries. Weather components recently came to notice as the northern hemisphere was hit by escalated incidence in winter. This study investigated the association between COVID-19 cases and two components, average temperature and relative humidity, in the 16 states of Germany. Three main approaches were carried out in this study, namely temporal correlation, spatial auto-correlation, and clustering-integrated panel regression. It is claimed that the daily COVID-19 cases correlate negatively with the average temperature and positively with the average relative humidity. To extract the spatial auto-correlation, both global Moran's $\mathcal{I}$ and global Geary's $\mathcal{C}$ were used whereby no significant difference in the results was observed. It is evident that randomness overwhelms the spatial pattern in all the states for most of the observations, except in recent observations where either local clusters or dispersion occurred. This is further supported by Moran's scatter plot, where states' dynamics to and fro cold and hot spots are identified, rendering a traveling-related early warning system. A random-effects model was used in the sense of case-weather regression including incidence clustering. Our task is to perceive which ranges of the incidence that are well predicted by the existing weather components rather than seeing which ranges of the weather components predicting the incidence. The proposed clustering-integrated model associated with optimal barriers articulates the data well whereby weather components outperform lag incidence cases in the prediction. Practical implications based on marginal effects follow posterior to model diagnostics.

Viral diseases emerge with complex transmission dynamics, and they are hard to eradicate challenging capacity of testing, diagnosis, and cure[1,2]. Such complexity is generated by various factors such as genetic changes of the virus, environmental influences, and host behavior[3,4]. COVID-19 caused by the coronavirus SARS-CoV-2 has also shown its revolutionary dynamics via all those routes, leaving the world at a standstill in many aspects. The transmission of coronavirus occurs and escalates in diverse means. Most notable drivers include direct contact with infectious individuals[5], fomite transmission via contaminated surfaces[6,7], transmission via virus-carrying aerosols[8,9], congested living and mobility leading to superspreading events[10–13], and lack of compliance to health guidelines[14–17]. Though both direct and indirect transmission are recognized, the influence of outdoor aerosol transmission is not properly understood[18,19]. Meanwhile, within-household is much higher compared to cross-household transmission leaving home quarantine also at risk[20]. Thus, planning healthcare and interventions has also become challenging. It is further problematic due to the presence of asymptomatic cases[21].

Transmission and morbidity of COVID-19 can be worsened when co-infections with other respiratory viruses are present. Several clinical studies from different countries have observed the co-infection of COVID-19 with other viral infections[22–24]. The most common respiratory viruses are influenza virus, respiratory syncytial virus, parainfluenza viruses, metapneumovirus, rhinovirus, adenoviruses, bocaviruses, and coronaviruses[25]. These

[1]Department of Mathematics, University of Sri Jayewardenepura, Nugegoda 10250, Sri Lanka. [2]Mathematical Institute, University of Koblenz, 56070 Koblenz, Germany. [3]Department of Mathematics and Physics, Lappeenranta University of Technology, 53851 Lappeenranta, Finland. [4]Department of Mathematics, University of Colombo, Colombo 00300, Sri Lanka. [5]Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia. ✉email: aldiladipo@sci.ui.ac.id

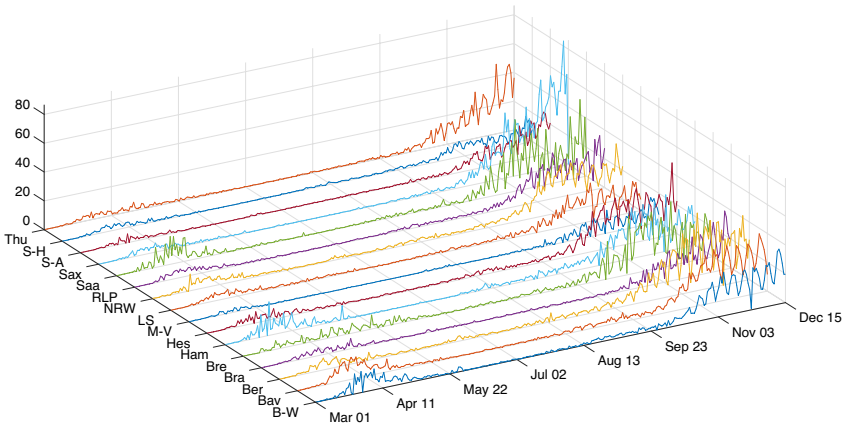
viral infections share common symptoms such as sneezing, cough, sore throats, and fever while following similar ways of transmission[26,27]. Influenza viruses that cause seasonal flu would easily co-exist with COVID-19 in the winter season[28]. This is motivated by the fact that most respiratory pathogens are seasonal[29,30]. Thus, given that many COVID-19 infected cases are undetected[31], sneezing and cough due to another infection may allow passing respiratory droplets carrying SARS-CoV-2 too. Although the information is still limited, one cannot set aside the possible risk of excessive COVID-19 spread due to co-infection[32,33]. In this regard, timely detection is important to curtail issues of missed diagnoses[34].

The influence of weather components such as temperature and relative humidity on the transmission of SARS-CoV-2 is investigated recently. Related studies have been motivated by the fact that temperature and relative humidity also regulated the survival of coronaviruses of SARS[35–38] and MERS[39,40]. Respiratory droplets play a key role in transmission, subsequently more structured with aerosols and fomites[41,42]. Due to other confounding factors related to specific geographical areas, mixed findings can be expected with different levels of temperature and relative humidity[43–46]. Using panel regressions, a study of 20 countries having the most number of confirmed cases[47] suggested that high temperature and relative humidity reduce transmission, while low temperatures are contributory for activation and infectivity of the virus. A low temperature range (− 6.28 °C to + 14.51 °C) has been identified as favorable to COVID-19 growth in[48] via a statistical estimation. This study also found that a 1 °C rise in temperature can reduce the number of cases by 13–17 per day. On the contrary, a study covering many cities in China[49] using a generalized additive model found no evidence supporting the decrease in the number of cases in warmer weather. Moreover, an SEIR model calibrated for 202 locations in 8 countries[50] showed no significant changes in the number of COVID-19 confirmed cases with a broad range of meteorological conditions. Another study in New South Wales, Australia[51], revealed a weak correlation between COVID-19 cases and temperature, but a negative correlation between cases and relative humidity. Studies using data for the earlier infections in Jakarta with average temperature (26.1–28.6 °C)[52] and Bangladesh with average temperature (23.6–31.1 °C) and minimum temperature (17.3–29.3 °C)[45] indicated significantly positive correlation. In addition, COVID-19 cases in China showed negative correlations with both temperature and relative humidity as investigated in[53] while those in 190 countries revealed non-linear correlations with both daily temperature and relative humidity as in[54]. In Iran, also according to[55], there was no clear evidence to relate the number of confirmed cases with warm or cold weather in different provinces, leaving population size to be a determinant factor. A related study for India was carried out using minimum temperature, maximum temperature, average temperature, and specific humidity (ratio of the mass of water vapor to the total mass of the air parcel) as the weather components[56]. The results showed a high positive correlation between COVID-19 cases and temperature measures and a low positive correlation between COVID-19 cases and specific humidity. In Germany, the confirmed cases hit 17 million by the first week of January 2021. The second wave escalation began in autumn and continued in winter. Daily cases exceed 20,000 in many days at the latter stage, where it was over 15,000 for other days in the last two months of 2020. The long-standing plateau of total deaths has also altered since November to a sharp increase and reached 35,000 at the beginning of 2021.

Motivated by the increase of morbidity during autumn and winter, this study employed panel COVID-19 incidence data from Germany and scrutinized their relationship with weather data. In some studies, weather components like temperature were collected in categories such as average, maximum, and minimum level[52,56–58], while others used daily average extracted on a defined regular interval[50,59]. Furthermore, in some other studies, either absolute humidity[59,60] or specific humidity[56] was employed instead of relative humidity. Ours utilized the average of daily average temperature and relative humidity from January 31, 2020 to December 15, 2020, from three representative weather stations in Germany. Besides data availability and similarity with other studies[61,62], the rationale behind the choice of the weather components lies in their readability throughout academia and the fact that no prior and posterior transformation are needed to obtain marginal effects. Extensive investigation on Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ statistics then followed so as to cover spatial auto-correlation and related practical implications. The difference with previous studies is that the temporal progression of the statistics is presented. Subsequently, this study brought forward a random-effects model with a clustering strategy. Our holistic idea lies in which ranges of the incidence are well predicted by the weather components. This is somewhat contrasting to determining the ranges of the weather components that can predict the incidence. Our clustering is based on the method of stratifying incidence data into an arbitrary number of clusters, separated by barriers. The temperature and relative humidity data were also grouped corresponding to the clustered incidence data. This not only improves fitting by providing more explanatory variables but also screens incidence clusters where the weather components fail to predict. Relevant implications using marginal effects for sample cases then followed posterior to model diagnostics.

Data and methods

COVID-19 and weather situation in Germany. According to the official 2018 census, the German states considerably vary in population, with North Rhine-Westphalia and Bremen having the highest and lowest population size of about 17,932,651 and 682,986, respectively, out of the total population size of 83,019,213. The states also have varied economic capacities in business, industries, tourism, and education, which affect their population size. For instance, the largely populated states like Bavaria and Baden-Württemberg have booming economy and offer plenty of employment opportunities due to the situation of renowned business centers and industries, whereas low-populated states e.g. Bremen are laid behind (see[64,65]). Apparently, the number of cases and fatalities relatively depends on the population size. For instance, based on the report from Robert Koch Institute (RKI) on December 16, 2020, the largest populated state shared the highest 7-day incidence cases, and the smallest populated state shared the lowest. Given that the cases are population-driven, the dataset used for this study includes the daily confirmed COVID-19 cases for all the states from the official website of RKI[66], which was later normal-



State	B-W	Bav	Ber	Bra	Bre	Ham	Hes	M-V
Min	0	0	0	0	0	0	0	0
Max	38.05	40.95	53.77	40.53	44.22	39.16	49.75	18.20
Mean	5.96	6.89	7.55	3.80	5.93	5.57	6.04	1.79
StDev	8.43	9.77	11.68	6.91	9.37	8.01	9.44	3.34
Population	11,069,533	13,076,721	3,644,826	2,511,917	682,986	1,841,179	6,265,809	1,609,675
State	LS	NRW	RLP	Saa	Sax	S-A	S-H	Thu
Min	0	0	0	0	0	0	0	0
Max	26.47	39.22	36.77	54.01	86.37	29.34	18.26	48.25
Mean	3.73	6.19	4.85	5.67	7.34	2.97	2.19	94.34
StDev	5.46	8.79	7.72	9.73	15.06	5.65	3.12	8.45
Population	7,982,448	17,932,651	4,084,844	990,509	4,077,937	2,208,321	2,896,712	2,143,145

Figure 1. Daily COVID-19 cases per 100,000 inhabitants from all 16 states in Germany from March 01 until December 15, 2020: B-W (Baden-Württemberg), Bav (Bavaria), Ber (Berlin), Bra (Brandenburg), Bre (Bremen), Ham (Hamburg), Hes (Hesse), M-V (Mecklenburg-Vorpommern), LS (Lower Saxony), NRW (North Rhine-Westphalia), RLP (Rhineland-Palatinate), Saa (Saarland), Sax (Saxony), S-A (Saxony-Anhalt), S-H (Schleswig-Holstein), Thu (Thuringia). Population data come from the 2018 census by the Federal Statistical Office of Germany[63].

ized per 100,000 inhabitants using the 2018 population census, see Fig. 1. This dataset spans the time window from March 01, 2020 to December 15, 2020. The normalization was intentional toward making the number of cases comparable across the states so as to allow for appropriate comparison with weather components that do not depend on the population (see similar treatments in[59,67,68]). Here, the daily cases were defined as the difference of the confirmed cases since the earliest time of the report. As for the accompanying weather components, temperature and relative humidity data were retrieved from climate environment open data[69]. Time series of average temperature and relative humidity were obtained using the records of three weather stations Berlin-Marzahn (Berlin), München-Stadt (Bavaria) and Stuttgart-Schnarrenberg (Baden-Württemberg). This choice was motivated by data availability and the fact that the weather pattern throughout Germany is more or less the same, except in the alps where a negligible percentage of humans live. Average temperature ranges from − 0.766 to 27.13, and average relative humidity ranges from 39.38 to 93.53%. It seems the two weather components have a negative correlation showing equivalence between low temperature and high relative humidity or vice versa. Moreover, looking at the plot of cases by month in Fig. 1 in comparison with the weather components in Fig. 2, it can be seen that cases are generally higher in colder season and considerably reduce during the hot season.

In addition to the reported incidence, the spatial movement of the largest outbreak over the 16 states is also worth investigating. As depicted in Fig. 3, several stages in the timeline can be identified according to the dominance shown by different states. In the first three weeks in March, the largest incidence mainly altered between Hamburg and Baden-Württemberg. Bavaria and Saarland replaced them in the next three weeks. Bavaria hold a local election on March 15, and in the next day, a state of emergency was declared for 14 days with mobility restrictions[70]. Moreover, it is the first state to declare curfew that was imposed on March 20[71]. Saarland neighboring with badly affected French region Grand Est also incurred the same situation at midnight on the same day[72]. Lack of protective clothing and closure of medical practices were also reported from Bavaria[73]. Thus, Bavaria owed the largest incidence from time to time, even after the first few weeks. Outbreaks in initial reception facilities also contributed to the increase of cases in Bavaria. The largest incidence in May and in the first two weeks of June was dominated by Bremen. It was followed by Berlin and North Rhine-Westphalia until the end

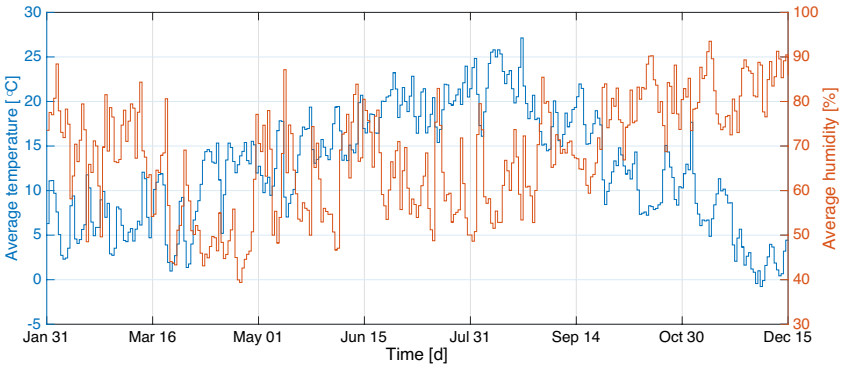


Figure 2. Average from the daily average temperature and relative humidity from the three weather stations in Germany: Berlin-Marzahn (Berlin), München-Stadt (Bavaria), Stuttgart-Schnarrenberg (Baden-Württemberg). Time window spans from January 31 until December 15, 2020. The tuples (Min, Max, StDev) are given by (– 0.766 °C, 27.13 °C, 6.45 °C) for the temperature and (39.38%, 93.53%, 12.71%) for the relative humidity, respectively.

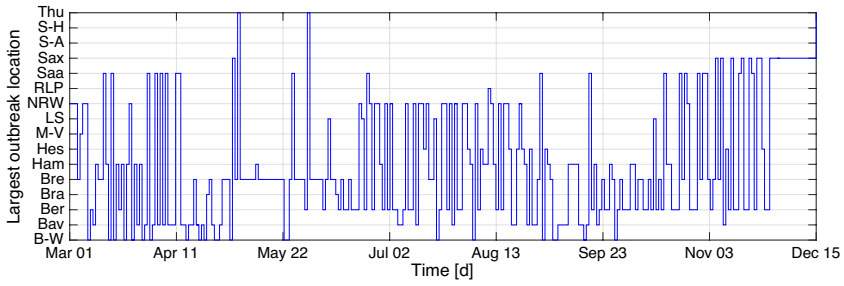


Figure 3. Spatial concurrence of the largest outbreak.

of August. A sudden increase of cases was reported in North Rhine-Westphalia due to proactive case tracing, in particular at a meat factory in Coesfeld[74]. Later another cluster occurred on June 17 in a slaughterhouse in Gütersloh, North Rhine-Westphalia, leaving superspreading the main cause of spread[75]. Hamburg and Bremen also came to notice in September and October. The latter stage of October was dominated by Saarland and Berlin. In November, the largest incidence altered between Saxony and Berlin, while Saxony kept the dominance for the first two weeks of December. Saxony had shown early signs of vulnerability, prohibiting residents from leaving their dwellings similar to Bavaria and Saarland. Berlin prevailed as the most responsible state in the latter two-third of the timeline. A large-scale protest was held on August 1 in Berlin against preventive measures. This hints lack of compliance to wearing face masks and keeping physical distance that supports increasing incidence[76].

Correlation studies. Referred studies in "Introduction" illustrate how meteorological factors correlated with the transmission of COVID-19. Highly transmissible disease like COVID-19 requires pathogens to remain active outside of the host body and relative humidity and temperature affect the virus's survival in the environment[44,77]. Another study engineering a SARS-CoV-2 isolate came across the fact that the virus can survive at least 28 days at ambient temperature 20 °C and 50% relative humidity on non-porous surfaces and is sensible to the variation of the weather components[78]. Therefore, it is considered noteworthy to examine the interrelationship between COVID-19 cases and meteorological factors. Many statistical methods have been used in earlier studies. According to the recent review in[61], applicable methods other than descriptive analysis are Pearson correlation coefficient, linear, and non-linear regression, LOESS, two-way ANOVA, etc. Wavelet coherency analysis was used in[50]. This study used the Spearman-rank correlation so as to evaluate both the linear and monotonic relationship between two tested covariates. Additionally, auto-correlation between reported COVID-19 cases was also done by piling the spatiotemporal data into one time series, considering that normalized data vary in relatively small numbers. Lags up to 7 days from presently were selected. Therefore, every

covariate augments 16 times 283 observations where the lag-0 time series consists of time window from March 8, 2020 to December 15, 2020. Both the Pearson and Spearman-rank correlation coefficients were computed.

Spatial pattern. Of special interest in this study is the degree of interconnection between all states in raising or decreasing the number of cases. The global Moran's $\mathscr{I}$[79] in comparison with the global Geary's $\mathscr{C}$[80,81] and its local decomposition known as Moran's scatter plot were used. The global measures serve to indicate the overall correlation between daily COVID-19 cases per 100,000 inhabitants in every state with the weighted average of the cases in neighboring states, which refers to the *spatial lag* of the state[82]. The spatial pattern is commonly seen to lie between three extreme cases: *locally clustered*, *random*, and *locally dispersed*. Locally clustered refers to the situation where neighboring states are similar in the level of daily new cases, under which spatial dependency rules out the spatial pattern. Locally dispersed refers to the inverse spatial dependency where neighboring states are dissimilar. Something in between is then referred to as random. Representation of these spatial patterns can be understood with the aid of a chessboard. If the spatial profile of daily cases in all states resembles the chessboard, then the spatial pattern is completely locally dispersed. If all the black cells would have gathered in one spot, then the spatial pattern is completely locally clustered. The random spatial pattern is then recognized from the way the black and white cells locate randomly on the board. This is extreme binary stratification that could never occur in the realism of epidemics, from which the corresponding global measure rarely reaches its bounds.

Let us suppose that time is fixed and the daily cases from all states are reported as $C = (c_1, \ldots, c_S)^\top$ with mean $\bar{c}$. The other main ingredient in spatial auto-correlation is the spatial weight matrix $W = (w_{ij})$, which measures the degree of contiguity among all the states. This study used the binary adjacency matrix, where w_{ij} is 1 in case i and j share a common border or 0 in case otherwise (including diagonal entries). This definition is commonly used in the literature (referred to as "queen case") in contrast to distance-based proximity measure where central locations play a significant role as well as a definition of being a "center" is required to define the distances. Let us write $Z = (z_1, \ldots, z_S)^\top := C - \bar{c}$ and define $|W| := \sum_{ij} w_{ij}$. The global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ statistic are given by

$$\mathscr{I} := \frac{S}{|W|} \cdot \frac{Z^\top W Z}{Z^\top Z} \quad \text{and} \quad \mathscr{C} := \frac{S-1}{2|W|} \cdot \frac{\sum_{ij} w_{ij}(c_i - c_j)^2}{Z^\top Z}$$

respectively. According to the formulas, the global Moran's $\mathscr{I}$ represents the standardized spatial autocovariance by the variance of the data, while the global Geary's $\mathscr{C}$ replaces the autocovariance by the sum of the squared differences in all data values. Both formulas then differ in sensitivity controlled by the autocovariance. In terms of stability against uncertainty in the data, Wijaya et al. in[68] describe how Geary's $\mathscr{C}$ tends to vary less significantly than Moran's $\mathscr{I}$ when data are perturbed using noise of any kind. The current study presented Geary's $\mathscr{C}$ only for the sake of comparison. A measurement $0 < \mathscr{I} \to 1$ (similarly $1 > \mathscr{C} \to 0$) indicates the direction toward locally structured spatial pattern; $\mathscr{I} = 0$ (or $\mathscr{C} = 1$) random spatial pattern; and $0 > \mathscr{I} \to -1$ (or $1 < \mathscr{C} \to 2$) locally dispersed spatial pattern. Statistical inference is usually done under a total randomization assumption to have a decision outcome based on the values of the statistics[83]. The p-value is generated after normalization using the expected values $\mathbb{E}(\mathscr{I}) = -1/(S-1)$, $\mathbb{E}(\mathscr{C}) = 1$ and variances $\mathbb{V}(\mathscr{I})$, $\mathbb{V}(\mathscr{C})$ reported in the original studies[79,80]. The null hypothesis is that there is no spatial auto-correlation of the daily cases on the observed S states, meaning that $\mathscr{I} \simeq \mathbb{E}(\mathscr{I})$ and $\mathscr{C} \simeq \mathbb{E}(\mathscr{C})$. Therefore, a p-value smaller than a predefined significance level α rejects the null hypothesis whereby either a locally structured or a locally dispersed spatial pattern occurs.

In contrast to the global measures, Moran's scatter plot measures the extent to which a state is considered a "hot spot" or "cold spot" or something in between[83]. It reports the coordinates $(Z/\sigma_C, WZ/\sigma_C)$ for all states, with $\sigma_C = \sqrt{Z^\top Z/S}$ denoting the standard deviation of C. As a row-standardized weight matrix is utilized, i.e., $|W| = S$, the pooled estimator of the regressing linear line for these coordinates passing through the origin is given by $(0, \mathscr{I})$. In the present context, a hot spot is defined as a state with a large number of daily cases surrounded by those with large numbers of cases (*high-high*). In the 2-dimensional Euclidean space, the coordinates of hot spots locate in the upper-right quadrant Q1. A cold spot, on the contrary, defines a state with a small number of cases surrounded by those with small numbers of cases (*low-low*). The coordinates of cold spots gather in the lower-left quadrant Q3. Other than these, local dispersion may occur falling into the following categories: a state with a small number of cases surrounded by those with large numbers (*low-high*) in the upper-left quadrant Q2, and a state with a large number of cases surrounded by those with small numbers (*high-low*) in the lower-right quadrant Q4. From the practical point of view, being a hot spot or cold spot may only rely on the health care capacity to ameliorate the disease burdens without imposing further restrictions to travel around neighboring states, except for those who travel across the border between scattered hot spots and cold spots. A state in a high-low or low-high spatial pattern, however, requires more restriction in traveling to neighboring states as the disease may diffuse (in case of high-low) or be absorbed (in case of low-high).

Simple case–weather relation. Let i and j denote the state and time index where $i \in \{1, \ldots, S = 16\}$ and $j \in \{1, \ldots, N\}$. Our approach to modeling daily COVID-19 cases in all states in Germany was based on directly relating collected entities. These include presently (lag-0) reported cases $C := (c_{ij})$, cases reported on the past seven days (lag-1, $\cdots$, lag-7) from presently $C_{-1} := (c_{ij-1}), \ldots, C_{-7} := (c_{ij-7})$, average air temperature $T := \mathbb{1}_S \otimes (t_j)$, and lag average relative humidity $H := \mathbb{1}_S \otimes (h_{j-25})$ corresponding to the cross-correlation result in Fig. 4. The notations $\mathbb{1}_S$ and $\otimes$ denote the column vector of size S whose entries are 1 and the Kronecker product between two matrices, respectively. The final size of our observations is the entire time window length minus the maximal autoregressive lag, which is $N := 290 - 7 = 283$ (i.e. from March 8 until December 15, 2020). Let us denote β_0 as the intercept, $\beta_{\text{ind}} := (\beta_1, \ldots, \beta_{S-1})$ as the individual-specific effects (cut down by

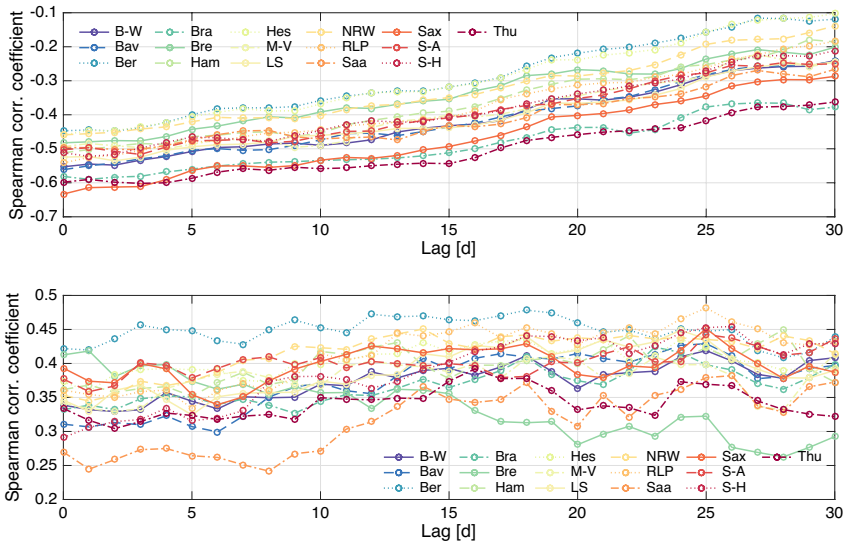


Figure 4. Spearman-rank correlation coefficients between daily cases from all states in Germany with the average temperature (above) and average humidity (below) on a moving window of 290 observations. Averaging throughout the states obtains the minimum of − 0.5223 (temperature) and maximum of 0.4194 (humidity) corresponding to the lags 0 and 25, respectively.

one term to avoid linear dependence with the intercept), β_{-i} (for $i = 1, \ldots, 7$) as the marginal effects of the lag incidence cases, β_T as the marginal effect of the temperature, β_H as the marginal effect of the relative humidity, and $\varepsilon = (\varepsilon_{ij})$ as the idiosyncratic error. The direct relationship among these covariates intends to not only skip additional transformations but also return direct marginal effects represented by the coefficients of the corresponding explanatory variables. This reads as

$$C = \beta_0 \mathbb{1}_{S\times N} + \sigma^{(0)} \mathbb{1}_N^\top \otimes [\beta_{\text{ind}}\, 0]^\top + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} C_{-i} + \sigma^{(8)} \beta_T T + \sigma^{(9)} \beta_H H + \varepsilon, \tag{1}$$

which folds

$$\begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{S1} & \cdots & c_{SN} \end{pmatrix} = \begin{pmatrix} \beta_0 & \cdots & \beta_0 \\ \vdots & \ddots & \vdots \\ \beta_0 & \cdots & \beta_0 \end{pmatrix} + \sigma^{(0)} \begin{pmatrix} \beta_1 & \cdots & \beta_1 \\ \vdots & \ddots & \vdots \\ \beta_{S-1} & \cdots & \beta_{S-1} \\ 0 & \cdots & 0 \end{pmatrix} + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} \begin{pmatrix} c_{1,1-i} & \cdots & c_{1,N-i} \\ \vdots & \ddots & \vdots \\ c_{S,1-i} & \cdots & c_{S,N-i} \end{pmatrix}$$
$$+ \sigma^{(8)} \beta_T \begin{pmatrix} t_1 & \cdots & t_N \\ \vdots & \ddots & \vdots \\ t_1 & \cdots & t_N \end{pmatrix} + \sigma^{(9)} \beta_H \begin{pmatrix} h_{-24} & \cdots & h_{N-25} \\ \vdots & \ddots & \vdots \\ h_{-24} & \cdots & h_{N-25} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1N} \\ \vdots & \ddots & \vdots \\ \varepsilon_{S1} & \cdots & \varepsilon_{SN} \end{pmatrix}.$$

The indicator parameters $\sigma^{(i)}$ take binary values and will serve to drop certain variables in the model specification (by value 0), whenever necessary. This model represents, perhaps, the simplest panel regression model in the following sense. The marginal effects of the lag incidence cases and those of the weather components could have been raised to matrices like in vector autoregression with exogenous variables (VAR-X) models[84]. Besides appending too many parameters (entries of the endogeneous matrices), which may lead to overfitting, VAR-X models also require all the explanatory variables to be covariance stationary (see[85] for details), which is rarely the case for disease and weather data in the subtropics. As the only random spatial pattern was observed from the incidence data for almost all observations, no essential state-crossing marginal effects were expected. State-dependent marginal effects for the weather components were also not considered due to data aggregation and limitation, also to the intention to have unified marginal effects that work on the national level. Moreover, all lags smaller than the optimal values for the weather components were not considered for complexity reduction. For the reason of having straight-forward marginal effects, prior transformations were not applied to any of the variables. Despite its simplicity, the model (1) treats omitted variable bias by including individual-specific effects. These are the simplest terms assuming that the omitted variables only have constant effects on the daily

COVID-19 cases in all the states. After all, the present study draws forth an outlook for compiling temperature and relative humidity data from all eligible stations as well as data of other confounding factors (e.g. other weather components, human mobility, employment opportunities, mapping of manufactures or public gatherings, etc) that not only add more explanatory variables but also clear up the heteroscedasticity issue.

Model including incidence clustering. Previous studies based their investigation on asking which ranges of weather components correctly predict incidence cases. This study asks a slightly different question: which ranges of incidence cases are correctly predicted by the existing values of the weather components. The values that fail to predict certain incidence cases due to insignificance would deem dropping. In[68], this clustering strategy was designed to eliminate the weather dependency on the zero incidence cases, handling the zero-inflation problem appropriately. In the context of COVID-19, some extreme cases might have never been related to weather, for example superspreading events[10–13] and indoor aerosol transmission[8,9]. The basic aim of the clustering is then to correctly place the role of weather where it should have never predicted such events. The use of a transient function to replace this functionality was inapplicable to us, for which bias may arise from the functional choice and its related extension strategy for prediction.

The clustering idea departs from stratifying the incidence data into M clusters $(\Omega_k)_{k=1}^{M}$ separated by barriers $\theta := (\theta_k)_{k=1}^{M-1}$. In the closed forms, the clusters are given by $\Omega_k = \{c : \max\{0, \theta_{k-1}\} \leq c < \min\{\theta_k, \max_{i,j} c_{ij}\}\}$. Let us define the function $\delta_k(C;\theta) := (\mathbb{1}_{\Omega_k} c_{ij})$, where $\mathbb{1}_{\Omega_k}$ denotes the characteristic function, taking value 1 in case c_{ij} belongs to Ω_k or 0 in case otherwise. Let us denote $P \circ Q = (p_{ij} q_{ij})$ as the Hadamard product between two matrices and define $T^k = T^k(\theta) := \delta_k(C;\theta) \circ T$, $H^k = H^k(\theta) := \delta_k(C;\theta) \circ H$. The latter return the original entries of the matrices T, H in case their pairing incidence cases belong to the corresponding cluster or 0 in case otherwise. Under this decomposition it always holds $\sum_k T^k = T$ and $\sum_k H^k = H$. Including clustering, a new model revises model (1) in the following fashion

$$C = \beta_0 \mathbb{1}_{S\times N} + \sigma^{(0)} \mathbb{1}_N^\top \otimes \beta_{\text{ind}}^\top + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} C_{-i} + \sum_{i=1}^{3} \sigma^{(7+i)} \beta_T^i T^{(i)} + \sum_{i=1}^{3} \sigma^{(10+i)} \beta_H^i H^{(i)} + \varepsilon. \tag{2}$$

Here, the incidence data were classified into three clusters ($M = 3$) on the basis of practicality to call for lower, middle, and upper cluster. In principle, the specification is not bound to such a small number as fitting would be better with more explanatory variables. However, questions regarding complexity and practical interpretations might arise when using a large number of clusters. On the present choice, when for instance $T^{(2)}$ has to be dropped due to insignificance, this simply means that the average temperature fails to predict incidence cases in the range defined by the middle cluster Ω_2. This model then allows the lone cases to be "unexplained by temperature".

The fact that T^k and H^k change with the lower and upper barrier $\theta = (\theta_\text{l}, \theta_\text{u})$, so does the pooled estimator $\hat{\beta} = \hat{\beta}(\theta)$ where $\beta = (\beta_0, \beta_{\text{ind}}, \beta_{-1}, \ldots, \beta_{-7}, \beta_T^1, \ldots, \beta_H^3)$. Our aim is to find the optimal barriers such that the squared error between data $C = (c_{ij})$ and the model approximate $C[\hat{\beta}](\theta)$ achieves its minimum. Mathematically, the preceding statement translates to the following problem

$$\min_{\theta} \qquad \sum_{i,j} (c_{ij}[\hat{\beta}](\theta) - c_{ij})^2 \tag{3a}$$

$$\text{subject to} \quad \min_{i,j} c_{ij} \leq \theta_\text{l} \leq \theta_\text{u} \leq \max_{i,j} c_{ij}. \tag{3b}$$

The pooled estimator $\hat{\beta}$ follows from the straightforward formula in terms of matrix inverse and multiplication involving explanatory and response variable.

Results

Case–weather cross-correlation and case-specific auto-correlation. Figure 4 represents the correlation coefficients on a moving window of 290 observations with time lags from 0 to 30 days for each state. Notice that the reported daily COVID-19 cases correlated negatively with the average temperature and positively with the average relative humidity. The magnitude of the correlation coefficient with average temperature shows decreasing trends with lag for all the states. With no lag introduced, the correlations are negative and significant for all the states (p-values from 6.27×10^{-34} to 1.17×10^{-15}). Averaging the correlation coefficients throughout the states, the minimum of − 0.5223 was obtained. This negative correlation is comparable up to certain ranges of minimum, maximum and average temperature to the studies in Brazil (with both average ranging from 20.9 to 27 °C and maximum temperature from 23.1 to 34.2 °C in[57] and with average temperature ranging from 16.8 to 27.4 °C in[86]) as well as the data in 127 countries (with average temperature from − 17.8 to 42.9 °C in[87]). In New York[88], the correlation was positive and insignificant for average and minimum temperature but positive and insignificant for the maximum temperature. In Oslo, Norway[89], the correlation was negative and insignificant for all maximum, minimum, and average temperature with 14 days time lag, but positive and significant correlation was obtained for normal temperature with 0, 5, 6, and 14 days lag. The temperature in Oslo ranged from − 7.5 to 21.9 °C during the study period. COVID-19 cases in Russian Federation exhibited positive significant correlation with minimum (− 17.78 °C to 8.89 °C), maximum (0.56 °C to 27.2 °C) and average temperature (− 2.78 °C to 16.1 °C)[46].

As far as relative humidity is concerned, it can be observed from Fig. 2 that its average varies from 39.38 to 93.53%. The best lag was found 25 days with the correlation coefficient value of 0.4194 from averaging throughout

ρ	lag-0	lag-1	lag-2	lag-3	lag-4	lag-5	lag-6	lag-7
lag-0	1							
lag-1	0.87, 0.83	1						
lag-2	0.83, 0.81	0.87, 0.83	1					
lag-3	0.80, 0.79	0.83, 0.81	0.87, 0.83	1				
lag-4	0.79, 0.79	0.81, 0.79	0.83, 0.79	0.87, 0.83	1			
lag-5	0.82, 0.80	0.80, 0.79	0.81, 0.79	0.83, 0.80	0.87, 0.83	1		
lag-6	0.87, 0.82	0.83, 0.80	0.80, 0.79	0.80, 0.79	0.83, 0.80	0.87, 0.83	1	
lag-7	0.89, 0.83	0.87, 0.82	0.83, 0.79	0.79, 0.78	0.80, 0.79	0.83, 0.80	0.87, 0.83	1

Table 1. Pearson and Spearman-rank correlation coefficients from the incidence data, rounded to two digits after comma.

the states. With this lag, the correlations are positive and significant for all states (p-values from 2.98×10^{-18} to 1.92×10^{-8}). For the relative humidity, different results preceded ours. A previous study in New York[88] concluded that average relative humidity was insignificantly negatively correlated with the daily new cases. It was found that average humidity was significantly negatively correlated and relative humidity was insignificantly negatively correlated with the number of the ICU daily patients, according to data from Milan (14–100% for relative humidity, 1–23 g m^{-3} for average humidity), Florence (10% to 100% for relative humidity, 1 to 23 g m^{-3} for average humidity) and Trento (16–100% for relative humidity, 1 to 25 g m^{-3} for average humidity) in Italy[90]. Data from Brazil ranging from 69.5 to 90.8% with no lag[50,57] showed that the correlation was positive but not significant with minimum and maximum average humidity. Data from 127 countries[87] led to the conclusion that the relative humidity was correlated negatively and insignificantly with daily new cases.

Table 1 shows the case-specific auto-correlations. Generally, Pearson is higher than Spearman-rank correlation coefficient. In addition, both Pearson and Spearman-rank correlation coefficient are significant with minimum 0.78 (p-values $\simeq 0$). From the column of lag-0, the auto-correlation generally swings from a large value at lag-1, then minima at either lag-3 or lag-4, to another large value at lag-7. The same behavior can be observed from the columns lag-1 until lag-3 where decrement rules out the first 4 lags and minima were found at either lag 3 or 4 days from the time series. This finding will set a basis for those in the panel regression models, as seen shortly.

Spatial auto-correlation. Meanwhile previous studies much focused on aggregated data and variation of distances in the spatial weight matrix, this study computed the global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ for all time to see how the spatial pattern changes seasonally since the earliest infection. The corresponding computation results together with the 95% confidence interval $[\mathscr{I} - 1.93\sqrt{\mathbb{V}(\mathscr{I})}, I + 1.93\sqrt{\mathbb{V}(\mathscr{I})}]$ (respectively for $\mathscr{C}$) are presented in Fig. 5. Although the spatial pattern of the daily cases in all the states changes around with time, it is evident that randomness overwhelms the pattern for most of the time. The progression of p-values (especially below α) indicates that, generally, no significant difference between Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ was observed except on the duration from November until mid of December where Geary's $\mathscr{C}$ shows more locally clustered spatial pattern.

The Moran's scatter plot for all the states in Germany was determined for all observations, see Fig. 6. For the sake of serial presentation, indexing the coordinates based on the quadrants is more favorable than plotting them. Overall, the results suggest that all the states show randomness with time in to which spatial pattern they belong. If one solely focuses on the recent observations (November 1 to December 15, 2020), then the following states have the tendency to occupy the following quadrants: Baden-Württemberg, Bavaria, Hesse, Thuringia (Q1); Brandenburg, Rhineland-Palatinate, Saxony-Anhalt (Q2); Hamburg, Mecklenburg-Vorpommern, Lower Saxony, Schleswig-Holstein (Q3); Berlin, Bremen, North Rhine-Westphalia, Saxony (Q4).

Panel regression models. Variable choices for model specification were investigated. The criteria are based on not only fit and complexity (information-type criterion) but also insignificance, negative marginal effects, and multicollinearity driven by certain variables. For the fit and complexity, a minimal value of Bayesian Information Criterion $\mathrm{BIC} = -2\log(L) + \log(N) \cdot k$[91] was sought. The first term of this criterion expresses maximization over the likelihood function L generated from our model and the second term includes the observation size N as well as the number of parameters k. Unlike Akaike Information Criterion (AIC)[92] that would have replaced $\log(N)$ by 2, BIC penalizes the number of parameters much more, especially for large observation sizes. Our study aims to drop certain variables toward cutting down BIC and amending insignificance as well as multicollinearity. The standard t-test was used for the significance test. Checking for multicollinearity follows from computing the Inverse Variance Inflation Factor (1/VIF) values for all explanatory variables except the constant. A 1/VIF measures one minus the coefficient of determination derived from an OLS-regression whereby the variable under test serves as the response while the others as the explanatory variables. In this sense, 1/VIF of a value smaller than the rule of thumb 0.1 shows multicollinearity driven by the tested variable[93]. In addition, the p-value of the F-statistic is monitored, which measures if the overall variables are simultaneously significant; of which smaller than $\alpha = 0.05$ indicates that they are. Not only can the model be designated to be better than just a constant, but multicollinearity can also be diagnosed. Johnston in[94] hinted the existence of multicollinearity as some p-values from t-tests are large while that from F-test is radically small, which agrees to the analytical inves-

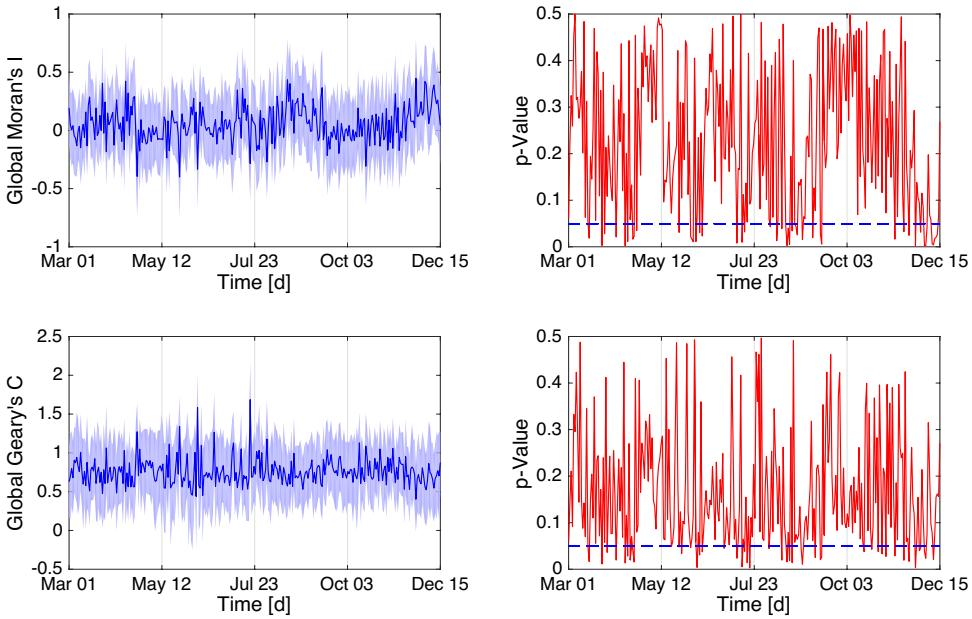
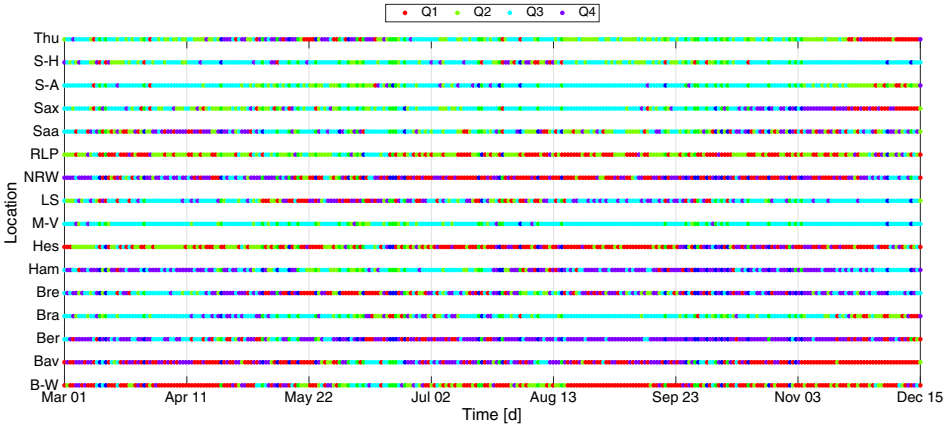


Figure 5. Global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ computed on a daily basis together with the corresponding 95% confidence interval and p-Value (right) for significance. The blue dashed line represents the significance level $\alpha = 0.05$.



State	B-W	Bav	Ber	Bra	Bre	Ham	Hes	M-V	LS	NRW	RLP	Saa	Sax	S-A	S-H	Thu
Q1	62.22%	88.88%	24.44%	22.22%	0%	0%	55.55%	0%	0%	28.88%	28.88%	17.77%	42.22%	11.11%	0%	46.66%
Q2	24.44%	4.44%	4.44%	35.55%	6.66%	0%	11.11%	0%	2.22%	4.44%	51.11%	22.22%	2.22%	57.77%	0%	35.55%
Q3	6.66%	0%	8.88%	35.55%	40%	53.33%	11.11%	100%	91.11%	24.44%	8.88%	26.66%	2.22%	28.88%	100%	13.33%
Q4	6.66%	6.66%	62.22%	6.66%	53.33%	46.66%	22.22%	0%	6.66%	42.22%	11.11%	33.33%	53.33%	2.22%	0%	4.44%

Figure 6. Classification into four quadrants (Q1, Q2, Q3, Q4) equivalent to Moran's scatter plot and the concurrence percentages from November 1 to December 15, 2020.

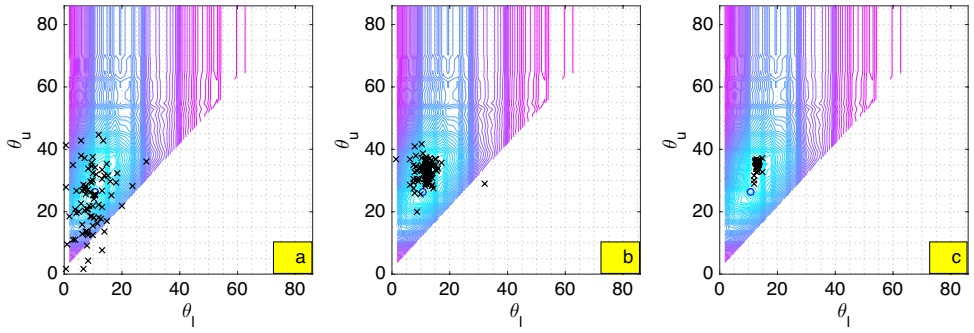


Figure 7. Computation of optimal barriers $(\theta_l, \theta_u) \approx (13.3645, 36.0597)$ for the clustering. Blue circle encodes the optimal barriers found by the brute-force computations on the 50×50 grid. The figures show the evolution of the locations of 100 players (black $\times$) converging to an optimal solution that does not overlap with the grid: (**a**) 5th iteration, (**b**) 10th iteration, (**c**) 20th iteration.

Model (1)							
$\sigma^{(i)} = 0$ for i		0	0, 3	0, 4	0, 3, 4		
BIC	24,049.74	23,933.03	23,924.93	23,953.11	23,946.74		
Issue	S0, S3	S3, N4	N4	S3, N3			
Model (2)							
$\sigma^{(i)} = 0$ for i		0, 12	0, 12, 10	0, 12, 3	0, 12, 10, 3	0, 12, 10, 4	0, 12, 10, 3, 4
BIC	21,006.56	21,578	21,569.72	21,570.42	21,562.14	21,587.59	21,580.1
Issue	S0, S3, S10, M12	S3, S10	S3	S10	N4	N3, S3	

Table 2. Model specification under variable dropping. BIC values as well as corresponding issues leading to model exclusion are reported: S*i*, N*i*, M*i* stand for insignificance, negative marginal effect, and multicollinearity driven by the corresponding variable ordered by $\sigma^{(i)}$, respectively.

tigation in[95,96]. Besides these aspects, if certain marginal effects would be consistent with our auto-correlation study were also checked. From Table 1, it is seen how cases in the past 7 days positively predict present cases with the least auto-correlations found from cases from the past 3 and 4 days. This led to dropping negative marginal effects corresponding to lag incidence cases that may occur due to a certain model specification.

To deal with the model including incidence clustering (2), the computation of optimal lower and upper barrier (θ_l, θ_u) as in (3) is necessary. The characteristic functions embedded in the objective function make the optimization problem non-smooth. The brute-force computations of the objective function in the upper-left triangle of the 50×50 grid in the domain $[\min_{i,j} c_{ij}, \max_{i,j} c_{ij}]^2$ and a PSO algorithm[68] were put in comparison. From Fig. 7, PSO outperforms the brute-force computations in locating the optimal barriers that minimize the objective function, also in terms of computation time.

According to Table 2, the BIC value for the simple model (1) is relatively large, exacerbated by large degrees of freedom. The model including incidence clustering (2) gives the least BIC value due to a minimal likelihood function. Additionally, the insignificance of the entire individual-specific effects for both models was spotted. The rationale behind this can be connected to the fact that the entire profile of global and local spatial auto-correlation as well as the largest outbreak ("COVID-19 and weather situation in Germany" and "Spatial pattern") show randomness for almost all observations. Therefore, no state was worth constant recruitment (weighting) for its neighborhood to show a consistent spatial pattern throughout the observations.

Post-estimation diagnostics for all the models including those investigated during model specification were performed. Additional to the models including lag incidence cases and weather components, this study considered the models where either of these entities is present. The fitting results are presented in Table 3. For straightforward marginal effects and computation of optimal barries, the pooled estimator was considered subject to its inefficiency. The test was conducted via the comparison between fixed-effects and random-effects estimator and that between random-effects and pooled estimator. To the former, the two estimators were compared using Durbin–Wu–Hausman test[97,98], where the fixed-effects estimator is assumed to be consistent, and the random-effects estimator is efficient and assumed to follow a normal distribution. The null hypothesis suggests that the random-effects estimator is a consistent estimator regardless of the size of the data. According to Table 3, the p-value corresponding to the statistic greater than $\alpha = 0.05$ indicates that the random-effects estimator is equally

	Val	StDev	*t* p-Val	1/VIF	*F* p-Val	R^2	Adj R^2	D-W-H	Wo	B-P LM
Model (1)										
β_0	−.8742	.3787	.021		0	.8558	.8556	.5355	0	1
β_{-1}	0.1827	0.0142	0	0.1603						
β_{-2}	0.0984	0.0128	0	0.2011						
β_{-5}	0.0514	0.0135	0	0.2033						
β_{-6}	0.2736	0.0149	0	0.1716						
β_{-7}	0.4145	0.0155	0	0.1645						
β_T	−0.0295	0.0099	0.003	0.6390						
β_H	0.0246	0.0049	0	0.7054						
β_0	0.1949	0.0593	0.001		0	0.8544	0.8543	0.5556	0	1
β_{-1}	0.1918	0.0142	0	0.1619						
β_{-2}	0.1054	0.0128	0	0.2026						
β_{-5}	0.0604	0.0135	0	0.2056						
β_{-6}	0.2791	0.0150	0	0.1723						
β_{-7}	0.4166	0.0155	0	0.1646						
β_0	−2.0767	0.7908	0.009		0	0.3694	0.3691	1	0.0094	0
β_T	−0.5681	0.0185	0	0.7997						
β_H	0.2256	0.0097	0	0.7997						
Model (2)										
β_0	5.9089	0.2162	0		0	0.9148	0.9146	0.7646	0	1
β_{-1}	0.1378	0.0109	0	0.1590						
β_{-2}	0.0716	0.0098	0	0.1998						
β_{-5}	0.0337	0.0104	0	0.2031						
β_{-6}	0.1636	0.0117	0	0.1667						
β_{-7}	0.2866	0.0123	0	0.1543						
β_T^l	−0.1261	0.0076	0	0.4755						
β_T^m	0.3158	0.0224	0	0.4380						
β_H^l	−0.0528	0.0026	0	0.3687						
β_H^u	0.2033	0.0047	0	0.6981						
β_0	1.8381	0.3594	0		0	0.8682 (within) 0.9558 (between) 0.8692 (overall)		0	0.0097	0
β_T^l	−0.2088	0.0092	0	0.4927						
β_T^2	−0.1010	0.0292	0.001	0.3878						
β_T^3	−0.6897	0.1037	0	0.4472						
β_H^l	0.0524	0.0046	0	0.1785						
β_H^2	0.2627	0.0051	0	0.1243						
β_H^3	0.5608	0.0085	0	0.3258						

Table 3. Fitting results and diagnostics for the models (1) and (2). The abbreviations stand for the following: Val (value), StDev (standard deviation), *t* p-Val (p-value of the *t*-test for the variable significance), 1/VIF (Inverse Variance Inflation Factor for multicollinearity), *F* p-Val (p-value of the *F*-test for the overall variable significance), R^2 (coefficient of determination), Adj R^2 (adjusted coefficient of determination), D–W–H (p-value of Durbin–Wu–Hausman test for random-effects vs. fixed-effects estimator), Wo (p-value of Wooldridge test for the serial correlation), B–P LM (p-value of Breusch–Pagan test for random effect vs pooled estimator).

consistent as the fixed-effects estimator. The two estimators for all presented models confirm equivalence except for model (2) where only weather components are present. For this case, the fixed-effects estimator was kept to handle consistency and panel effect. To the latter, Breusch–Pagan Lagrange Multiplier test was done under no panel effect as the null hypothesis[99], i.e., the model under the random-effects estimator returns zero variance in the state-dependent errors. Apparently, no panel effect was observed for all models except for those that include only weather components, in which case either random-effects or fixed-effects estimator is preferable. The inefficiency of the presented pooled, random-effects, and fixed-effects estimator is confirmed as serial correlation in all the state-dependent errors occurred. Wooldridge test[100] showed this. Therefore, a caveat remains for all models that their standard deviations of the coefficients are smaller and R^2's are larger than they should be. After all, the pooled estimator is always consistent, even for a relatively small data size. As final practical remarks from the models, all the lag incidence cases give the waving effects in terms of lag where the cases 5 days and

7 days from presently predict the present cases the least and the most, respectively. Keeping the lag incidence cases, the weather components from model (1) give a consistent prediction with that from the cross-correlation study. Together with clustering, the marginal effects of weather were corrected for model (2). It was observed that temperature fails to predict cases in the upper cluster while relative humidity fails to cases in the middle cluster. Temperature seems to give a larger positive marginal effect for the middle cluster while relative humidity a negative smaller marginal effect for the lower cluster.

As far as predictive performance is concerned, several findings can be highlighted. As the larger models exhibit no more issues with insignificance and multicollinearity, neither do the smaller models. For the model variant (1), the smaller models gain $R^2 \approx 0.8544$, BIC $\approx$ 23,972.15 (only lag incidence cases) and $R^2 \approx 0.3694$, BIC $\approx$ 30585.35 (only weather components), respectively. Meanwhile the model including only weather components shows the poorest performance; its BIC value is also radically larger than that of the model including only lag incidence cases. For the model (1), the impact of weather is rather small, as the decrease of temperature from a reference value e.g. $T \approx 20\,^{\circ}\mathrm{C}$ to $T \approx 10\,^{\circ}\mathrm{C}$ (i.e. by 50%) is associated to the increase of COVID-19 cases for all states from e.g. $C \approx 20$ by $(|\beta_T|10/20) \cdot 100\% \approx 1.475\%$. When the lag incidence cases were dropped, the increase changes to $(0.5681 \cdot 10/20) \cdot 100\% \approx 28\%$. Moreover, the increase of relative humidity from 60 to 80% (by 33%) is associated to the increase of the cases from $C \approx 20$ by 2.46% (with lag incidence cases) and 22.56% (without lag incidence cases). The overall impression indicates the superiority of the model with only lag incidence cases when one designates fit to significantly matter than the number of parameters. For the model including incidence clustering (2), a different profile was obtained when only using non-dropped weather components: $R^2 \approx 0.7948$, BIC $\approx$ 25517.61. Here, a significant improvement under incidence clustering becomes evident. Surprisingly, the model including the entire weather components even outperforms that including only lag incidence cases by fit and complexity: $R^2 \approx 0.8692$, BIC $\approx$ 23494.94. All marginal effects corresponding to the temperature matrices are negative, and those corresponding to the relative humidity matrices are positive. It was observed that the temperature returns the smallest marginal effect on the COVID-19 cases in the middle cluster and relative humidity in the lower cluster. Besides the significance of the marginal effects, even no multicollinearity was observed. Apart from this, when the predictive ability is evaluated by R^2 and BIC amending multicollinearity and inconsistent predictors, it is still argued that combining lag incidence cases and weather components serve as the best models as presented in Table 3. The corresponding graphical fitting can be seen in Fig. 8.

Discussion

In this study, lags from the cross-correlation between average temperature and relative humidity were extracted to synthesize suitable variables in the regression models. Additionally, case-specific auto-correlation supports the model specification where lag-3 and lag-4 incidence would rather be insignificant predictors for the present incidence. Spatial auto-correlation using global Moran's $\mathcal{I}$ and global Geary's $\mathcal{C}$ was investigated in the framework of analyzing the spatial effect in COVID-19 transmission. The global measures indicate random spatial patterns most of the time, except there were either local clusters or dispersion in recent observations from November 1 to December 15, 2020. Moran's scatter plot was then used to disclose the local behavior of the spatial pattern. The result shows that the distribution of the hot spots and cold spots generally changed with time. The random spatial pattern justifies the model specification where the individual- or state-specific effects that would have served to endow specific states with constant weighting factors, were dropped.

In the simple random-effects model, the average temperature and lag relative humidity were shown to affect the incidence significantly, however, the resulting coefficient of determination is comparably much smaller than whenever only lag incidence cases were used; panel effect also raises in the former case. For the reason of placing the correct role of weather in predicting certain ranges of incidence, the weather components were grouped with the aid of a clustering strategy. The new clustering-integrated model accompanied by optimal barriers shows good agreement with the data whereby weather components outperform lag incidence cases in the prediction. On this matter, the fixed-effects estimator was the only presumably consistent estimator that also tackles the panel effect. For all models, it was observed that every explanatory variable competes against the others to be a significant predictor. Therefore, model choice together with its consequences (marginal effects), depend entirely on the decision-maker. Marginal effects can be guidance when a model is chosen a priori. When R^2 and BIC matter a lot, our recommendation is to opt for the clustering-integrated model with lag incidence cases and lag weather components. There it was found that temperature and relative humidity have negative, relatively small marginal effects on the cases in the lower cluster (below 13 cases per 100,000 inhabitants); the temperature has a large positive marginal effect on the cases in the middle cluster (between 13 and 36 cases per 100,000 inhabitants) and no marginal effect on the upper cluster (above 36 cases per 100,000 inhabitants); relative humidity has a large positive marginal effect on the upper cluster but none on the middle cluster. The clustering-integrated model with only weather components is recommended when weather receives more privilege than lag incidence cases. Our result is consistent with the cross-correlation study that temperature has negative marginal effects while relative humidity has positive marginal effects on the incidence in all clusters. The middle cluster receives the smallest marginal effect from temperature and the lower cluster from relative humidity. This hints physical consequences that temperature can only predict incidence cases during hot (summer) and cold season (winter), where cases clearly distinguish against each other from the data, not during transitional seasons (spring and fall). Relative humidity, on the other hand, is less likely to predict sinking cases during the hot season.

Conclusion

This study focused on the interrelationship between two weather components overlapping in many previous studies (average temperature and relative humidity) and COVID-19 incidence in Germany. Cross-correlation, case-specific auto-correlation, and spatial auto-correlation analysis were done to determine suitable variables

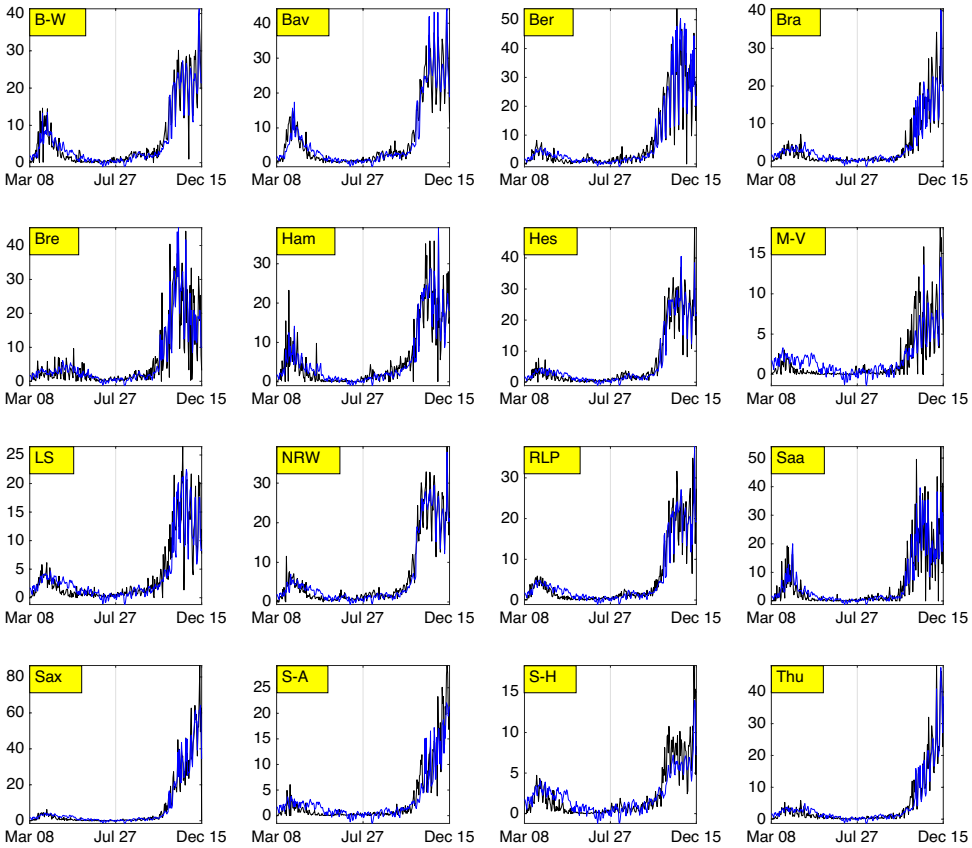


Figure 8. Fitting result (in blue) for the model including incidence clustering.

and to explain the negligible panel effect in the panel random-effects models. In addition, the findings from the spatial auto-correlation provide the placement of the 16 states in the four quadrants from Moran's scatter plot and appropriate policy regarding traveling restrictions. The increasing demand for confounding factors to explain various incidence levels has been neutralized by the aid of incidence clustering. This strategy supports the idea of considering only certain hypothetical factors predicting COVID-19 incidence and general regression modeling wherein explanatory variables are limited. This localization of incidence that is correctly predicted by the two weather components has profound implications for public health authorities. The modeling does not only determine the extent of the prediction via marginal effects but also paves the way for precautionary actions amidst upcoming weather.

Data availability

All the data sources have been included in "COVID-19 and weather situation in Germany".

Received: 18 January 2021; Accepted: 16 May 2021
Published online: 28 May 2021

References

1. Belser, J. A., Eckert, A. M., Tumpey, T. M. & Maines, T. R. Complexities in ferret influenza virus pathogenesis and transmission models. *Microbiol. Mol. Biol. Rev.* **80**, 733–744 (2016).
2. Storch, G. A. Diagnostic virology. *Clin. Infect. Dis.* **31**, 739–751 (2000).

3. Steinmeyer, S. H., Wilke, C. O. & Pepin, K. M. Methods of modelling viral disease dynamics across the within- and between-host scales: The impact of virus dose on host population immunity. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 1931–1941 (2010).
4. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487 (2008).
5. World Health Organization. Coronavirus disease (COVID-19): How is it transmitted? https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted (2020). Accessed 19 December 2020.
6. Azuma, K. *et al.* Environmental factors involved in SARS-CoV-2 transmission: Effect and role of indoor environmental quality in the strategy for COVID-19 infection control. *Environ. Health Prev. Med.* **25**, 1–16 (2020).
7. Wijaya, K. P. *et al.* An epidemic model integrating direct and fomite transmission as well as household structure applied to COVID-19. *J. Math. Ind.* **11**, 1–26 (2021).
8. Karia, R., Gupta, I., Khandait, H., Yadav, A. & Yadav, A. COVID-19 and its modes of transmission. *SN Compr. Clin. Med.*, 1–4 (2020).
9. Morawska, L. & Milton, D. It is time to address airborne transmission of Coronavirus Disease 2019 (COVID-19). *Clin. Infect. Dis.* **71**, 2311–2313 (2020).
10. Bouffanais, R. & Lim, S. Cities - try to predict superspreading hotspots for COVID-19. *Nature* **583**, 352–355 (2020).
11. Wong, F. & Collins, J. J. Evidence that coronavirus superspreading is fat-tailed. *Proc. Natl. Acad. Sci.* **117**, 29416–29418 (2020).
12. Kain, P. M., Childs, M. L., Becker, A. D. & Mordecai, E. A. Chopping the tail: How preventing superspreading can help to maintain COVID-19 control. *Epidemics* **34**, 100430 (2020).
13. Wang, L. *et al.* Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat. Commun.* **11**, 5006 (2020).
14. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
15. Ebrahim, S. H. & Memish, Z. A. COVID-19—The role of mass gatherings. *Travel Med. Infect. Dis.* **34**, 101617 (2020).
16. World Health Organization. WHO mass gathering COVID-19 risk assessment tool—Generic events. https://www.who.int/publications/i/item/10665-333185 (2020). Accessed 25 October 2020.
17. Assche, J. V., Politi, E., Dessel, P. V. & Phalet, K. To punish or to assist? Divergent reactions to ingroup and outgroup members disobeying social distancing. *Br. J. Soc. Psychol.* **59**, 594–606 (2020).
18. Belosi, F., Conte, M., Gianelle, V., Santachiara, G. & Contini, D. On the concentration of SARS-CoV-2 in outdoor air and the interaction with pre-existing atmospheric particles. *Environ. Res.* **193**, 110603 (2021).
19. Tung, N. T. *et al.* Particulate matter and SARS-CoV-2: A possible model of COVID-19 transmission. *Sci. Total Environ.* **750**, 141532 (2021).
20. Lei, H., Xu, X., Xiao, S., Wu, X. & Shu, Y. Household transmission of COVID-19—A systematic review and meta-analysis. *J. Infect.* **81**, 979–997 (2020).
21. Ooi, E. E. & Low, J. G. Asymptomatic SARS-CoV-2 infection. *Lancet Infect. Dis.* **20**, 996–998 (2020).
22. Lin, D. *et al.* Co-infections of SARS-CoV-2 with multiple common respiratory pathogens in infected patients. *Sci. China Life Sci.* **63**, 1–4 (2020).
23. Richardson, S. *et al.* Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* **323**, 2052–2059 (2020).
24. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of co-infection between SARS-CoV-2 and other respiratory pathogens. *JAMA* **323**, 2085–2086 (2020).
25. Boncristiani, H. F., Criado, M. F. & Arruda, E. Respiratory viruses. *Encycl. Microbiol.* **2009**, 500–518 (2009).
26. Dasaraju, P. V. & Liu, C. Infections of the respiratory system. in: *Medical Microbiology* 4th edn (ed Baron, S.) (University of Texas Medical Branch at Galveston, 1996).
27. Azekawa, S., Namkoong, H., Mitamura, K., Kawaoka, Y. & Saito, F. Co-infection with SARS-CoV-2 and influenza A virus. *IDCases* **20**, e00775 (2020).
28. Mossad, S. B. COVID-19 and flu: Dual threat, dual opportunity. *Clevel. Clin. J. Med.* **87**, 651–655 (2020).
29. Dowell, S. F. & Ho, M. S. Seasonality of infectious diseases and severe acute respiratory syndrome—What we don't know can hurt us. *Lancet Infect. Dis.* **4**, 704–708 (2004).
30. Shi, P. *et al.* Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Sci. Total Environ.* **728**, 138890 (2020).
31. Kronbichler, A. *et al.* Asymptomatic patients as a source of COVID-19 infections: A systematic review and meta-analysis. *Int. J. Infect. Dis.* **98**, 180–186 (2020).
32. Ozaras, R. *et al.* Influenza and COVID-19 coinfection: report of six cases and review of the literature. *J. Med. Virol.* **92**, 2657–2665 (2020).
33. Singh, B., Kaur, P., Reid, R. J., Shamoon, F. & Bikkina, M. COVID-19 and influenza co-infection: Report of three cases. *Cureus J. Med. Sci.* **12**, e9852 (2020).
34. Pormohammad, A. *et al.* Comparison of influenza type A and B with COVID-19: A global systematic review and meta-analysis on clinical, laboratory and radiographic findings. *Rev. Med. Virol.*, e2179 (2020).
35. Cai, Q. C. *et al.* Influence of meteorological factors and air pollution on the outbreak of severe acute respiratory syndrome. *Public Health* **121**, 258–265 (2007).
36. Chan, K. H. *et al.* The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* **2011**, 734690 (2011).
37. Casanova, L. M., Jeon, S., Rutala, W. A., Weber, D. J. & Sobsey, M. D. Effects of air temperature and relative humidity on coronavirus survival on surfaces. *Appl. Environ. Microbiol.* **76**, 2712–2717 (2010).
38. Sun, Z., Thilakavathy, K., Kumar, S. S., He, G. & Liu, S. V. Potential factors influencing repeated SARS outbreaks in China. *Int. J. Environ. Res. Public Health* **17**, 1633 (2020).
39. Gardner, E. G. *et al.* A case-crossover analysis of the impact of weather on primary cases of Middle East respiratory syndrome. *BMC Infect. Dis.* **19**, 1–10 (2019).
40. Altamimi, A. & Ahmed, A. E. Climate factors and incidence of middle east respiratory syndrome coronavirus. *J. Infect. Public Health* **13**, 704–708 (2020).
41. Cai, J. *et al.* Indirect virus transmission in cluster of COVID-19 cases. *Emerg. Infect. Dis.* **26**, 1343–1345 (2020).
42. Yeo, C., Kaushal, S. & Yeo, D. Enteric involvement of coronaviruses: Is faecal-oral transmission of SARS-CoV-2 possible?. *Lancet Gastroenterol. Hepatol.* **5**, 335–337 (2020).
43. Chin, A. W. H. *et al.* Stability of SARS-CoV-2 in different environmental conditions. *Lancet Microbe* **1**, e10 (2020).
44. Ahlawat, A., Wiedensohler, A. & Mishra, S. K. An overview on the role of relative humidity in airborne transmission of SARS-CoV-2 in indoor environments. *Aerosol Air Qual. Res.* **20**, 1856–1861 (2020).
45. Islam, A. R. T. *et al.* Effect of meteorological factors on COVID-19 cases in Bangladesh. *Environ. Dev. Sustain.*, 1–24 (2020).
46. Lasisi, T. T. & Eluwole, K. K. Is the weather-induced COVID-19 spread hypothesis a myth or reality? Evidence from the Russian federation. *Environ. Sci. Pollut. Res.*, 1–5 (2020).
47. Sarkodie, S. A. & Owusu, P. A. Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. *Environ. Res.* **191**, 110101 (2020).
48. Sil, A. & Kumar, V. N. Does weather affect the growth rate of COVID-19, a study to comprehend transmission dynamics on human health. *J. Saf. Sci. Resil.* **1**, 3–11 (2020).

49. Xie, J. & Zhu, Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci. Total Environ.* **724**, 138201 (2020).
50. Pan, J. *et al.* Warmer weather unlikely to reduce the COVID-19 transmission: an ecological study in 202 locations in 8 countries. *Sci. Total Environ.* **753**, 142272 (2020).
51. Ward, M. P., Xiao, S. & Zhang, Z. The role of climate during the COVID-19 epidemic in New South Wales, Australia. *Transbound. Emerg. Dis.* **67**, 2313–2317 (2020).
52. Tosepu, R. *et al.* Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Sci. Total Environ.* **725**, 138436 (2020).
53. Qi, H. *et al.* COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci. Total Environ.* **728**, 138778 (2020).
54. Guo, C. *et al.* Meteorological factors and COVID-19 incidence in 190 countries: An observational study. *Sci. Total Environ.* **757**, 143783 (2020).
55. Jahangiri, M., Jahangiri, M. & Najafgholipour, M. The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran. *Sci. Total Environ.* **728**, 138872 (2020).
56. Sharma, P., Singh, A. K., Agrawal, B. & Sharma, A. Correlation between weather and COVID-19 pandemic in India: An empirical investigation. *J. Public Affairs* **20**, e2222 (2020).
57. Rosario, D. K. A., Mutz, Y. S., Bernardes, P. C. & Conte-Junior, C. A. Relationship between COVID-19 and weather: Case study in a tropical country. *Int. J. Hyg. Environ. Health* **229**, 113587 (2020).
58. Mofijur, M. *et al.* Relationship between weather variables and new daily COVID-19 cases in Dhaka, Bangladesh. *Sustainability* **12**, 8319 (2020).
59. Bukhari, Q., Massaro, J., D'Agostino, R. & Khan, S. Effects of weather on coronavirus pandemic. *Int. J. Environ. Res. Public Health* **17**, 5399 (2020).
60. Rashed, E. A., Kodera, S., Gomez-Tames, J. & Hirata, A. Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: Multi-prefecture study in Japan. *Int. J. Environ. Res. Public Health* **17**, 5354 (2020).
61. Mecenas, P., Baston, R., Vallinoto, A. & Normando, D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. *PLoS One* **15**, e0238339 (2020).
62. Malki, Z. *et al.* Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos Solitons Fractals* **138**, 110137 (2020).
63. Federal Statistical Office. Current population. https://www.destatis.de/EN/Home/_node.html (2020). Accessed 05 January 2021.
64. Statistische Ämter des Bundes und der Länder. Bruttoinlandsprodukt (VGR) Ergebnisse der Volkswirtschaftlichen Gesamtrechnungen der Länder. https://www.statistikportal.de/en/node/649 (2020). Accessed 04 January 2021.
65. statista. Arbeitslosenquote in Deutschland nach Bundesländern. https://de.statista.com/statistik/daten/studie/36651/umfrage/arbeitslosenquote-in-deutschland-nach-bundeslaendern/ (2020). Accessed 04 January 2021.
66. Robert Koch Institute. Coronavirus disease 2019 (COVID-19): Daily situation report of the Robert Koch Institute. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html (2020). Accessed 31 December 2020.
67. Adams, A., Chen, X., Li, W. & Zhang, C. The disguised pandemic: The importance of data normalization in COVID-19 web mapping. *Public Health* **183**, 36–37 (2020).
68. Wijaya, K. *et al.* Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia. *Stoch. Environ. Res. Risk Assess.*, 1–20 (2020).
69. CDC-OpenData. Index of /climate_environment/CDC/. https://opendata.dwd.de/climate_environment/CDC/ (2020). Accessed 31 December 2020.
70. Xinhuanews. Germany's Bavaria declares emergency situation effective on Tuesday. http://www.xinhuanet.com/english/2020-03/16/c_138884534.htm (2020). Accessed 09 March 2021.
71. J. Mladek (Nordkurier). Bavaria imposes curfew! https://www.nordkurier.de/politik-und-wirtschaft/bayern-verhaengt-ausgangssperre-2038792303.html (2020). Accessed 04 January 2021.
72. Richard Connor. German states move closer to near-total lockdowns. https://www.dw.com/en/german-states-move-closer-to-near-total-lockdowns/a-52863482 (2020). Accessed 09 March 2021.
73. WELT. First major German city introduces mandatory masking. https://www.welt.de/politik/deutschland/article206911189/Coronavirus-Erste-deutsche-Grossstadt-fuehrt-Maskenpflicht-ein.html (2020). Accessed 04 January 2021.
74. L. Riekhoff and A. Sommer (streiflichter). Coronavirus in the Coesfeld district: 59 new infections with the coronavirus. https://www.streiflichter.com/lokales/coesfeld/coronavirus-kreis-coesfeld-aktuelle-fallzahlen-region-13643612.html (2020). Accessed 04 January 2021.
75. Deutsche Welle (DW). Coronavirus: Over 600 people test positive at German slaughterhouse. https://www.dw.com/en/coronavirus-over-600-people-test-positive-at-german-slaughterhouse/a-53846038 (2020). Accessed 04 January 2021.
76. BBC. Coronavirus: Thousands protest in Germany against restrictions. https://www.bbc.com/news/world-europe-53622797 (2020). Accessed 04 January 2021.
77. Das, P. & Choudhuri, T. Decoding the global outbreak of COVID-19: The nature is behind the scene. *Virus Dis.* **31**, 1–7 (2020).
78. Riddell, S., Goldie, S., Hill, A., Eagles, D. & Drew, T. W. The effect of temperature on persistence of SARS-CoV-2 on common surfaces. *Virol. J.* **17**, 1–7 (2020).
79. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
80. Geary, R. C. The contiguity ratio and statistical mapping. *Inc. Stat.* **5**, 115–146 (1954).
81. Cliff, A. & Ord, J. *Spatial Autocorrelation. Monographs in Spatial and Environmental Systems Analysis* (Pion, 1973).
82. Anselin, L. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Spat. Anal. Perspect. GIS* **4**, 111–116 (1996).
83. Sokal, R. R., Oden, N. L. & Thomson, B. A. Local spatial autocorrelation in a biological model. *Geogr. Anal.* **30**, 331–354 (1998).
84. Ocampo, S. & Rodríguez, N. An introductory review of a structural VAR-X estimation and applications. *Revista Colombiana de Estadística* **35**, 479–508 (2012).
85. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis* (Springer, 2005).
86. Prata, D. N., Rodrigues, W. & Bermejo, P. H. Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil. *Sci. Total Environ.* **729**, 138862 (2020).
87. Yuan, J. *et al.* Non-linear correlation between daily new cases of COVID-19 and meteorological factors in 127 countries. *Environ. Res.* **193**, 110521 (2020).
88. Bashir, M. F. *et al.* Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci. Total Environ.* **728**, 138835 (2020).
89. Menebo, M. M. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. *Sci. Total Environ.* **737**, 139659 (2020).
90. Lolli, S., Chen, Y. C., Wang, S. H. & Vivone, G. Impact of meteorological conditions and air pollution on COVID-19 pandemic transmission in Italy. *Sci. Rep.* **10**, 1–15 (2020).
91. Raftery, A. E. Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).

92. Akaike, H. Information theory and an extension of the maximum likelihood principle. in *Selected Papers of Hirotugu Akaike*. Springer Series in Statistics (Perspectives in Statistics) (eds. Parzen, E. *et al.*) (Springer, 1998).
93. Mansfield, E. R. & Helms, B. P. Detecting multicollinearity. *Am. Stat.* **36**, 158–160 (1982).
94. Johnston, J. *Econometric Methods* 2nd edn (McGraw Hill Higher Education, 1972).
95. Farrar, D. E. & Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **49**, 92–107 (1967).
96. Willis, C. E. & Perlack, R. D. Multicollinearity: Effects, symptoms, and remedies. *J. Northeast. Agric. Econ. Council* **7**, 55–61 (1978).
97. Hausman, J. A. Specification tests in econometrics. *Econometrica* **46**, 1251–1271 (1978).
98. Davidson, R. & MacKinnon, J. G. *Estimation and Inference in Econometrics. OUP Catalogue* (Oxford University Press, 1993).
99. Baltagi, B. H. & Li, Q. A lagrange multiplier test for the error components model with incomplete panels. *Econom. Rev.* **9**, 103–107 (1990).
100. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data* (MIT Press, 2002).

Acknowledgements
This research has been supported by the Ministry of Research, Technology, and Higher Education/National Research and Innovation Agency, Indonesia through the PUPT research grant scheme 2021.

Author contributions
N.C.G. and K.P.W. drafted the work and performed the computations; D.A, M.A., K.K.W.H.E, and N.C.G. interpreted data and conducted preliminary analysis; D.A. and K.P.W. managed funding acquisition. All authors reviewed earlier drafts and approved its final version.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to D.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publication IV

Ganegoda, N., Wijaya, K.P., Chavez, J.P., Aldila, D., Erandi, K.K.W., and Amadi, M.
Reassessment of contact restrictions and testing campaigns against COVID-19 via spatio-temporal modeling

Reprinted with permission from
Nonlinear Dynamics
Vol. 107, pp. 3085–3109, 2022.

Nonlinear Dyn
https://doi.org/10.1007/s11071-021-07111-w

ORIGINAL PAPER

Reassessment of contact restrictions and testing campaigns against COVID-19 via spatio-temporal modeling

Naleen Chaminda Ganegoda · Karunia Putra Wijaya · Joseph Páez Chávez · Dipo Aldila · K. K. W. Hasitha Erandi · Miracle Amadi

Received: 7 July 2021 / Accepted: 28 November 2021

Abstract Since the earliest outbreak of COVID-19, the disease continues to obstruct life normalcy in many parts of the world. The present work proposes a mathematical framework to improve non-pharmaceutical interventions during the *new normal* before vaccination settles herd immunity. The considered approach is built from the viewpoint of decision makers in developing countries where resources to tackle the disease from both a medical and an economic perspective are scarce. Spatial auto-correlation analysis via global Moran's index and Moran's scatter is presented to help modulate decisions on hierarchical-based priority for healthcare capacity and interventions (including possible vaccination), finding a route for the corresponding deployment as well as landmarks for appropriate border controls. These clustering tools are applied to sample data from Sri Lanka to classify the 26 Regional Director of Health Services (RDHS) divisions into four clusters by introducing convenient classification criteria. A metapopulation model is then used to evaluate the intra- and inter-cluster contact restrictions as well as testing campaigns under the absence of confounding factors. Furthermore, we investigate the role of the basic reproduction number to determine the long-term trend of the regressing solution around disease-free and endemic equilibria. This includes an analytical bifurcation study around the basic reproduction number using Brouwer Degree Theory and asymptotic expansions as well as related numerical investigations based on path-following techniques. We also introduce the notion of *average policy effect* to assess the effectivity of contact restrictions and testing campaigns based on the proposed model's transient behavior within a fixed time window of interest.

Keywords COVID-19 · Spatial auto-correlation · Metapopulation model · Bifurcation theory · Path-following-based continuation

N. C. Ganegoda
Department of Mathematics, University of Sri Jayewardenepura, Nugegoda 10250, Sri Lanka

K. P. Wijaya ( )
Mathematical Institute, University of Koblenz, D–56070 Koblenz, Germany
e-mail: karuniaputra@uni-koblenz.de

J. Páez Chávez
Center for Applied Dynamical Systems and Computational Methods (CADSCOM), Faculty of Natural Sciences and Mathematics, Escuela Superior Politécnica del Litoral, P.O. Box 09-01-5863, Guayaquil, Ecuador

J. Páez Chávez
Center for Dynamics, Department of Mathematics, TU Dresden, D–01062 Dresden, Germany

D. Aldila
Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

K. K. W. H. Erandi
Department of Mathematics, University of Colombo, Colombo 00700, Sri Lanka

M. Amadi
Department of Mathematics and Physics, Lappeenranta University of Technology, FI–53851 Lappeenranta, Finland

Published online: 20 December 2021

1 Introduction

COVID-19 outbreaks have been curtailing socio-economic activities around the globe. Over 150 million total confirmed cases had been reported by Apr 29, 2021, and the number of deaths exceeded 3 million by Apr 16, 2021, reflecting the burden of the pandemic [1]. This unprecedented health crisis has shown how far time and spatial propagation of incidence matter to each individual on a micro-scale and subsequently to a country on a macro-scale. Toward the ultimate herd immunity, several vaccines have been introduced; however, their efficacy must be scrutinized amidst virus mutations [2,3]. World Health Organization sets a minimum efficacy of 50% with a preferable threshold of 70% [4]. Although many of the vaccines are well above these efficacy levels, effectiveness in the field might be different due to the variations in affordability, public compliance, healthcare planning, etc. [5,6]. Moreover, equitable access to vaccines, in particular for developing countries, is also a challenging task [7]. Therefore, all the non-pharmaceutical interventions (NPIs) by means of contact restrictions (physical distancing, wearing face masks, washing hands, crowd clearance, workplace clearance, school closure, lockdown, public curfew, mobility restriction), and testing campaigns (including contact tracing) must be maintained until vaccination programs take substantial control over the further spread [8–10]. Many developing countries are still subject to financial restrictions against the import of vaccines [7], and at the same time, NPIs give a variable impact due to wavering laws and public compliance that mostly weigh upon socio-economic reasons [11]. As far as the spatial aspect is concerned, these NPIs should be implemented considering disease and societal impact according to international, national, and regional epidemiological situations [12]. Research on the actual performance of NPIs in developing countries is limited, and thus related government decisions usually are over- or underestimated [13,14]. It further creates a dilemma on what is more important between intra-regional and inter-regional contact restrictions, in particular for reopening the economy [15].

As vaccines with yet unknown success rates toward herd immunity are not even equally affordable across different economic classes, the only alternatives are enforcing laws and reshaping public awareness toward upholding NPIs. In relation to the spatial aspect, we start our investigation with the following questions:

(RQ1) On what sense may the decision maker appropriately perform the prioritization of healthcare capacity (e.g., hospital beds, ICU units, testing capacity, monitoring quarantine, including limited vaccines) among all spatial units in a country?

(RQ2) Under limited data of confounding factors, how can the decision maker value and reassess the flow of epidemics as well as the impeding NPIs?

This work puts up not only the prioritization of healthcare capacity and NPIs among spatial units but also the route for them in a more robust way than incidence-driven approaches. Endeavor to this has been known from the field of spatial mapping, namely to group spatial units into meaningful clusters.

In Sec. 2, we adopt global Moran's index and Moran's scatter to measure the timely spatial pattern of COVID-19 incidence in a country as well as to set the grouping. Particularly in developing countries, prioritizing high-risk areas or hotspots is driven by careful utilization of healthcare capacity [16]. The two aforementioned tools stand out among simplistic case mappings for their power to localize and group hotspots. Accordingly, priority for intra-cluster NPIs remains the same within a cluster but sequential between clusters. This strategy is important for developing countries like Sri Lanka that has not yet been covered by a holistic spatial analysis of this caliber. In addition to prioritization and route, the clustering study can bear the locations for placing border controls, which in this case are those in the main inter-cluster mobility streams. There remains, however, one caveat from these tools. That is, they are not able to parameterize the ongoing government decisions in terms of numbers and thus fail to impart how sensitive the incidence is against changes in those decisions.

Focusing on Sri Lanka, in Sec. 3 we propose a metapopulation model for Moran's clusters determined from available panel COVID-19 incidence data. The preference of the dynamic model over functional regression models stems from integrable mechanistic processes behind COVID-19 infection and that no spatio-temporal data of confounding factors were found. A complexity reduction is proposed based on the unavailability of related field data, resulting in a simple model but rational enough such that contact restrictions and testing campaigns are mediated. Sects. 3.2–3.4 are

then devoted to studying the likelihood if the incidence persists for a long time. To this, the model solution is compared with certain equilibria in the local sense whereby the basic reproduction number and effective reproduction numbers play the key role within. The model fitting as in Sec. 4 will provide a proxy to not only the approximate reproduction numbers but also non-observable dynamics, including contact matrix and the ongoing government decision on testing campaigns.

Finally, Sec. 5 extends the bifurcation analysis numerically using a path-following technique for the case where according to the fitting, the clusters are not strongly connected. In addition to this, the performance of the government decisions on contact restrictions and testing campaigns during the observations is reassessed via maximal *average policy effect*, which measures the average number of individuals per 1,000,000 inhabitants that could have been saved from COVID-19 infection on the virtue of better interventions. Scenarios to cost-to-benefit ratio are also presented alongside.

2 Spatio-temporal analysis

This section is devoted to answering question 1 in the context of Sri Lanka. Particularly under consideration is prioritization of healthcare capacity and NPIs as well as classification of the 26 Regional Director of Health Services (RDHS) divisions into Moran's clusters.

2.1 Study area and observation period

Sri Lanka is a South Asian island country situated in the Indian Ocean between latitudes 5°55' and 9°50' N and between longitudes 79°31' and 81°53' E. Sri Lanka has a population of about 21.9 million [17]. From an administrative perspective, the country is divided into 9 provinces that cater to 25 districts. In health administration, there are 26 Regional Director of Health Services (RDHS) divisions that mainly coincide with administrative districts, except the district Ampara that is covered by two RDHS divisions. The primary units of health administration are called Medical Officer of Health (MOH) areas. There are 356 MOH areas wherein the health surveillance activities are carried out [18]. Over 100,000 total confirmed cases and 600 deaths had been reported in Sri Lanka by Apr 24 and by Apr 13, 2021, respectively [19]. The public has been asked to follow health guidelines such as wearing masks, washing hands, and keeping one-meter distance since the early stage of the outbreak [20]. All the confirmed cases are directed to hospitals, and close contacts in addition to overseas returnees are requested to be quarantined [20]. The data used in this study are the daily new cases recorded by the Epidemiology Unit, Ministry of Health of Sri Lanka, spanning over the period from Nov 14 until Mar 31, 2021 [21]. Recording data in RDHS level began from Nov 14, which lies within the post-curfew period after major super-spreading events (apparel factory cluster [22] and fish market cluster [23]). Earlier to that, the data had been listed only according to the clusters arisen from super-spreading events and quarantine centers. This is due to that only several clusters were significant rather than a community level spread up to the end of Oct 2020 [21]. The RDHS-wise normalized daily new cases (per 1,000,000 inhabitants) are illustrated in Fig. 1. Note that no major mobility restrictions had been imposed within the observation period.

2.2 Global Moran's index and Moran's scatter

For spatial auto-correlation, interconnectivity between spatial units indexed by i and j is usually represented by a spatial weight matrix $W = (w_{ij})$. These weights can be designed according to shared boundaries of spatial units or distance between centers. The usual adjacency matrix can be an option, but a distance measure may better articulate connectivity since adjacency only captures interaction among neighbors. In our case, the distances d_{ij} among RDHS divisions R_i are based on placing appropriate centers $(x_i^{\mathrm{R}}, y_i^{\mathrm{R}})$, which are taken from averaging those from MOH areas M_k, namely $(x_k^{\mathrm{M}}, y_k^{\mathrm{M}})$, weighted by their population P_k. The centers consist of the latitude x_k^{M} and longitude y_k^{M} of the most attractive points, for example a city center, main administrative/commercial building, transport hub, main junction, etc. It then follows

$$(x_i^{\mathrm{R}}, y_i^{\mathrm{R}}) := \frac{\sum_{k:\mathrm{M}_k\in\mathrm{R}_i} \left(P_k x_k^{\mathrm{M}}, P_k y_k^{\mathrm{M}}\right)}{\sum_{k:\mathrm{M}_k\in\mathrm{R}_i} P_k}. \tag{1}$$

Now that the distances d_{ij} are computable by the standard Haversine formula, we take the power functional form [24] of the weight

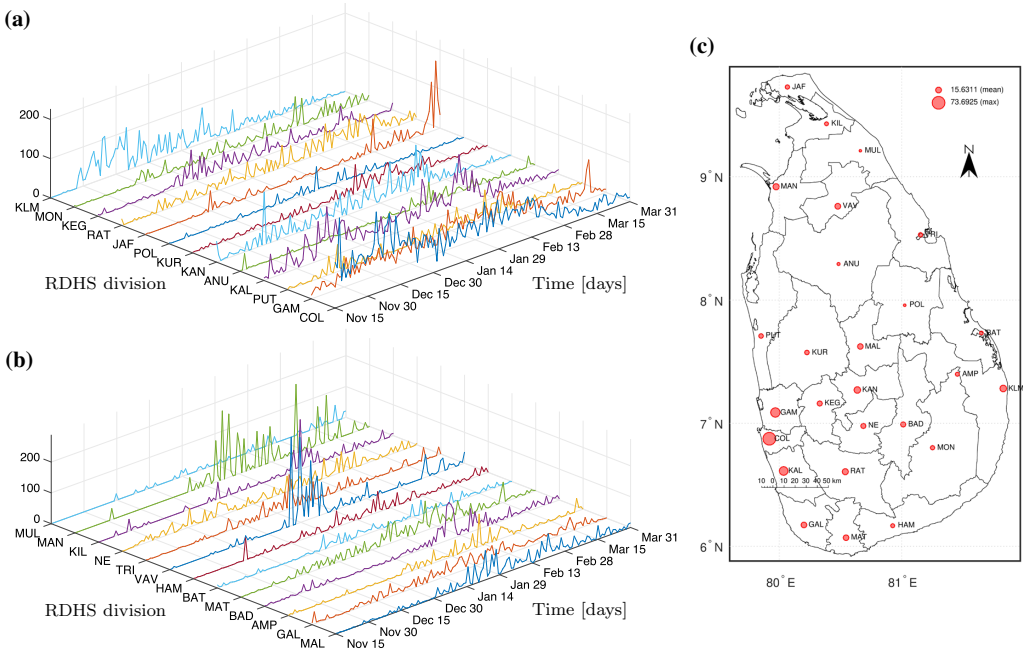


Fig. 1 Daily COVID-19 new cases in RDHS divisions per 1,000,000 inhabitants (**a–b**) and their timely average including spatial average and maximum (**c**). The RDHS divisions are COL (Colombo), GAM (Gampaha), PUT (Puttalam), KAL (Kalutara), ANU (Anuradhapura), KAN (Kandy), KUR (Kurunegala), POL (Polonnaruwa), JAF (Jaffna), RAT (Ratnapura), KEG (Kegalle), MON (Moneragala), KLM (Kalmunai), MAL (Matale), GAL (Galle), AMP (Ampara), BAD (Badulla), MAT (Matara), BAT (Batticaloa), HAM (Hambantota), VAV (Vavuniya), TRI (Trincomalee), NE (Nuwaraeliya), KIL (Kilinochchi), MAN (Mannar), and MUL (Mullaitivu)

$$w_{ij} := \begin{cases} \dfrac{d_{ij}^{-\delta}}{\sum_j d_{ij}^{-\delta}}, & d_{ij} < d,\ i \neq j \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

The exponential decay parameter $\delta > 0$ serves to scale the influence of the distance while the threshold distance $d > 0$ cuts the inessential interconnectivity. It is important to note that sufficiently large d values help make W irreducible, i.e., all regions become *strongly connected*.

Suppose that time is frozen and the mean normalized cases over the period shown in Fig. 1 for $S = 26$ RDHS divisions are reported as $C = (c_1, \cdots, c_S)$ with mean $\bar{c}$. Taking $Z = (z_1, \cdots, z_S) := C - \bar{c}\mathbb{1}$, the global Moran's index $\mathscr{I}$ [25] with a row stochastic matrix W as in (2) is given by

$$\begin{aligned} \mathscr{I} &:= \frac{Z^\top W Z}{Z \cdot Z} = \left(\frac{Z}{\|Z\|_2}\right)^\top W \frac{Z}{\|Z\|_2} \\ &= \left(\frac{Z}{\|Z\|_2}\right)^\top \left(\frac{W + W^\top}{2}\right) \frac{Z}{\|Z\|_2}. \end{aligned} \tag{3}$$

The global Moran's index basically is the Rayleigh quotient of $(W + W^\top)/2$ evaluated at Z, which brings the spatial autocovariance standardized by the variance of the data. The interpretation of the index usually comes in connection with the so-called Moran's scatter $(Z/\sigma_C, WZ/\sigma_C)$ where $\sigma_C := \sqrt{Z \cdot Z/S}$. It is quite apparent that the latter compares every spatial unit's self-incidence magnitude against the mean with the weighted magnitudes from its corresponding neighbors as *spatial lags* of the unit. The four Moran's clusters are then the cluster Q1 (first quadrant in 2-dimensional Euclidean space) referring to a set of spatial units of high incidence surrounded by their spatial lags of high

incidence (*high–high*, *hotspots*), the cluster Q2 (second quadrant) for spatial units of low incidence surrounded by their spatial lags of high incidence (*low–high*), the cluster Q3 (third quadrant) for those of low incidence surrounded by their spatial lags of low incidence (*low–low*, *coldspots*), and the cluster Q4 (fourth quadrant) for those of high incidence surrounded by their spatial lags of low incidence (*high–low*). We obtain two facts accordingly. First, the regressing line of $(Z/\sigma_C, WZ/\sigma_C)$ that passes through the origin has the slope $\mathscr{I}$. Second, if $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalue of the symmetric matrix $(W + W^\top)/2$, then the standard Rayleigh–Ritz (min–max) theorem (see e.g., [26] or [27]) suggests that $\lambda_{\min} := \min_{\|u\|_2=1} u \cdot (W + W^\top)u/2 \leq \mathscr{I} \leq \max_{\|u\|_2=1} u \cdot (W+W^\top)u/2 =: \lambda_{\max}$. This gives somewhat the tightest range due to $|\lambda_{\min}| < \lambda_{\max} = \rho((W + W^\top)/2) \leq \|W\|_2 \leq \sqrt{\|W\|_1\|W\|_\infty} = \sqrt{\|W\|_1} \leq (1+\|W\|_1)/2$. Since the diagonal entries of W are 0, evaluating the Rayleigh quotient at any of vectors in the standard basis of $\mathbb{R}^S$, namely $u = (0, \cdots, 0, 1, 0, \cdots, 0)$, yields $\lambda_{\min} \leq 0$. If $\mathscr{I} \to \lambda_{\max}$, then more points are aligned with the regressing line of that slope, making Q1 and Q3 full of points leaving out Q2 and Q4 scarce. A *locally clustered spatial pattern* is then observed. If $\mathscr{I} \to \lambda_{\min}$ and in case $\lambda_{\min} < 0$, then points are more concentrated in Q2 and Q4, indicating a *locally dispersed spatial pattern*. In between, under $\mathscr{I} \to 0$, there is no relation between self-incidence magnitudes and those from their spatial lags, leading to a *random spatial pattern*. We shall comment that the case $|\mathscr{I}| \leq 1$ may be observed in many cases where $\lambda_{\max} \leq 1$ but generally not always true. Besides assuring the upper bound 1 to any of the aforementioned bounds of $\lambda_{\max}$, sufficient conditions for this may include: W is symmetric (doubly stochastic) such that $\|W\|_1 = \|W\|_\infty = 1$, W and $W^\top$ commute in which case $\rho(W + W^\top) \leq \rho(W) + \rho(W^\top)$ [28], and W is diagonalizable since then $\rho(W + W^\top) = \rho(W) + \rho(W^\top)$.

As far as Sri Lankan data are concerned, a technical question arises: which values of δ and d in the weight matrix are suitable for the data? We answer this question by computing the smallest absolute elasticity indices of Moran's index on the average new cases. Now suppose that δ is decreased to a certain percentage ε_δ from its current value, i.e., $\delta \mapsto \delta - \varepsilon_\delta\delta$, where $0 < \varepsilon_\delta \leq 1$. In this way, $(\delta - \varepsilon_\delta\delta)/\delta = 1 - \varepsilon_\delta$ represents the total percentage post perturbation and ε_δ the percentage of increment. Taking this definition of ε_δ is more technically sound for a comparison among parameters as they may live in disparate scales. In response, $\mathscr{I} = \mathscr{I}(\delta, d)$ also changes from its initial data in the same fashion

$$\frac{\mathscr{I}(\delta - \varepsilon_\delta\delta, d)}{\mathscr{I}(\delta, d)} = 1 - \frac{\partial_\delta \mathscr{I}(\delta, d)}{\mathscr{I}(\delta, d)}\varepsilon_\delta\delta + \mathcal{O}(\varepsilon_\delta^2),$$
$$\frac{\mathscr{I}(\delta, d - \varepsilon_d d)}{\mathscr{I}(\delta, d)} = 1 - \frac{\partial_d \mathscr{I}(\delta, d)}{\mathscr{I}(\delta, d)}\varepsilon_d d + \mathcal{O}(\varepsilon_d^2).$$

For "fair" treatment, one usually designates $\varepsilon_\delta = \varepsilon_d = \varepsilon$, which is sufficiently small. Therefore, the first-order terms from the previous expressions take the role in determining which parameter, to which $\mathscr{I}$ is more sensitive. We then say $\mathscr{I}$ is more sensitive to the increase of δ than d in the regime

$$\frac{\partial_\delta \mathscr{I}(\delta, d)}{\mathscr{I}(\delta, d)}\delta > \frac{\partial_d \mathscr{I}(\delta, d)}{\mathscr{I}(\delta, d)}d. \tag{4}$$

In the literature, e.g., [29,30], these two compared expressions in (4) are called the (first-order) *elasticity indices*. There is one issue, namely the non-smoothness of the index with respect to d limits the definition to its approximation; see Fig. 2a–d. Apparently, Moran's index $\mathscr{I}$ is highly sensitive to d in case δ is relatively small ($1 \leq \delta \lesssim 5$) but insensitive to d as $\delta \gtrsim 9$. By $d = 1\text{e+}05$ m and $\delta = 9$, the elasticity indices are roughly zero, meaning that Moran's index changes only very slightly under the variation of (d, δ) in a neighborhood of these values. Additionally, plotting Moran's scatter on a daily basis (Fig. 2f) gives maximal concurrence percentages across RDHS divisions that agree with Moran's scatter on the average data (Fig. 2e). To the latter, we obtain $\mathscr{I} \approx 0.5687$ (p value ≈ 0.000324) indicating a locally clustered spatial pattern for the average data.

Accordingly, we classify the RDHS divisions as follows: cluster Q1 (COL, GAM, and KAL); cluster Q2 (KEG, MAL, GAL, and NE); cluster Q3 (PUT, KUR, ANU, POL, JAF, MON, AMP, BAD, MAT, BAT, HAM, VAV, TRI, KIL, and MUL); cluster Q4 (KAN, RAT, KLM, and MAN). From the application point of view, the cluster Q1 amasses all the hotspots. Ameliorating the burdens of infection follows from putting a first-level priority on healthcare capacity and possible vaccinations as well as providing more strict border controls that would reduce mobility *from and to* its spatial lags, i.e., neighbors in the sense of the weight matrix.

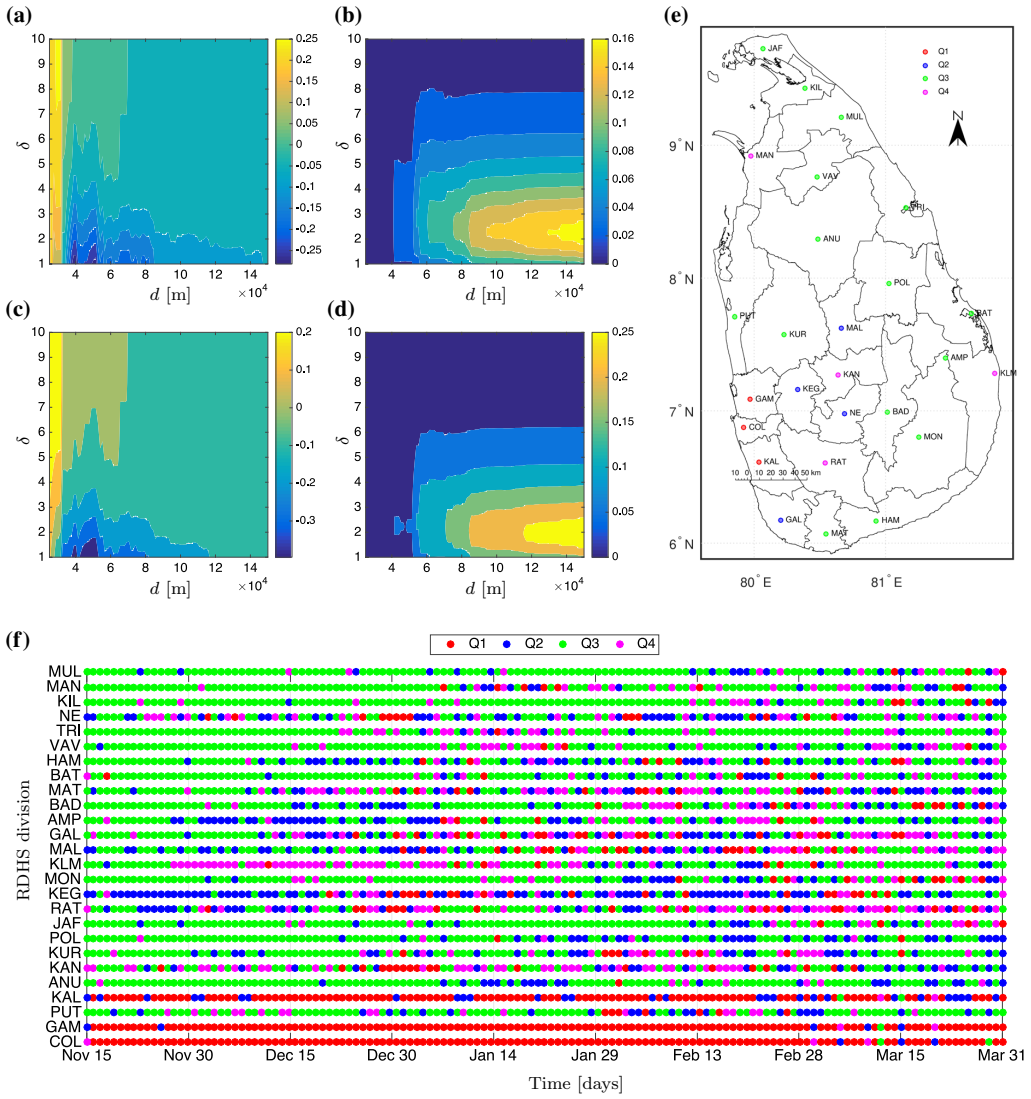


Fig. 2 Approximates of the elasticity indices $[\mathscr{I}(d+\varepsilon d,\delta)-\mathscr{I}(d,\delta)]/[\varepsilon\mathscr{I}(d,\delta)]$ using (**a**) $\varepsilon = 50\%$ and (**b**) $\varepsilon = 100\%$ and $[\mathscr{I}(d,\delta+\varepsilon\delta)-\mathscr{I}(d,\delta)]/[\varepsilon\mathscr{I}(d,\delta)]$ using (**c**) $\varepsilon = 50\%$ and (**d**) $\varepsilon = 100\%$. Moran's scatter plot using $d =$ 1e+05 m and $\delta = 9$ for the average daily new cases is presented in (**e**) while (**f**) gives the timely Moran's scatter

Intra-cluster border controls cannot change the situation much, but interventions can be realized through the applications of NPIs including public curfew and testing campaigns. We argue that the intra-cluster prioritization as well as deployment route for NPIs can be left to the decision maker, which can rely on the available resources. The cluster Q4 requires not only a second-level priority on healthcare capacity but also mobility restrictions *to* its spatial lags, otherwise the epidemics outwardly *diffuses*. Meanwhile, the cluster

Q2 may receive a third-level priority as well as isolation *from* its spatial lags such that it does not *attract* epidemics. Lives in the cluster Q3 can be the easiest in terms of mobility restrictions as long as reasonable hygiene practices and physical distancing are upheld. Border controls can now be localized to any point that shares the borders between clusters, which could be an intersection point on the main road or an intersecting railway station. For example, no border controls are required between KUR and ANU, but between KUR and COL. If the authorities follow a clustering based on administrative provinces, the border between KUR and ANU should be controlled as they belong to separated provinces. Thus, our analysis suggests more scientific clustering that may overrule general administrative choices.

3 Modeling-based reassessment of NPIs

Now that the 26 RDHS divisions are classified into the four Moran's clusters, this section is devoted to modeling the incidence dynamics on the clusters and parameterizing ongoing government decisions on contact restrictions and testing campaigns toward answering 1. Here is the idea: Once the essential performance measures for contact restrictions and testing campaigns are gained through the model fitting, we can optimize the model toward specific goals whereby different magnitudes of the measures are tested. We focus on three goals, namely minimizing the basic reproduction number, maximizing the average policy effect, and minimizing the associated policy cost. All forms of NPIs can be reassessed toward these goals.

The nature of standard metapopulation models suggests that, in contrast to the kinematic models, the whereabouts of every single individual are no longer concerned. Among first generations of metapopulation model, two-patch models were proposed for their accessibility to sophisticated analytical investigation on the disease endemicity via the basic reproduction number $\mathscr{R}_0$ [31–34]. These studies shared similar results: the disease-free equilibrium is globally asymptotically stable if $\mathscr{R}_0 < 1$, and all the state variables are uniformly persistent if $\mathscr{R}_0 > 1$ leading to the existence of an endemic equilibrium, which was proven to be globally asymptotically stable. The SIS model in [34] stands out among the cited models as it incorporates infections during travels. General n-patch SIR-type models considering mobile humans with *memory* over their origin zones admit short visits to other zones that allow them to infect other humans or to acquire infections ex-situ [35–37]. The notion of *transit time* becomes the key determinant to the latter. Models without memory were proposed in [38,39] where in [38], a more generalized population growth function was used, taking into account the relationship between $\mathscr{R}_0$ and the disease extinction and persistence. Metapopulation models for COVID-19 have also appeared recently. Citron *et. al.* [40] consider metapopulation versions of an SIR, an SIS, and a Ross-Macdonald model integrating Eulerian movement (direct out-flux) and Lagrangian movement model (net out-flux and influx). The two movement models were analyzed to synthesize conditions under which one model can be superior against the other with respect to epidemiological outcomes. A model including transit time and infection due to exposed cases was proposed in [41] With known network attributes of the tested case, the study was brought to determine the infection rates and the ratio between asymptomatic and symptomatic cases. Recently, metapopulation models including vaccinated compartments [42] and age structure [43] were proposed and validated using field data. A memory-less migration (or diffusion) model including human mobility [44] was used for modeling daily confirmed cases on a network of 343 cities in China.

In this study, our model is concerned with the COVID-19 epidemics that naturally include undetected and deceased cases. We use the concept memory in the model, but unlike in [35–37], the number of humans from cluster i that are in cluster j and thus at which cluster the contacts happen, are not displayed. In contrast to [42], we combine the infection rate and the matrix representing the fraction of total daily time for i-residents to be in j-region into what we called a *contact matrix*.

3.1 Metapopulation model for COVID-19 epidemics and biological assumptions

We divide the regional population N_i $(i = 1, \cdots, D)$ into five subpopulations based on their health status: susceptible S_i (healthy but vulnerable to infection), detected *active* cases I_i comprising some portions of asymptomatic and symptomatic (hospitalized) cases, undetected cases U_i (*dark figures*, mostly asymp-

tomatic), recovered R_i, and deceased cases D_i. Due to a small incubation duration, we count the intermediate exposed (pre-symptomatic) cases to the susceptible subpopulation to simplify the model presentation. Net population growth due to imports, migrations, natural births, and deaths is assumed to be negligible during the observation period, inducing constant total cluster population N_i. The point of departure from our modeling is concerned with

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = \tilde{\mu}(S_i + I_i + U_i + R_i) - S_i \sum_{j=1}^{D} \tilde{\beta}_{ij} \frac{(I_j + \varrho U_j)}{N_j} + \tilde{\eta} R_i - \tilde{\mu} S_i, \tag{5a}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = \alpha S_i \sum_{j=1}^{D} \tilde{\beta}_{ij} \frac{(I_j + \varrho U_j)}{N_j} - (\tilde{\gamma} + \tilde{\mu}) I_i, \tag{5b}$$

$$\frac{\mathrm{d}U_i}{\mathrm{d}t} = (1-\alpha) S_i \sum_{j=1}^{D} \tilde{\beta}_{ij} \frac{(I_j + \varrho U_j)}{N_j} - (\tilde{\gamma} + \tilde{\mu}) U_i, \tag{5c}$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = \tilde{\gamma}(1-m) I_i + \tilde{\gamma} U_i - (\tilde{\eta} + \tilde{\mu}) R_i, \tag{5d}$$

$$\frac{\mathrm{d}D_i}{\mathrm{d}t} = \tilde{\gamma} m I_i. \tag{5e}$$

In this basic model, $\tilde{\beta}_{ij}$ denotes the infection rate that determines the likelihood of a susceptible person from ith region to meet with an infected person from jth region. In the standard SIR models for airborne diseases, the infection rates depend on many factors including sneezing rate, probability of sneezing during encounters [45], infectiousness measure (viral load, case index) [46,47], effectiveness measure determining how probable an average susceptible person contracts infection (health condition, age) [45], human mobility for bearing corrections of the possible number of encounters [45], influence of media reports on public awareness [48], and possibly weather factors that enhance aerosol transmissions [49–51]. As the exposed cases were gathered in S_i, the infection rates also give another correction as the individuals cannot both be infected and spread the viruses. After contracting infection, the remaining time known as *viral shedding period* $1/\tilde{\gamma}$ determines the average duration from the onset of symptoms until the cessation of viral shedding (when viruses can no longer be released from an infected person), indicating the end of the infectiousness period [52,53]. The parameters $1/\tilde{\eta}$ and m denote the duration of temporary immunity and the fatality rate from the detected cases. We impose a strong assumption that during the limited observations, the entire infected cases are timely distributed into the detected and undetected cases with the average proportions α and $1-\alpha$, respectively. To accommodate different transmission scales from detected and undetected cases, we have used the parameter $\varrho > 1$. Finally, $\tilde{\mu}$ denotes the natural birth or death rate.

Our next task is to simplify the model even further. Due to unknown dark figures U_i, several ideas and estimates have been appearing in the literature, see e.g., [54]. Ours is based on the assumption that the dark figures proportionate the detected cases to a certain constant, i.e., $U_i = pI_i$ where $0 < p < 1$ for all clusters i and time. By the range of p, we impose that most cases are detected. As we specify $\alpha = 1/(1+p)$, Eqs. (5b) and (5c) apparently become equivalent. This choice justifies the idea that the constant ratio between undetected and detected cases requires constant detection rate (in the sense of averaging) and that the detection rate also holds $0 < \alpha < 1$. Apart from this, we bring forward the non-observability assumption to R_i due to data credibility. Our study designates R_i as to proportionate D_i to a certain constant from time to time, namely $R_i \simeq \eta D_i/(\tilde{\mu} + \tilde{\eta} - \eta)$ for a new parameter $0 < \eta < \tilde{\mu} + \tilde{\eta}$ and all i. It is straightforward to see that $(\tilde{\mu} + \tilde{\eta}) R_i = \eta(R_i + D_i) = \eta[N_i - S_i - (1+p) I_i]$. In the next model presentation, we would like to use the re-scaled variables $S_i \leftarrow cS_i/N_i$ and $I_i \leftarrow cI_i/N_i$ with $c = 10^6$ as well as the following notations $\mu := \tilde{\mu}(1+p)$, $\gamma := \tilde{\gamma} + \tilde{\mu}$, $S := (S_1, \cdots, S_4)^\top$, $I := (I_1, \cdots, I_4)^\top$, $\beta := (\beta_{ij})$ as the *contact matrix*, $\mathrm{diag}(S)$ as the diagonal matrix whose main diagonal is S, and $\mathbb{1}$ as a matrix or a vector whose entries are 1. We acquire the final model

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu I - (1+\varrho p)\mathrm{diag}(S)\frac{\beta}{c} I + \eta[c\mathbb{1} - S - (1+p) I], \tag{6a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{(1+\varrho p)}{1+p}\mathrm{diag}(S)\frac{\beta}{c} I - \gamma I, \tag{6b}$$

with an initial value (S_0, I_0). This model portrays the situation where all infected cases are distributed to the detected classes in case $p = 0$, i.e., when the quality of the testing campaigns is extremely good. Moreover, as much as half of the infected cases will be distributed to

the detected cases when no essential tests are done, i.e., when $p = 1$. In Sri Lanka, PCR and antigen tests are carried out on a random and targeted basis [55]. However, limited financial allocations may curtail arbitrary increase in testing capacity [56]. Another factor for a large p is the compromised public compliance to tracing technology that tolerates the effectiveness [57]. As a result, lack of tests retards the process of unraveling possibly infected close contacts and thus hotspot identification, which eventually delays blocking the routes of transmission [58].

3.2 Basic reproduction number

We study the basic reproduction number to determine the local behavior of model solution around the disease-free (DFE) and endemic equilibrium (EE) for fleeting observations. Therefore, given that optimal parameters are subject to data availability, the predictive power of this behavioral analysis is limited to a short-range prediction window. Let x^* be a point of interest that is compared to the solution of the model $\mathrm{d}x/\mathrm{d}t = f(x)$ represented by (6). For simplicity, we assume that x^* is either DFE or EE such that $f(x^*) = 0$. The error measure $z := x - x^*$ then follows $\mathrm{d}z/\mathrm{d}t = \nabla f(x^*)z + \mathcal{O}(\|z\|^2) \approx \nabla f(x^*)z$ providing that z is close enough to 0 or x is close to x^*. A compelling property of such a linearized system is that the short-term trend of the model solution around 0 can be predicted by the local (even global) stability. The basic reproduction number $\mathscr{R}_0$ will then be used to parameterize a condition for the maximal real part of the eigenvalues of Jacobian matrix $\nabla f(x^*)$, which eventually determines the local stability of z around 0.

When $x^* = \mathrm{DFE} = (c\mathbb{1}, 0)$, we obtain

$$\nabla f(x^*) = \begin{pmatrix} -\eta\mathrm{id} & [\mu - \eta(1+p)]\mathrm{id} - (1+\varrho p)\beta \\ 0 & \frac{(1+\varrho p)}{1+p}\beta - \gamma\mathrm{id} \end{pmatrix}, \tag{7}$$

where id denotes the identity matrix. The next generation matrix as well as the basic reproduction number can now be formulated as

$$\begin{aligned} G &:= FV^{-1} = \frac{F}{\gamma} \\ &\text{with } F := \frac{(1+\varrho p)}{1+p}\beta,\ V := \gamma\mathrm{id} \text{ and} \\ \mathscr{R}_0 &:= \rho(G), \end{aligned} \tag{8}$$

respectively. Here, $\rho(G)$ denotes the spectral radius of G. According to Berman and Plemmons [59], $V - F$ becomes a nonsingular M–matrix if and only if $\gamma > \rho(F)$ or $1 > \mathscr{R}_0$. The fact that λ being an eigenvalue of G is equivalent to $\lambda - 1$ being the corresponding eigenvalue of $G - \mathrm{id}$ (with the non-changing eigenvectors), Perron–Frobenius Theorem on simplicity and dominance of $\mathscr{R}_0$ also guarantees that $\mathscr{R}_0 < 1$ holds true if and only if all other eigenvalues of $G - \mathrm{id}$ (or $F - V$) lie in the open disk of center -1 and radius $\mathscr{R}_0$ (or center $-\gamma$ and radius $\gamma\mathscr{R}_0$). Additionally, we acquire another fact that $G - \mathrm{id}$ (or $F - V$) becomes singular if and only if $\mathscr{R}_0 = 1$, in which case a zero eigenvalue occurs. Due to the equivalence relations, the final case $\mathscr{R}_0 > 1$ happens if and only if there exists at least one eigenvalue of $F - V$ with a positive real part. We obtain the following summary: z is attracted to 0 or DFE becomes locally attractive to the solution x of (6) if $\mathscr{R}_0 < 1$ and it becomes locally repelling to x if $\mathscr{R}_0 > 1$.

3.3 Existence and attractiveness of an endemic equilibrium

Computing an endemic equilibrium (EE) from model (6) also returns complexity on its own. The second subsystem (6b) gives the equilibrium equation

$$S_i = \frac{(1+p)\gamma c}{(1+\varrho p)} \cdot \frac{I_i}{\sum_j \beta_{ij} I_j}, \tag{9}$$

for all j. Substituting this to the first subsystem (6a) together with $(1+\varrho p)\mathrm{diag}(S)\beta I/c = (1+p)\gamma I$ also multiplying every i-th entry with $\sum_j \beta_{ij} I_j/\eta c$ gives us

$$\begin{aligned} &\frac{[\mu - (1+p)(\gamma+\eta)]}{\eta c} I_i \cdot \sum_j \beta_{ij} I_j \\ &\quad + \sum_j \beta_{ij} I_j - \frac{(1+p)\gamma}{(1+\varrho p)} I_i = 0, \end{aligned} \tag{10}$$

for all i. The preceding system of equations folds under multiplication by $-(1+\varrho p)/(1+p)\gamma$ into

$$\begin{aligned} \varphi(I) &:= I - GI \\ &\quad + \underbrace{\left[\frac{[(1+p)(\gamma+\eta) - \mu]}{\eta c} \cdot \frac{(1+\varrho p)}{(1+p)\gamma}\right]}_{=:K} \mathrm{diag}(I)\beta I \\ &= 0, \end{aligned} \tag{11}$$

where G denotes the next generation matrix as in (8). This is a multidimensional quadratic equation whose solutions cannot be derived explicitly. In order to guarantee the existence of EE, we first see if the point $(\mathscr{R}_0 = 1, I = 0)$ is indeed a branch point of the quadratic equation. Since $\varphi : \Omega \to \mathbb{R}^D$ is a quadratic function defined on some open subset $\Omega \subseteq \mathbb{R}^D$, it is verifiable that $\varphi \in C^\infty(\Omega)$. Let $q \in \mathbb{R}^D$ be a point such that $q \notin \varphi(\partial\Omega)$, where $\partial\Omega$ denotes the boundary of Ω. The point q is said to be *regular* if either $\varphi^{-1}(q) = \phi$ for all points $I^* \in \varphi^{-1}(q)$ return invertible $\nabla_I \varphi(I^*)$. Otherwise, q is called *critical*.

Adopting definitions from [60,61], the map $\mathcal{B} : C^1(\Omega) \times \Omega \times \mathbb{R}^D \to \mathbb{Z}$ defined as

$$\mathcal{B}(\varphi, \Omega, q) := \begin{cases} \sum_{I^* \in \varphi^{-1}(q)} \text{sign det} \nabla_I \varphi(I^*), & q \text{ regular} \\ \mathcal{B}(\varphi, \Omega, \tilde{q}), & q \text{ critical} \end{cases} \tag{12}$$

with $\tilde{q}$ being regular such that $\|q - \tilde{q}\| < \inf_{s \in \varphi(\partial\Omega)} \|q - s\|$, denotes the Brouwer degree of φ in Ω with respect to a reference point q. Another convention narrows the singular value down to $I = 0$ with the neighborhood Ω of 0 is chosen in such a way that $I = 0$ is isolated. In this case, the map

$$\text{ind}(\varphi, 0) := \mathcal{B}(\varphi, \Omega, 0) \tag{13}$$

defines the index of φ at the isolated singular value $I = 0$. According to the last two references, $(\mathscr{R}_0 = 1, I = 0)$ is a branching point providing that $\text{ind}(\varphi, 0)$ changes values around $\mathscr{R}_0 = 1$.

In case $\mathscr{R}_0 < 1$, the fact that the multiplication between complex conjugate numbers return a positive number, gives us $\det \nabla_I \varphi(0) = \det(\text{id} - G) = \Pi_i(1 - \lambda_i) > 0$. We can always impose continuous perturbation on parameters $s = s(\varepsilon) \in \{\mu, \varrho, p, \beta_{11}, \cdots, \beta_{44}, \eta, \gamma\}$ such that $s(0)$ solves $\mathscr{R}_0(0) = 1$ and $s(\varepsilon)$ is equivalent to $\mathscr{R}_0(\varepsilon) > 1$ for $0 < \varepsilon < \tilde{\varepsilon}$ and some $\tilde{\varepsilon}$. In case $\varepsilon = 0$, the eigenvalue of $\text{id} - G$ with the largest real part apparently returns $1 - \mathscr{R}_0 = 0$ and the other eigenvalues lie in the open disk of center -1 and radius 1. We can appoint the eigenvalue $\hat{\lambda}$ of $\text{id} - G$ with the largest negative real part and of algebraic multiplicity $a_m(\hat{\lambda}) \geq 1$, and define $\hat{r} := 1 - \Re\hat{\lambda}$. The function $\Phi(\lambda, \varepsilon) := \det([1 - \lambda]\text{id} - G(\varepsilon))$ with $G(0)$ corresponding to $\mathscr{R}_0(0) = 1$ is holomorphic in λ and continuous in ε. We can appoint $r < \hat{r}$ such that $\hat{\lambda}$ is the only root in the closed disk $\overline{\mathbb{D}(\hat{\lambda}, r)}$. There must now exist $\hat{\varepsilon} \leq \tilde{\varepsilon}$ such that

$$|\Phi(\lambda, \varepsilon) - \Phi(\lambda, 0)| < |\Phi(\lambda, 0)|$$

holds for all $\lambda \in \partial\mathbb{D}(\hat{\lambda}, r)$ and $0 < \varepsilon < \hat{\varepsilon}$. According to Rouché's Theorem [62], $\Phi(\lambda, \varepsilon)$ has roots in $\mathbb{D}(\hat{\lambda}, r)$ of counting multiplicities $a_m(\hat{\lambda})$ when $0 < \varepsilon < \hat{\varepsilon}$. The same continuity argument can be used to show that all the remaining eigenvalues can never have largest negative real part which exceeds $\Re\hat{\lambda} + r$. In summary, as $0 < \varepsilon < \hat{\varepsilon}$ for a new $\hat{\varepsilon}$ it holds that 1 is not an eigenvalue of G and $\mathscr{R}_0$ slightly increases from 1 such that

$$\mathscr{R}_0 > 1 > |\lambda| \tag{14}$$

for all eigenvalues $\lambda \neq \mathscr{R}_0$ of G. This returns two results: (i) $\text{id} - G$ becomes non-singular such that $I = 0$ serves as an isolated singular value of φ in its neighborhood Ω due to $\varphi(I) = \varphi(0) + \nabla_I \varphi(0) \cdot I + \mathcal{O}(\|I\|^2) \approx I - GI$ there and (ii) $\det \nabla_I \varphi(0) = \det(\text{id} - G) = (1 - \mathscr{R}_0)\Pi_{i:\lambda_i \neq \mathscr{R}_0}(1 - \lambda_i) < 0$. The index of φ at the singular value $I = 0$ now reads as

$$\text{ind}(\varphi, 0) = \text{sign } \det \nabla_I \varphi(0) = \begin{cases} 1, & \mathscr{R}_0 < 1 \\ -1, & \mathscr{R}_0(\varepsilon) > 1 \end{cases},$$

for $0 < \varepsilon < \hat{\varepsilon}$. This confirms that that $(\mathscr{R}_0 = 1, I = 0)$ is indeed a branching point.

The next task is to verify the positivity of the local branch. We took the asymptotic expansion of $\mathscr{R}_0$ from 1 [63,64], i.e., the coefficient of $-G$ in the quadratic equation (11) via the direct relation between $\mathscr{R}_0$ and ε:

$$1 = \frac{1}{\mathscr{R}_0} + \frac{\mathscr{R}}{\mathscr{R}_0}\varepsilon + \mathcal{O}(\varepsilon^2), \quad 0 < \varepsilon \ll 1 \tag{15}$$

such that the branch I takes the expansion

$$I = \psi_1 \varepsilon + \psi_2 \varepsilon^2 + \mathcal{O}(\varepsilon^3). \tag{16}$$

Substituting the preceding expressions to the quadratic equation (11) returns

$$\begin{aligned} 0 = &\left[\psi_1 \varepsilon + \psi_2 \varepsilon^2 + \mathcal{O}(\varepsilon^3)\right] - \left[\frac{1}{\mathscr{R}_0} + \frac{\mathscr{R}}{\mathscr{R}_0}\varepsilon + \mathcal{O}(\varepsilon^2)\right] \\ &G\left[\psi_1 \varepsilon + \psi_2 \varepsilon^2 + \mathcal{O}(\varepsilon^3)\right] \\ &+ K \text{diag}\left[\psi_1 \varepsilon + \psi_2 \varepsilon^2 + \mathcal{O}(\varepsilon^3)\right] \beta \left[\psi_1 \varepsilon + \psi_2 \varepsilon^2 + \mathcal{O}(\varepsilon^3)\right]. \end{aligned}$$

Zeroing the first-order term (ε) gives us

$$G\psi_1 = \mathscr{R}_0 \psi_1. \tag{17}$$

This means that ψ_1 is the eigenvector of G associated with $\mathscr{R}_0$, whose existence and positivity are guaranteed

by Perron–Frobenius Theorem. The latter also guarantees the existence and positivity of the left eigenvector ξ_1 associated with $\mathscr{R}_0$. Now, multiplying the second-order term (ε^2) with $\xi_1^\top$ from the left gives us

$$\underbrace{\left[\xi_1^\top - \frac{1}{\mathscr{R}_0}\xi_1^\top G\right]}_{=0}\psi_2 - \mathscr{R}\xi_1^\top\psi_1 + K\xi_1^\top \mathrm{diag}(\psi_1)\beta\psi_1 = 0 \tag{18}$$

whereby

$$\mathscr{R} = \frac{K\xi_1^\top \mathrm{diag}(\psi_1)\beta\psi_1}{\xi_1^\top\psi_1} > 0 \tag{19}$$

by $K > 0$. Moreover, substituting $(1+\varrho p)\mathrm{diag}(S)\beta I/c = (1+p)\gamma I$ from (6b) to (6a) leads us to the following summary

$$\begin{aligned} S &= c\mathbb{1} - \frac{(1+p)(\gamma+\eta)-\mu}{\eta}\psi_1\varepsilon + \mathcal{O}(\varepsilon^2),\\ I &= \psi_1\varepsilon + \mathcal{O}(\varepsilon^2),\\ \mathscr{R}_0 &= 1 + \mathscr{R}\varepsilon + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{20}$$

These parametric expressions suggest that as ε increases from 0, $\mathscr{R}_0$ increases from 1 and a unique local branch I takes off from 0 with the initial direction ψ_1 with respect to ε whereby the susceptible state decreases from c simultaneously for all clusters. Finally, one yields

$$\begin{aligned} &\partial_{\mathscr{R}_0} S\big|_{\mathscr{R}_0=1} = -\frac{[(1+p)(\gamma+\eta)-\mu]}{\eta\mathscr{R}}\psi_1 < 0\\ &\text{and}\quad \partial_{\mathscr{R}_0} I\big|_{\mathscr{R}_0=1} = \frac{\psi_1}{\mathscr{R}} > 0. \end{aligned} \tag{21}$$

This indicates the existence of a continuum of endemic equilibria in the neighborhood of $\mathscr{R}_0 = 1$ and in the direction of increasing $\mathscr{R}_0$. For $0 < \varepsilon \ll 1$, let us write one endemic equilibrium EE as $x^* = x^*(\varepsilon)$ with the expression given in (20). The Jacobian matrix evaluated at EE takes the form

$$\nabla f(x^*;\varepsilon) = \begin{pmatrix} -\eta\mathrm{id} & [\mu-\eta(1+p)]\mathrm{id} - (1+\varrho p)\beta \\ 0 & \frac{(1+\varrho p)}{1+p}\beta - \gamma\mathrm{id}\end{pmatrix} + \underbrace{\begin{pmatrix} -(1+\varrho p)\mathrm{diag}(\frac{\beta}{c}\psi_1) & (1+p)K\mathrm{diag}(\psi_1)\beta \\ \frac{1+\varrho p}{1+p}\mathrm{diag}(\frac{\beta}{c}\psi_1) & -K\mathrm{diag}(\psi_1)\beta\end{pmatrix}\varepsilon}_{=:\mathcal{E}} + \mathcal{O}(\varepsilon^2).$$

The matrix in the leading order ε^0 has eigenvalues $-\eta$ of algebraic multiplicity 4 and $\gamma(\lambda-1)$ where λ are the eigenvalues of G. Due to simplicity and dominance of $\mathscr{R}_0 = 1$, all the eigenvalues $\gamma(\lambda-1)$ locate in the open disk of center $-\gamma$ and radius $\gamma\mathscr{R}_0$ where only $\gamma(\mathscr{R}_0-1)$ is in the origin. We can use Rouché's Theorem one more time with the function $\Phi(\lambda,\varepsilon) := \det(\nabla f(x^*;\varepsilon) - \lambda\mathrm{id})$ to show that all eigenvalues of $f(x^*;\varepsilon)$, except the one that corresponds to $\mathscr{R}_0$, stay in the open left-half plane in $\mathbb{C}$ for a sufficiently small ε.

The fate of this last eigenvalue can be analyzed as follows. The eigenvalue $\gamma(\mathscr{R}_0-1)$ of $\nabla f(x^*;0)$ associates with the right and left eigenvector (by $\gamma\xi_1$ and $\gamma\psi_1$ of γG):

$$v_L := \gamma\begin{pmatrix}0\\ \xi_1\end{pmatrix} \quad\text{and}\quad v_R := \gamma\begin{pmatrix}\frac{[\mu-\eta(1+p)]\psi_1-(1+\varrho p)\beta\psi_1}{\eta+\mathscr{R}_0}\\ \psi_1\end{pmatrix}$$

respectively. Using Taylor expansion on a simple eigenvalue of a perturbed matrix [65], we obtain the eigenvalue of the Jacobian matrix that corresponds to $\mathscr{R}_0$:

$$\begin{aligned} \mathrm{eig}(\nabla f(x^*;\varepsilon);\mathscr{R}_0) &= \gamma(\mathscr{R}_0-1) + \frac{v_L^\top \mathcal{E} v_R}{v_L^\top v_R} + \mathcal{O}(\|\mathcal{E}\|^2)\\ &\le \mathscr{R}_0 - 1 + \frac{1}{\xi_1^\top\psi_1}(0 \quad \xi_1)^\top\\ &\quad \begin{pmatrix} -(1+\varrho p)\mathrm{diag}\left(\frac{\beta}{c}\psi_1\right) & (1+p)K\mathrm{diag}(\psi_1)\beta \\ \frac{1+\varrho p}{1+p}\mathrm{diag}\left(\frac{\beta}{c}\psi_1\right) & -K\mathrm{diag}(\psi_1)\beta\end{pmatrix}\\ &\quad \begin{pmatrix}\frac{[\mu-\eta(1+p)]\psi_1-(1+\varrho p)\beta\psi_1}{\eta+\mathscr{R}_0}\\ \psi_1\end{pmatrix}\varepsilon + \mathcal{O}(\varepsilon^2)\\ &= \frac{K\xi_1^\top\mathrm{diag}(\psi_1)\beta\psi_1}{\xi_1^\top\psi_1}\varepsilon\\ &\quad - \left[\frac{K\xi_1^\top\mathrm{diag}(\psi_1)\beta\psi_1}{\xi_1^\top\psi_1} + \frac{1+\varrho p}{1+p}\xi_1^\top\mathrm{diag}\left(\frac{\beta}{c}\psi_1\right)\right.\\ &\quad \left.\times\frac{[-\mu+\eta(1+p)]\psi_1 + (1+\varrho p)\beta\psi_1}{\eta+\mathscr{R}_0}\right]\varepsilon + \mathcal{O}(\varepsilon^2)\\ &= -\left[\frac{1+\varrho p}{1+p}\xi_1^\top\mathrm{diag}\left(\frac{\beta}{c}\psi_1\right)\right.\\ &\quad \left.\frac{[-\mu+\eta(1+p)]\psi_1 + (1+\varrho p)\beta\psi_1}{\eta}\right]\varepsilon + \mathcal{O}(\varepsilon^2) \end{aligned}$$

by substituting $\mathscr{R}_0$ from (20) and taking Taylor expansion over $1/(1+\mathscr{R}_0/\eta)$. This shows the existence of $\bar{\varepsilon} \le \hat{\varepsilon}$ where $0 < \varepsilon < \bar{\varepsilon}$ implies all eigenvalues of $\nabla f(x^*;\varepsilon)$ having negative real part. Combining with (20) and (21), we acquire a forward bifurcation of the model system (6) at $\mathscr{R}_0 = 1$. This means that the local branch of EEs becomes locally attractive as $\mathscr{R}_0 > 1$.

3.4 Effective reproduction numbers

Providing epidemics are going on, $I > 0$, we have from (26):

$$\frac{1}{\gamma}\mathrm{diag}(I)^{-1}I' = \frac{(1+\varrho p)}{(1+p)\gamma}\mathrm{diag}(I)^{-1}\mathrm{diag}(S)\frac{\beta}{c}I - \mathbb{1} \tag{22}$$

Observe that $I' \lessgtr 0$ if and only if the local *instantaneous reproduction numbers*

$$\mathscr{R}_i(t) := \frac{(1+\varrho p)}{(1+p)\gamma}\frac{S_i(t)}{I_i(t)}\sum_j \frac{\beta_{ij}}{c}I_j(t) \lessgtr 1 \tag{23}$$

for all i. Practically speaking, if the active cases I determines the endemicity levels, $\mathscr{R}_i(t)$ speaks about the epidemics progression. The fact that $(1+\varrho p)\cdot \mathrm{diag}(S)\frac{\beta}{c}I/(1+p)$ gives the inflow of new cases in all clusters, $\mathscr{R}_i(t)I_i(t)$ gives the expected number of new cases from $S_i(t)$ at the timestamp t attributed to the entire infected individuals from $I_j(t)$ $(j = 1, \cdots, 4)$ per viral shedding period $1/\gamma$. Therefore, $\mathscr{R}_i(t)$ represents the expected number of new cases from $S_i(t)$ attributed to the normalized infected individuals $I_j(t)/I_i(t)$ $(j = 1, \cdots, 4)$ per viral shedding period $1/\gamma$. For a realistic approximation, we took the inflow of new cases from the data while the active cases, which serve as the denominators, will be taken from the fitting. Such a definition of instantaneous reproduction number has been used by Fraser in [66], except where the active cases (as the denominator) were taken from weighted new cases in the past n days for a fixed n. The weights were later known as *serial intervals* [67,68], estimating the distribution of delays taken from the onset of symptoms until hospital admission (i.e., when the data of new cases are usually recorded). Under the two facts that (1) the instantaneous reproduction number is, by the definition, too fluctuating and (2) infected individuals can already infect susceptible individuals from the onset of symptoms, Fraser also introduced some moving average such that the 'real' new cases at a certain timestamp should come from the new cases 'recorded by hospitals' in the future timestamps (up to n) weighted by the serial intervals, while the active cases come from the sum of those corresponding to the used timestamps, where again, at each timestamp the active cases are weighted sum of new cases in the past n days. Inspired by such refinement with, however, lack of serial interval data, ours becomes

$$\mathscr{R}_i(t) \approx \frac{(1+\varrho p)}{(1+p)\gamma}\cdot\frac{\frac{1}{\tau}\int_t^{t+\tau} S_i(s)\sum_j \frac{\beta_{ij}}{c}I_j(s)\,\mathrm{d}s}{\frac{1}{\tau}\int_t^{t+\tau} I_i(s)\,\mathrm{d}s} \tag{24}$$

for some averaging window size τ. The forward moving average thus allows the serial intervals to be of uniform distribution around τ days. In the numerical computations, we designate $\tau = 7$.

4 Data assimilation

The basic aim of parameter estimation is to find agreement between model solution for weekly new cases $C(t_k) := \frac{(1+\varrho p)}{(1+p)}\mathrm{diag}(S(t_k))\frac{\beta}{c}I(t_k)$ at all time points t_k with known data C_k^d subject to identifiability of unknown parameters $\theta = (\varrho, p, \eta, \beta, S_0, I_0)$. We assume that the fitting would be subject to time-invariant i.i.d. error of the weekly covariance Σ (for all the four clusters) and the prior was set to be Gaussian. The latter means that the parameter estimation will be based on minimizing the Mahalanobis distance between the model solution and empirical data. For simplicity, no correlation was imposed for the cluster-wise error such that $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_4^2)$. The non-degenerate joint likelihood function for one time point t_k is then given by

$$\mathcal{L}_k(\theta) := \frac{\exp\left[-\frac{1}{2}\left(C(t_k)-C_k^d\right)^\top \Sigma^{-1}\left(C(t_k)-C_k^d\right)\right]}{\sqrt{(2\pi)^4\det\Sigma}}.$$

Assuming timely i.i.d. measurement, the joint likelihood for the entire observations is then given by

$$\begin{aligned}\mathcal{L}(\theta) &:= K_G\prod_k \mathcal{L}_k(\theta)\\ &= \exp\left[-\frac{1}{2}\sum_k \left(C(t_k)-C_k^d\right)^\top \Sigma^{-1}\left(C(t_k)-C_k^d\right)\right]\end{aligned}$$

by taking $K_G = (2\pi)^{2|k|}(\det\Sigma)^{|k|/2}$ that serves to simplify the representation of the likelihood function [72]. Our study designates the variance terms as the mean of the data throughout the observations $(\sigma_1, \cdots, \sigma_4) = (1/|k|)\sum_k C_k^d$ so as to avoid a blow-up in the likelihood function.

As the parameter dimension is much smaller than the data size, the standard asymptotic confidence interval [73] has been suggested to delineate the parameter uncertainty [74,75]. The formula of the confidence

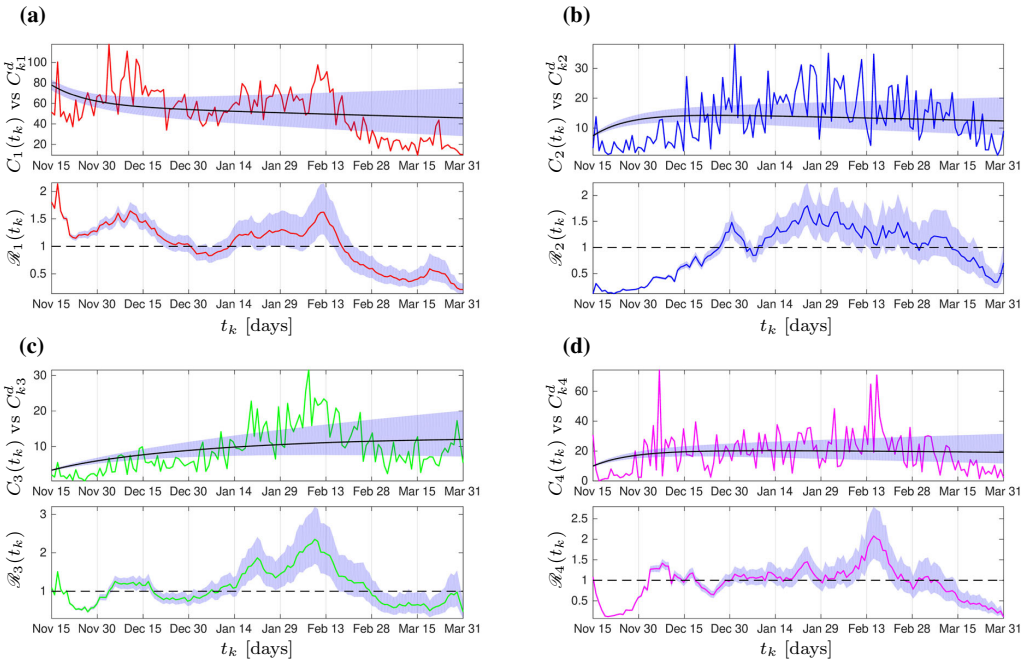


Fig. 3 Fitting results from the deterministic SI model. The vertical axes give the numbers of weekly new cases per 1,000,000 inhabitants and effective local reproduction numbers from all clusters. The fluctuating curves represent the data and the shaded region around the fitted model solution determine the variation of ϱ, p, η from their confidence interval

interval for each optimal parameter $\hat{\theta}_\ell$ takes the form

$$\left[\hat{\theta}_\ell - \varepsilon_\ell, \hat{\theta}_\ell + \varepsilon_\ell\right]$$
$$\text{where} \quad \varepsilon_\ell = \sqrt{2\chi^2(\alpha, df) \cdot \left(\nabla^{-2} \log \mathcal{L}(\hat{\theta})\right)_{\ell\ell}}. \tag{25}$$

The operator ∇^{-2} denotes the inverse of the Hessian, while the notation $\chi^2(\alpha, df)$ denotes the α quantile of the χ^2 distribution with the degree of freedom df. The degree of freedom can be chosen between two that further determines the type of confidence interval: $df = 1$ gives *pointwise asymptotic confidence interval* (PACI) that works on the individual parameter, $df =$ #parameters gives a *simultaneous asymptotic confidence interval* (SACI) that works jointly for all the parameters.

In the present study, the Hessian matrix in (25) will be approximated up to the second order using the queen-type stencil. Due to disparate scales of the parameters, the step size will be made dependent on the parameter's order of magnitude, i.e., $\Delta\theta_\ell := \delta\theta_\ell$ for a uniformly small δ. Our study uses $\delta \approx$ 1e–08. After all, the fitting will be done in MATLAB using the toolbox `fmincon` accompanied by `interior-point` as the core optimization solver. The fitting result together with the effective local reproduction numbers can be seen in Fig. 3. Meanwhile, we keep β, S_0, I_0 at the fitted values, we vary ϱ, p, η from their confidence interval to have a shaded region around the fitting curves. Due to model simplicity (no time-dependent parameters), we can only expect to see an almost stationary model solution to fit the almost variance-stationary dataset, also subject to the constraint on I_0 of the four clusters: $I_{10} \geq I_{40} \geq I_{20} \geq I_{30}$. The fitted parameter values can be seen in Table 1.

Table 1 Parameters of the SI model (26). All zero β-values were due to rounding numbers smaller than 1e–07. This is intentional against floating-point error in the numerical continuation, while at the same time, almost no visible difference in the model response in comparison to that using positive values was observed

Parameter	Description	Unit	Range	Value (ε PACI)	Ref.
Known:					
$1/\tilde{\mu}$	Average human lifespan	[d]	$70 - 80 \cdot 365$	$76.9 \cdot 365$	[69]
$1/\tilde{\gamma}$	Viral shedding period	[d]	$1 - 30$	20	[70]
γ	$\tilde{\mu} + \tilde{\gamma}$	[d^{-1}]			
c	Scaling factor	–		10^6	
Optimized:					
ϱ	Transmission scale from undetected to susceptible	–	$1 - 20$	11.6373 (0.6727)	
p	Case detection ratio	–	$0 - 35/65$	0.4698 (0.0148)	[71]
η	Rescaled loss-of-immunity rate	[d^{-1}]	$1/3 \cdot 365$	4.5666e–04 (2.0482e–05)	
β_{11}	Infection rate between S_1 and I_1	[d^{-1}]	$0 - 3/c$	0.0040 (2.4694e–05)	
β_{12}	Infection rate between S_1 and I_2	[d^{-1}]	$0 - 3/c$	0.0308 (5.7954e–04)	
β_{13}	Infection rate between S_1 and I_3	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{14}	Infection rate between S_1 and I_4	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{21}	Infection rate between S_2 and I_1	[d^{-1}]	$0 - 3/c$	0.0032 (1.3493e–04)	
β_{22}	Infection rate between S_2 and I_2	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{23}	Infection rate between S_2 and I_3	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{24}	infection rate between S_2 and I_4	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{31}	Infection rate between S_3 and I_1	[d^{-1}]	$0 - 3/c$	0.0012 (4.3181e–05)	
β_{32}	Infection rate between S_3 and I_2	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{33}	Infection rate between S_3 and I_3	[d^{-1}]	$0 - 3/c$	0.0126 (3.0470e–04)	
β_{34}	Infection rate between S_3 and I_4	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{41}	Infection rate between S_4 and I_1	[d^{-1}]	$0 - 3/c$	0.0046 (5.8647e–05)	
β_{42}	Infection rate between S_4 and I_2	[d^{-1}]	$0 - 3/c$	0 (0)	
β_{43}	Infection rate between S_4 and I_3	[d^{-1}]	$0 - 3/c$	0.0039 (4.4670e–05)	
β_{44}	Infection rate between S_4 and I_4	[d^{-1}]	$0 - 3/c$	0 (0)	
S_{10}	Initial condition for S_1	–	$0.01c - c$	8.9227e+05 (2.3154e+04)	
S_{20}	Initial condition for S_2	–	$0.01c - c$	9.3000e+05 (4.7650e+04)	
S_{30}	Initial condition for S_3	–	$0.01c - c$	6.5110e+05 (1.6834e+04)	
S_{40}	Initial condition for S_4	–	$0.01c - c$	8.1692e+05 (3.7530e+04)	
I_{10}	Initial condition for I_1	–	$0 - 0.1c$	569.9877 (18.4876)	
I_{20}	Initial condition for I_2	–	$0 - 0.1c$	569.9791 (10.1970)	
I_{30}	Initial condition for I_3	–	$0 - 0.1c$	37.4413 (0.8390)	
I_{40}	Initial condition for I_4	–	$0 - 0.1c$	569.9829 (17.9763)	
Adjusted:					
μ	$\tilde{\mu}(1 + p)$	[d^{-1}]		5.2365e–05	
T_{Ref}	Reference time for transient analysis	[d]		136	
p_{Ref}	Reference value for average policy effect	–		p	
ω_{Ref}	Reference value for average policy effect	–		1	

5 Numerical study of the COVID-19 model via path-following techniques

In this section, our main goal is to investigate the dynamical response of the model as certain selected parameters are varied. To evaluate the impact of reassessment on government policy against COVID-19 posterior to fitting, we shall introduce one more control parameter ω that hereafter is referred to as the *contact restriction factor*. This parameter will serve to decrease the intra- and inter-cluster contacts as so far portrayed by the fitted values of β_{ij}. From the application point of view, this parameter can be realized by enhancing NPIs and all possible interventions that likely reduce the contact between susceptible and infected persons. Furthermore, the parameter p (case detection ratio) will be interpreted as a factor determining the quality of COVID-19 testing campaigns in such a way that p close to zero represents an effective testing policy, while a large p indicates that a great number of infections are not detected and therefore are able to spread the disease at higher infections rates (according to the factor ϱ in the SI model (26)). Consequently, the reassessment yields a small modification in the model as

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= \mu I - (1+\varrho p)\omega \mathrm{diag}(S)\frac{\beta}{c} I \\ &\quad + \eta[c\mathbb{1} - S - (1+p)I], \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{(1+\varrho p)\omega}{1+p}\mathrm{diag}(S)\frac{\beta}{c} I - \gamma I.\end{aligned} \tag{26}$$

The numerical investigation will be based on the parameter fitting obtained in the previous section. There, the pair (S_i, I_i) represents the susceptible and infected population in the cluster i. In this way, our study will focus on the effect of the main disease control parameters $(p, \omega) \in (0, 1]^2$ on the model behavior including the basic reproduction number

$$\mathscr{R}_0 = \frac{(1+\varrho p)\omega}{(1+p)\gamma}\rho(\beta), \tag{27}$$

in such a way that a fixed combination of (p, ω) will be interpreted as a specific disease control policy determined by the decision makers. The numerical study will be carried out using the path-following software COCO (Computational Continuation Core [76]). This is an analysis and development platform for the numerical treatment of continuation problems using MATLAB. A remarkable feature of COCO is its set of toolboxes that covers, to a large extent, the functionality of available continuation packages, such as AUTO [77] and MATCONT [78]. In particular, we will make extensive use of the COCO-toolbox `ep`, which encompasses a set of numerical routines for the bifurcation analysis of parameter-dependent families of equilibria in smooth dynamical systems.

5.1 Monitor and cost functions

In this investigation, one of the main goals is to study the effectiveness of the disease control policy to reduce the number of COVID-19 cases in the proposed biological scenario, and for this purpose suitable performance measures will be considered in our numerical implementation. Let us assume that

$$S_{\mathrm{Ref}}(t) \quad \text{and} \quad I_{\mathrm{Ref}}(t), \quad 0 \le t \le T_{\mathrm{Ref}},$$

is a bounded reference solution of system (26) computed for the parameter values and initial conditions given in Table 1, with $T_{\mathrm{Ref}} > 0$ being a reference final time and $\omega = 1$. In this setting, we define the performance measure given by

$$\begin{aligned}M_{\mathrm{APE}}(p,\omega) &:= \\ &\frac{1}{T_{\mathrm{Ref}}}\int_0^{T_{\mathrm{Ref}}} \left\| \frac{\omega_{\mathrm{Ref}}(1+\varrho p_{\mathrm{Ref}})}{c(1+p_{\mathrm{Ref}})}\mathrm{diag}\left(S_{\mathrm{Ref}}(t)\right)\beta I_{\mathrm{Ref}}(t)\right\|_1 \mathrm{d}t \\ &- \frac{1}{T_{\mathrm{Ref}}}\int_0^{T_{\mathrm{Ref}}} \left\| \frac{\omega(1+\varrho p)}{c(1+p)}\mathrm{diag}(S(t))\beta I(t)\right\|_1 \mathrm{d}t,\end{aligned} \tag{28}$$

where $\omega_{\mathrm{Ref}} = 1$ and p_{Ref} is the p-value given in Table 1. In the above expression, $S(t)$, $I(t)$, $0 \le t \le T_{\mathrm{Ref}}$, stand for a solution of system (26) computed for the parameter values and initial conditions given in Table 1, but for arbitrary (p, ω). From a practical point of view, the quantity $M_{\mathrm{APE}}(p, \omega)$ (hereafter referred to as the *average policy effect*) represents the average COVID-19 cases that could have been free from infection on a daily basis by applying a particular disease control policy (p, ω), in comparison to the reference solution case $(p_{\mathrm{Ref}}, \omega_{\mathrm{Ref}})$ defined above. In connection to this definition, we introduce the associated *policy cost* given by

$$M_{\mathrm{Cost}}(p,\omega) := \lambda\frac{\omega_{\mathrm{Ref}} - \omega}{\omega} + (1-\lambda)\frac{p_{\mathrm{Ref}} - p}{p}, \tag{29}$$

where $0 \le \lambda \le 1$ is a coefficient that characterizes the cost distribution among contact restrictions and testing campaigns. As can be seen from (29), a strict mobility

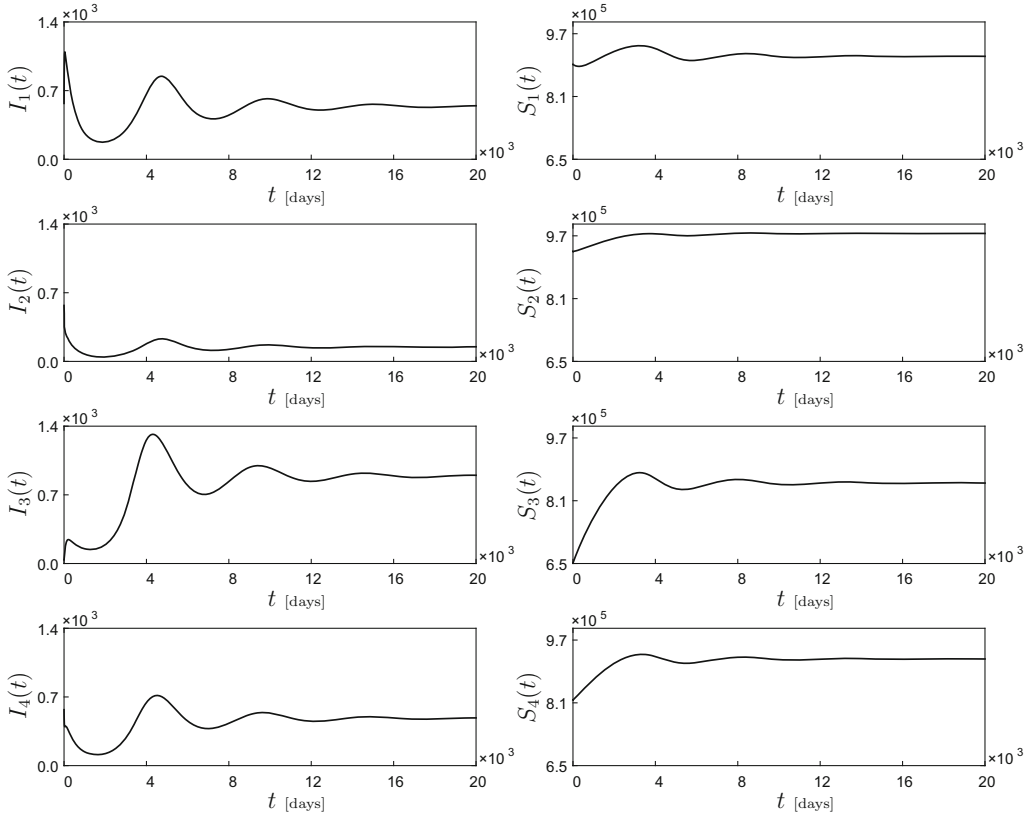


Fig. 4 Dynamical response of the COVID-19 model (26), computed for the parameter values and initial conditions given in Table 1. The picture shows time series for the infected ($I_i(t)$) and susceptible population ($S_i(t)$)

reduction ($\omega \approx 0$) implies a high policy cost, representing the bad impact on the economy and other negative effects associated with the mobility reduction. Similarly, a widely spread and effective COVID-19 testing campaign ($p \approx 0$) also produces very high costs, due to the personals required for implementation, expenditure on test kits and other logistics, media advertisement, organization, etc. In our investigation, the value $\lambda = 0.7$ will be assigned, which portrays a realistic distribution between the two cost terms in (29) according to our numerical simulations. Nevertheless, we give such a higher contribution from contact restrictions based on bad economic impact in Sri Lanka due to job and earning losses associated with mobility restriction and crowd clearance, which additionally force the government to spend much on welfare activities targeting low-income citizens [79]. Therefore, the cost function given in (29) takes not only the view of government spending but also the economic recession in the whole country into account.

5.2 Numerical investigation of the modified COVID-19 model

With the mathematical framework introduced in the earlier section, we can now move on to the numerical study of the modified COVID-19 model (26) using parameter values and initial conditions given in Table 1. Observe that the contact matrix β is no longer irreducible. As a result, the initial direction of the continuum of endemic equilibria ψ_1 as in (20) is only guaran-

teed to be nonnegative according to Perron–Frobenius Theorem. A preliminary system response can be seen in Fig. 4. The picture shows time series for the active cases $I_i(t)$ and susceptible population $S_i(t)$, corresponding to Moran's clusters Qi where $i = 1, 2, 3, 4$. As can be seen in the figure, for the selected parameter values the system shows a damped oscillatory behavior that settles down after a long time to an endemic equilibrium, i.e., a steady state where the COVID-19 infection is present in all clusters. This equilibrium state will then be used as starting point for our numerical investigation based on path-following techniques.

Let us begin our study with the numerical continuation of the endemic equilibrium found above with respect to the mobility restriction factor ω. The result of this process can be observed in Fig. 5, panels (c) and (e). Specifically, panels (c) and (e) present the behavior of I_1 (left vertical axis, in blue), I_3 (right vertical axis, in red) and I_2 (left vertical axis, in blue), I_4 (right vertical axis, in red), respectively, as the parameter ω varies. Panel (a) shows the dependency on ω of the basic reproduction number $\mathscr{R}_0$, given by formula (27). In this diagram, it can be seen that for low values of ω, the basic reproduction number is smaller than one, due to which the system presents a stable disease-free equilibrium corresponding to the solid horizontal branches shown in Fig. 5c and (e). As ω increases, $\mathscr{R}_0$ increases as well, and it crosses 1 from below at $\omega \approx 0.90535$, where a branching point BP1 is detected. Here, the disease-free equilibrium loses stability and an endemic branch is born (via a forward bifurcation). Interestingly, at this point a COVID-19 outbreak occurs only for clusters Q3 and Q4, while clusters Q1 and Q2 remain disease-free. If ω increases further, however, the disease for clusters Q1 and Q2 develops for $\omega \approx 0.93739$, where a branching point BP2 is found. From this point onward, the disease is present in all clusters, and the increment of the infected cases augments more rapidly as the mobility restriction factor grows.

A similar scenario is encountered when the case detection ratio p is considered as the bifurcation parameter, see Fig. 5b, d and f. A first branching point (from below) is found for $p \approx 0.38920$ (BP3), where a COVID-19 outbreak takes place, but only for clusters Q3 and Q4, as before. A full disease development is encountered at $p \approx 0.41585$ (BP4), where now clusters Q1 and Q2 show COVID-19 infection. This scenario is clearly depicted in Fig. 5d and f showing high infections for higher p (i.e., for inefficient testing campaigns). Cluster-oriented interpretation can be distinguished by locally targeted testing campaigns (higher p) and widespread random testing campaigns (lower p). Our model thus conjectures that it takes smaller reduction of (p, ω) from $(p_{\text{Ref}}, \omega_{\text{Ref}})$ in order to clean up the active cases in Q1 and Q2 than in Q3 and Q4.

As can be seen from the numerical study discussed above, both the mobility restriction factor ω and the case detection ratio p play a crucial role in controlling the disease. For instance, Fig. 5c reveals that the branching point BP1 is responsible for a first COVID-19 outbreak, occurring in clusters Q3 and Q4. Therefore, our next concern will be to investigate how this critical point varies in the p-ω control space. For this purpose, we will carry out a two-parameter continuation of this critical point, see Fig. 6a. The black curve represents a locus of branching points on the p-ω plane. The resulting curve divides the control space into two regions: one for stable disease-free equilibria (yellow) and one corresponding to stable endemic equilibria (blue). In this way, for a specific disease control policy represented by the pair (p, ω), we can determine *a priori* whether the policy will be effective or not in controlling a COVID-19 outbreak. This can be verified at the test points P1–P4 shown in Fig. 6a. For all these points, test trajectories are calculated using the data shown in Table 1, see Fig. 6b. As can be seen, the solutions computed at P1 and P3 (disease-free region, in yellow) decay to zero, while those computed for P2 and P4 (endemic region, in blue) settle down to an endemic equilibrium, where a long-term COVID-19 outbreak occurs.

5.3 Optimization of the disease control policies

In the previous section, we applied numerical continuation methods to study the effect of the mobility restriction factor ω and the case detection ratio p on the behavior of the modified COVID-19 model (26). In this way, we established critical values of the control parameters upon which a disease outbreak occurs. In this section, we will consider the effect of the control parameters on the average policy effect and the policy cost, as defined in Sec. 5.1. For this purpose, we will assume that the disease control policies represented by the pair (p, ω) are chosen from the yellow region in Fig. 6a.

To begin our study, we will carry out the numerical continuation of disease-free equilibria of model (26) with respect to ω and monitor the behavior of the

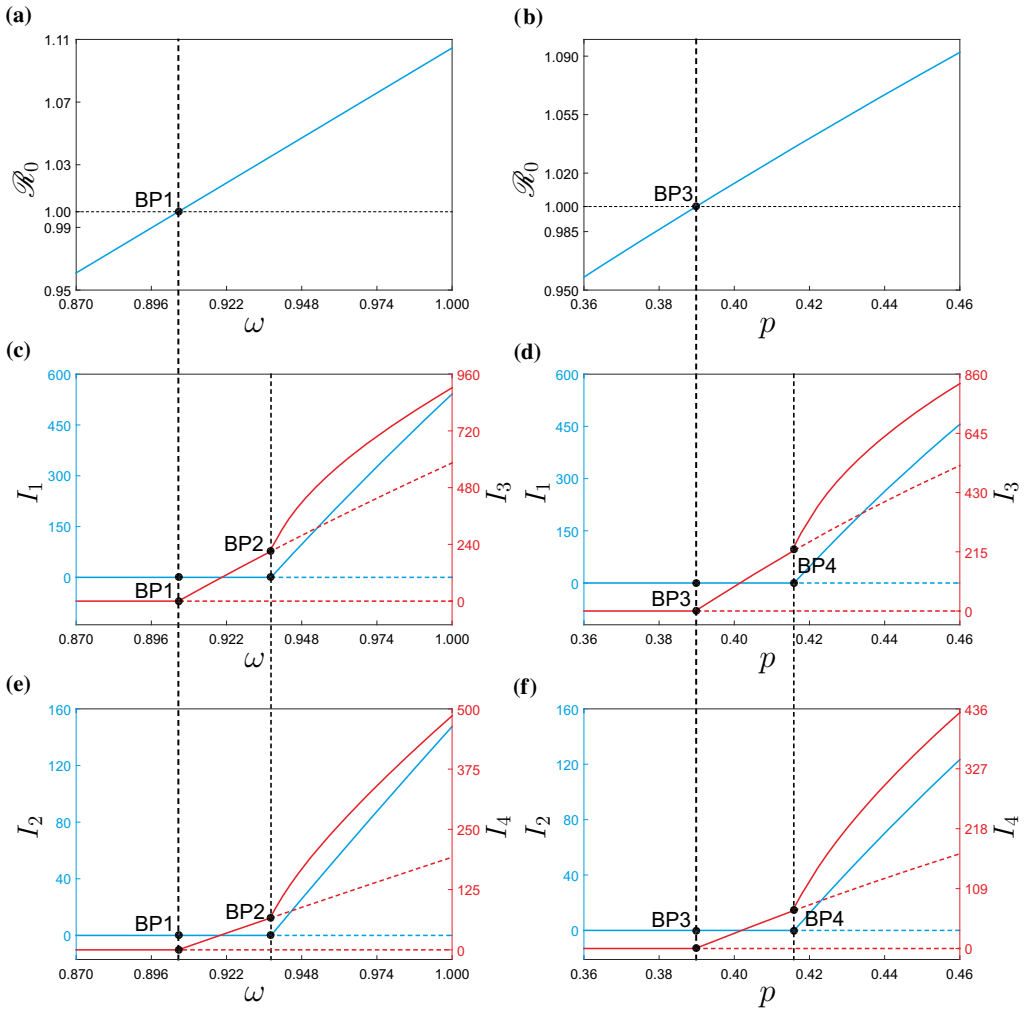


Fig. 5 One-parameter continuation of equilibria of system (26) with respect to the mobility restriction factor ω and the case detection ratio p, computed for the parameter values given in Table 1. Panels **a** and **b** depict the behavior of the basic reproduction number $\mathscr{R}_0$ given by formula (8). Panels **c** and **d** present the behavior of I_1 (left vertical axis, in blue) and I_3 (right vertical axis, in red). Similarly, panels **e** and **f** plot the variation of I_2 (left vertical axis, in blue) and I_4 (right vertical axis, in red) with respect to the corresponding parameters. In these diagrams, solid and dashed lines stand for branches of stable and unstable equilibria, respectively. During the computations, a series of branching points are detected for $\omega \approx 0.90535$ (BP1), $\omega \approx 0.93739$ (BP2) (depicted in panels (**a**), (**c**) and (**e**)) and $p \approx 0.38920$ (BP3), $p \approx 0.41585$ (BP4) (depicted in panels (**b**), (**d**) and (**f**))

average policy effect M_{APE} defined in (28). The result of this procedure can be seen in Fig. 7a. As can be expected, M_{APE} is a decreasing function of ω, since the average COVID-19 infections that can be avoided (as explained in Sect. 5.1) decrease if higher degrees of mobility between clusters are allowed. Moreover, panel (a) shows a series of points labeled P_L, which correspond to ω-values yielding $M_{\text{APE}} = L$. At these points, the resulting costs are shown in Fig. 7c, which depicts the behavior of the cost function M_{Cost} (see (29)) with

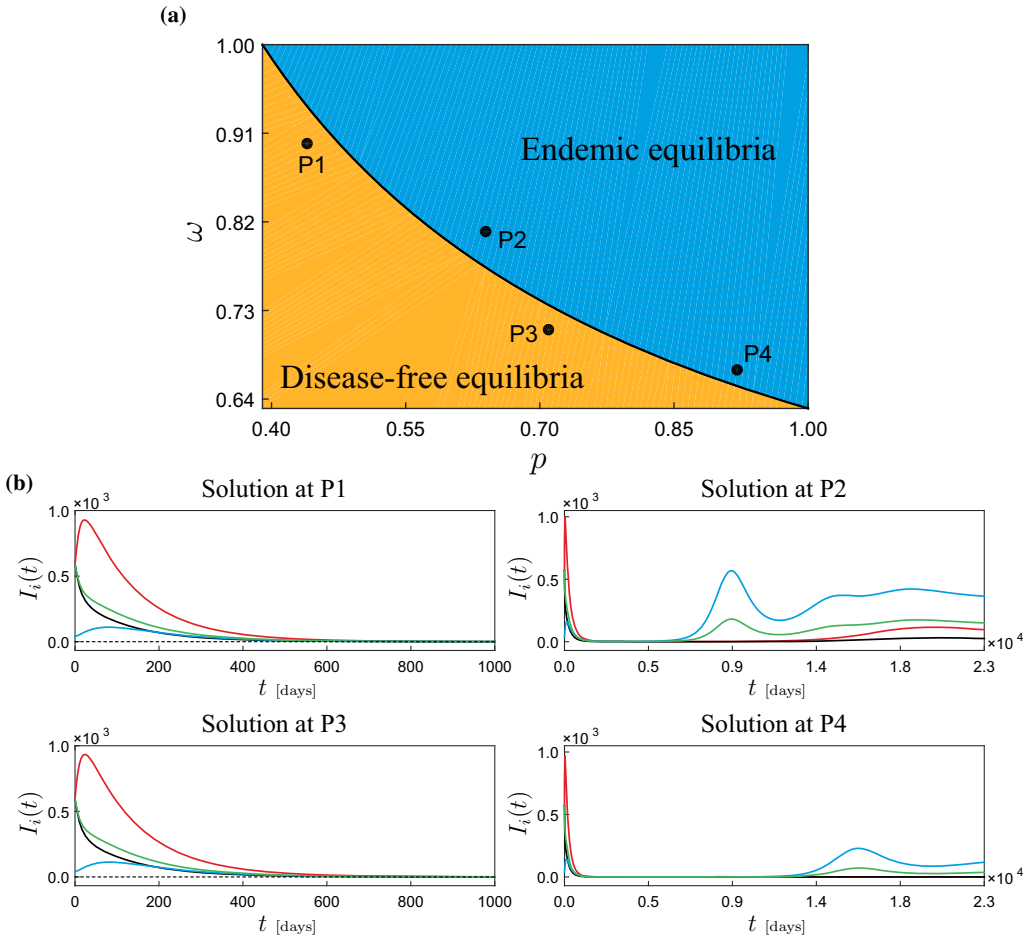


Fig. 6 **a** Two-parameter continuation of the branching point BP1 found in Fig. 5a with respect to p and ω. The resulting curve divides the parameter space into two regions: one for stable disease-free equilibria (yellow) and one corresponding to stable endemic equilibria (blue). **b** System responses obtained at the test points P1 ($p = 0.44$, $\omega = 0.90$), P2 ($p = 0.64$, $\omega = 0.81$), P3 ($p = 0.71$, $\omega = 0.71$) and P4 ($p = 0.92$, $\omega = 0.67$). The time plots present the behavior of I_1 (red), I_2 (black), I_3 (blue) and I_4 (green). All numerical simulations are calculated with the initial conditions specified in Table 1

respect to ω. As can be seen in the diagram, this function grows as ω decreases, which is consistent with the fact that stricter contact restrictions lead to higher policy costs. This observation then raises the question if for a desired fixed value of M_{APE}, a more convenient control policy (p, ω) can be found in terms of cost reduction. To tackle this question, we will employ two-parameter continuation with respect to p and ω to find loci of control points (p, ω) yielding fixed values of M_{APE}, monitoring the corresponding cost function. The result can be seen in Fig. 7b. Here, a family of curves in the p-ω plane are shown for which M_{APE} is kept fixed. Panel (d) presents the behavior of the cost function M_{Cost} along the curves obtained in panel (b). As can be seen, in all cases the cost function presents local minima, which can be interpreted as an optimal policy implementation for a desired fixed value of M_{APE}.

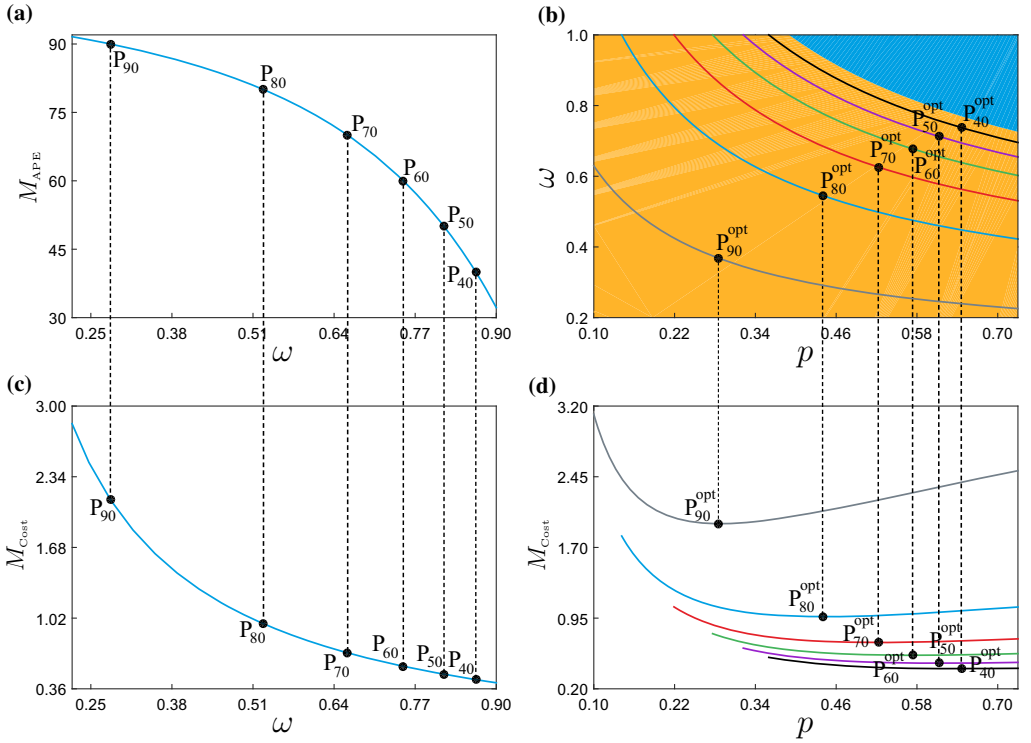


Fig. 7 **a** One-parameter continuation of equilibria as in Fig. 5c, showing the behavior of the average policy effect M_{APE} and the policy cost M_{Cost} (panel **c**) with respect to ω. The points P_L correspond to ω-values yielding $M_{\mathrm{APE}} = L$. These are found at $\omega \approx 0.28137$ (P_{90}), $\omega \approx 0.52689$ (P_{80}), $\omega \approx 0.66182$ (P_{70}), $\omega \approx 0.75098$ (P_{60}), $\omega \approx 0.81627$ (P_{50}) and $\omega \approx 0.86723$ (P_{40}). **b** Two-parameter continuation of equilibria of system (26) with respect to p and ω, keeping the average policy effect M_{APE} constant (at the values specified above). In this panel, the yellow and blue regions are the same as in Fig. 6a. Panel **d** shows the behavior of the policy cost M_{Cost} computed along the curves obtained in panel (**b**), for fixed $M_{\mathrm{APE}} = 90$ (grey curve), $M_{\mathrm{APE}} = 80$ (blue curve), $M_{\mathrm{APE}} = 70$ (red curve), $M_{\mathrm{APE}} = 60$ (green curve), $M_{\mathrm{APE}} = 50$ (purple curve) and $M_{\mathrm{APE}} = 40$ (black curve). In this panel, the cost for fixed M_{APE} attains a minimum at $(p, \omega) \approx (0.28511, 0.36847)$ ($\mathrm{P}_{90}^{\mathrm{opt}}$), $(p, \omega) \approx (0.43984, 0.54558)$ ($\mathrm{P}_{80}^{\mathrm{opt}}$), $(p, \omega) \approx (0.52256, 0.62619)$ ($\mathrm{P}_{70}^{\mathrm{opt}}$), $(p, \omega) \approx (0.57365, 0.67755)$ ($\mathrm{P}_{60}^{\mathrm{opt}}$), $(p, \omega) \approx (0.61264, 0.71263)$ ($\mathrm{P}_{50}^{\mathrm{opt}}$) and $(p, \omega) \approx (0.64590, 0.73768)$ ($\mathrm{P}_{40}^{\mathrm{opt}}$)

6 Concluding remarks

Analysis in this study covers spatio-temporal aspects of COVID-19 transmission in Sri Lanka with the aid of basic reproduction number and path-following continuation pertained to the role of NPIs (contact restrictions and testing campaigns). The daily new cases have been widely used data; however, we processed normalized cases via the populations of RDHS divisions. It suppresses unbiased estimates as higher numbers of cases are reported in highly populated RDHS divisions. Subsequently, all RDHS divisions were categorized using Moran's scatter into four clusters. Prioritization as well as route for interventions should be Q1 (high-high), Q4 (high-low), Q2 (low-high), and Q3 (low-low). One useful contribution is that the government can use such a route in vaccination programs started at the latter stage of the study period. Priority within a cluster may rely on logistics available within that cluster and temporary shift from the other clusters. Our result is also helpful in placing appropriate border controls for the sake of curtailing transmission waves. Even though Q1 and Q3 do not encounter different incidence levels in their spatial lags, Q2 is vulnerable for significant absorption

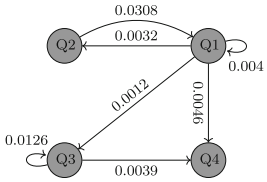


Fig. 8 Network based on the contact matrix β. The arrow directed from the cluster Qi to the cluster Qj translates the statement "the susceptible humans in Qj contract infection through contact with infected humans in Qi" or shortly "Qi causes infection in Qj"

while Q4 is responsible for significant diffusion. Therefore, border controls can be placed in every important intersecting point between two different clusters.

We extend the qualitative clustering analysis into a quantitative one by conducting an inverse problem using the cases data. A preliminary model of SIURSD type is proposed, carrying the metapopulation context with memory. Due to non-observable model variables, dimensional reduction leads us to an SI type. This final model may look parsimonious; however, it still explains essential mechanistic processes of COVID-19 transmission: cluster-wise contact matrix, viral shedding period, transmission scaling between detected and undetected cases, case detection ratio, contact restriction factor, and loss of immunity. Fitting to the data was done to reveal hidden dynamics including contract matrix, initial conditions for the active cases, case detection ratio, transmission scaling, and loss of immunity. Nonetheless, the SI type may provide beneficiary to big data analytics, especially when the observation period and network size are extended. Forward bifurcation for strongly connected network among clusters was found around the basic reproduction number 1. Numerical investigation was done for the case where the network is, according to rounding small β-values to zero, not strongly connected. Time-varying effective local reproduction numbers for all clusters are also presented. Their appearance supersedes clueless cases data when it comes to localizing time at which the current transmission is high (reproduction number greater than 1), suggesting for immediate interventions.

An interesting result is evident from one-parameter continuation of equilibria. Recalling the analytical framework in Sec. 3.3, the initial direction of the continuum of endemic equilibria at $\mathscr{R}_0 = 1$ is the Perron vector of the next generation matrix ψ_1 (see Eqs. (20) and (21)). As the network associated with the next generation matrix or the contact matrix β is not strongly connected (see Fig. 8), Perron–Frobenius Theorem (cf. [80]) only guarantees the nonnegativity of the Perron vector. Particularly to our case, we obtain

$$\psi_1 \approx \begin{pmatrix} 0 \\ 0 \\ 0.9553 \\ 0.2957 \end{pmatrix}.$$

This Perron vector indicates two findings: (1) the clusters Q1 and Q2 remain in the disease-free states when $\mathscr{R}_0$ shortly exceeds 1; meanwhile (2) the long-term number of active cases in the cluster Q3 jumps to larger extent than that in the cluster Q4 as observed in Fig. 5. If we read the bifurcation diagrams backward in ω and p, then these findings mean that Q1 and Q2 achieve disease-free states quicker than Q3 and Q4 under the reduction of p and ω from $p_{\text{Ref}} \approx 0.4698$ and $\omega_{\text{Ref}} = 1$, respectively. The network in Fig. 8 explains that Q2 receives a relatively small "injection" from Q1 but returns with a large injection to Q1; meanwhile there is no essential self-injection in Q2. Equipped with a small self-injection, Q1 also injects Q3 and Q4 at comparable rates. Meanwhile, Q3 admits a relatively large self-injection but spares an injection to Q4. On the overall picture, it is arguable that Q1 and Q2 lose endemicity faster than Q3 and Q4 if the entire injection rates (the nonzero entries of the contact matrix β) are reduced simultaneously. At a certain stage, there comes, on the one hand, a scenario where the self-injection in Q1 and thus the injection to Q2 are negligible, making Q2 non-reproductive. On the other hand, the negligible injection from Q1 is compensated by the self-injection in Q3 that withstand both Q3 and Q4 in the endemic states. Notwithstanding this interesting finding, we also observe that disease-free equilibrium (DFE) can be found by reducing p and ω not so far away from p_{Ref} and ω_{Ref}, respectively. Thus, we argue that the original interventions imposed by the government had been satisfactory during the observation period. From the point of view model transients, one should note that significant contact restrictions are both costly and not gainful in terms of average policy effect (APE). This is evident by concave behavior of APE and convex behavior of the cost against ω (see Fig. 7(a,c)). Therefore, reducing p and ω to arbitrarily small values does not make much sense. Scenarios for the optimal values of p and ω minimizing the cost under fixed magnitudes of

APE were proposed. As expected, even optimal results come with a price, as the optimal (p, ω)-values walk toward the third quadrant by increasing APE values; see Fig. 7(b).

Finally, this study leaves us some gaps for further improvement. First, several attributes in the original model can be modified to capture more complexities. For example, the average viral shedding period $1/\tilde{\gamma}$ for the undetected cases could have been different from that of the detected cases due to nonoccurrence of symptoms. Despite the averaging, taking the timely proportion of detected cases α to be a constant can be too stringent owing to the unknown dark figures (undetected cases). Future improvement may include time-dependent noise for such parameters with given (under guidance of field experts) or computationally tested priors. That recovered and deceased cases preserve a constant ratio is also worth of improvement. Second, the control parameters ω and p actually represent adjustment of contact restrictions and testing campaigns on the national level, meaning that the scaled susceptible cases in all clusters are enforced the same way toward endemicity reduction, irrespective of their local resources. Meanwhile, the reduction of β-values via the cluster-independent ω also serves as another limitation of the model. From the application point of view, this means that all actions entailed in the contact restrictions should simultaneously follow the adjustment of ω without proper consideration as to what actions are paramount among the others. For example, reduction of ω from $\omega_{\text{Ref}} = 1$ to 0.75 means that those who go out for activities should reduce the intensities to 75%, those who travel across clusters 4 days a week should change to 3 days a week, schools that are opened 4 days a week should be opened 3 days a week instead, working in the office 5 days a week should change to 3.75 equivalent working days, etc. Technically speaking, these changes may sound simple from the standpoint of the decision maker; however, different abilities and preferences among humans can make the implementation difficult to trace. Third, our SI model contains several parameters that multiply with others. A parameter identification analysis is worth considering if one were to reveal possible dependency among them and thus correct the model specification. Fourth, had regional data of confounding factors been there, we could have integrated these data e.g., in the β-values from time to time to capture the different cluster peaks and fluctuations. This is due to the fact that the β-values appear to be equivalent to the term of new cases. Incidence and meteorological data from other countries, for example, could be helpful toward this direction.

Acknowledgements The first author is thankful to the Epidemiology Unit, Medical Statistics Unit, Health Information Unit, Health Promotion Bureau of the Ministry of Health—Sri Lanka, for providing data and assistance; National Science Foundation of Sri Lanka for facilitating collaborations with the above units. The fourth author acknowledges the financial support from the Ministry of Education, Culture, Research, and Technology of Indonesia, 2021. The authors express their sincere gratitude to the anonymous reviewers who have provided suggestions and addressed challenging questions that significantly improve the reading of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data are available upon request, which can be directed to the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Worldometer, Coronavirus worldwide graphs. https://www.worldometers.info/coronavirus/worldwide-graphs/, 2021. Accessed: 15.05.2021
2. Hodgson, S.H., Mansatta, K., Mallett, G., Harris, V., Emary, K.R.W., Pollard, A.J.: What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. Lancet Infect. Dis. **21**(2), E26–E35 (2021)
3. Williams, T.C., Burgers, W.A.: SARS-CoV-2 evolution and vaccines: cause for concern? Lancet Respir. Med. **9**(4), 333–335 (2021)
4. World Health Organization, WHO target product profiles for COVID-19 vaccines. https://www.who.int/publications/m/item/who-target-product-profiles-for-covid-19-vaccines, 2021. Accessed: 25.02.2021

5. Fontanet, A., Autran, B., Lina, B., Kieny, M.P., Karim, S.S.A., Sridhar, D.: SARS-CoV-2 variants and ending the COVID-19 pandemic. Lancet **397**(10278), 952–954 (2021)
6. Cohen,J.: South Africa suspends use of AstraZeneca's COVID-19 vaccine after it fails to clearly stop virus variant. *Science*, (2021). published online Feb 8
7. Organisation for Economic Co-operation and Development, Coronavirus (COVID-19) vaccines for developing countries: An equal shot at recovery. https://www.oecd.org/coronavirus/policy-responses/coronavirus-covid-19-vaccines-for-developing-countries-an-equal-shot-at-recovery-6b0771e6/, 2021. Accessed: 02.04.2021
8. Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., Klimek, P.: Ranking the effectiveness of worldwide COVID-19 government interventions. Nat. Human Behav. **4**(12), 1303–1312 (2020)
9. Skegg, D., Gluckman, P., Boulton, G., Hackmann, H., Karim, S.S.A., Piot, P., Woopen, C.: Future scenarios for the COVID-19 pandemic. Lancet **397**(10276), 777–778 (2021)
10. Askitas, N., Tatsiramos, K., Verheyden, B.: Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study. Sci. Rep. **11**, 1–13 (2021)
11. Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., Jit, M., CMMID COVID-19 Working Group: The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. BMC Med. **19**, 1–12 (2021)
12. European Centre for Disease Prevention and Control, Guidelines for the implementation of nonpharmaceutical interventions against COVID-19. https://www.ecdc.europa.eu/sites/default/files/documents/covid-19-guidelines-non-pharmaceutical-interventions-september-2020.pdf, 2020. Accessed: 02.07.2021
13. Gargoum, S.A., Gargoum, A.S.: Limiting mobility during COVID-19, when and to what level? An international comparative study using change point analysis. J. Transp. Health **20**, 101019 (2021)
14. Gösgens, M., Hendriks, T., Boon, M., Steenbakkers, W., Heesterbeek, H., van der Hofstad, R., Litvak, N.: Trade-offs between mobility restrictions and transmission of SARS-CoV-2. J. Royal Soc. Interface **18**, 1–11 (2021)
15. International Monetary Fund, Policy Responses to COVID-19. https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19, 2021. Accessed: 02.07.2021
16. Krubiner, C., Keller, J.M., Kaufman,J .: Balancing the COVID-19 response with wider health needs: key decision-making considerations for low- and middle-income countries. https://www.cgdev.org/publication/balancing-covid-19-response-wider-health-needs-key-decision-making-considerations-low, (2020). Accessed: 23.06.2021
17. Central Bank of Sri Lanka, Annual Report 2020. https://www.cbsl.gov.lk/en/publications/economic-and-financial-reports/annual-reports/annual-report-2020, 2021. Accessed: 20.08.2021
18. Department of Census and Statistics - Sri Lanka, Population and housing. http://www.statistics.gov.lk/, 2020. Accessed: 15.04.2020
19. Worldometer, COVID-19 coronavirus pandemic: Sri Lanka. https://www.worldometers.info/coronavirus/country/sri-lanka/, 2021. Accessed: 15.05.2021
20. National Operation Centre for Prevention of COVID - 19 Outbreak, Official Website for Sri Lanka's Response to COVID-19 response. https://covid19.gov.lk/news/health.html, 2020. Accessed: 10.12.2020
21. Epidemiology Unit - Ministry of Health Sri Lanka, COVID - 19 daily situation report. https://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=225&lang=en, 2020. Accessed: 30.10.2020
22. The Government official news portal - Department of Government Information - Sri Lanka, Divulapitiya Covid cluster total- 1770. https://news.lk/news/political-current-affairs/item/30861-divulapitiya-covid-cluster-total-1770, 2020. Accessed: 09.11.2020
23. S. Perera (*The Island*), Peliyagoda fish market cluster big and widespread in its reach. https://island.lk/peliyagoda-fish-market-cluster-big-and-widespread-in-its-reach/, Oct 23, 2020. Accessed: 10.12.2020
24. Kondo, K.: Testing for global spatial autocorrelation in stata. 2018. MORANSI: Stata module to compute Moran's I, Statistical Software Components S458473, Boston College Department of Economics, revised 14 Jun (2021)
25. Moran, P.A.: Notes on continuous stochastic phenomena. Biometrika **37**(1/2), 17–23 (1950)
26. Monteiro, R.D.C.: Primal-dual following interior point algorithms for semidefinite programming. SIAM Journal on Optimization **7**(3), 663–678 (1997)
27. Bhatia, R.: Matrix Analysis. Springer-Verlag, New York (1997)
28. Zima, M.: A theorem on the spectral radius of the sum of two operators and its application. Bull. Aust. Math. Soc. **48**, [427-434 (1993)
29. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global Sensitivity Analysis: The Primer. John Wiley & Sons, US (2008)
30. Tomovic, R., Vukobratovic, M.: General Sensitivity Theory. North-Holland, Netherlands (1972)
31. Hethcote, H.W.: Qualitative analyses of communicable disease models. Math. Biosci. **28**(3–4), 335–356 (1976)
32. Wang, W.: Population dispersal and disease spread. Discr. Cont. Dyn. Syst.-B **4**(3), 797 (2004)
33. Sun, C., Yang, W., Arino, J., Khan, K.: Effect of media-induced social distancing on disease transmission in a two patch setting. Math. Biosci. **230**(2), 87–95 (2011)
34. Arino, J., Sun, C., Yang, W.: Revisiting a two-patch sis model with infection during transport. Math. Med. Biol. **33**(1), 29–55 (2016)
35. Sattenspiel, L., Dietz, K.: A structured epidemic model incorporating geographic mobility among regions. Math. Biosci. **128**(1–2), 71–91 (1995)
36. Arino, J., Van den Driessche, P.: A multi-city epidemic model. Math. Popul. Stud. **10**(3), 175–193 (2003)
37. Arino, J., Van Den Driessche, P.: The basic reproduction number in a multi-city compartmental epidemic model. In: Posit. Syst., pp. 135–142. Springer, Berlin (2003)
38. Wang, W., Zhao, X.-Q.: An epidemic model in a patchy environment. Math. Biosci. **190**(1), 97–112 (2004)

39. Li, M.Y., Shuai, Z.: Global stability of an epidemic model in a patchy environment. Canad. Appl. Math. Quart. **17**(1), 175–187 (2009)
40. Citron, D.T., Guerra, C.A., Dolgert, A.J., Wu, S.L., Henry, J.M., Smith, D.L.: Comparing metapopulation dynamics of infectious diseases under different models of human movement. Proc. Nat. Acad. Sci. **118**(18), e2007488118 (2021)
41. Calvetti, D., Hoover, A.P., Rose, J., Somersalo, E.: Metapopulation network models for understanding, predicting, and managing the coronavirus disease COVID-19. Front. Phys. **8**, 261 (2020)
42. Saldaña, F., Velasco-Hernández, J.X.: The trade-off between mobility and vaccination for COVID-19 control: a metapopulation modelling approach. Royal Soc. Open Sci. **8**(6), 202240 (2021)
43. Coletti, P., Libin, P., Petrof, O., Willem, L., Abrams, S., Herzog, S.A., Faes, C., Kuylen, E., Wambua, J., Beutels, P., et al.: A data-driven metapopulation model for the Belgian COVID-19 epidemic: assessing the impact of lockdown and exit strategies. BMC Infect. Dis. **21**(1), 1–12 (2021)
44. Zhang, B., Liang, S., Wang, G., Zhang, C., Chen, C., Zou, M., Shen, W., Long, H., He, D., Shu, Y., et al.: Synchronized nonpharmaceutical interventions for the control of COVID-19. Nonlinear Dyn. **106**, 1477–11489 (2021)
45. Wijaya, K.P., Ganegoda, N., Jayathunga, Y., Götz, T., Schäfer, M., Heidrich, P.: An epidemic model integrating direct and fomite transmission as well as household structure applied to COVID-19. J. Math. Ind. **11**(1), 1–26 (2021)
46. Jones, T.C., Mühlemann, B., Veith, T., Biele, G., Zuchowski, M., Hoffmann, J., Stein, A., Edelmann, A., Corman, V.M., Drosten, C.: An analysis of SARS-CoV-2 viral load by patient age. MedRxiv (2020). https://doi.org/10.1101/2020.06.08.20125484
47. Rozhnova, G., van Dorp, C.H., Bruijning-Verhagen, P., Bootsma, M.C.J., van de Wijgert, J.H.H.M., Bonten, M.J.M., Kretzschmar, M.E.: Model-based evaluation of school- and non-school-related measures to control the COVID-19 pandemic. Nat. Commun. **12**, 1–11 (2021)
48. Wijaya, K.P., Páez Chávez, J., Aldila, D.: An epidemic model highlighting humane social awareness and vector-host lifespan ratio variation. Commun. Nonlinear Sci. Numer. Simul. **90**, 105389 (2020)
49. Ganegoda, N.C., Wijaya, K.P., Amadi, M., Erandi, K.H., Aldila, D.: Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany. Sci. Rep. **11**(1), 1–16 (2021)
50. Tang, S., Mao, Y., Jones, R.M., Tan, Q., Ji, J.S., Li, N., Shen, J., Lv, Y., Pan, L., Ding, P., et al.: Aerosol transmission of SARS-CoV-2? Evidence, prevention and control. Environ. Int. **144**, 106039 (2020)
51. Editorial, Covid-19 transmission-up in the air. *Lancet Respir. Med.*, vol. 8, p. 1159, (2020)
52. Campioli, C.C., Cevallos, E.C., Assi, M., Patel, R., Binnicker, M.J., O'Horo, J.C.: Clinical predictors and timing of cessation of viral RNA shedding in patients with COVID-19. J. Clin. Virol. **130**, 104577 (2020)
53. Widders, A., Broom, A., Broom, J.: SARS-CoV-2: the viral shedding vs infectivity dilemma. Infect. Dis. Health **25**(3), 210–215 (2020)
54. Böhning, D., Rocchetti, I., Maruotti, A., Hollinge, H.: Estimating the undetected infections in the Covid-19 outbreak by harnessing capture-recapture methods. Int. J. Infect. Dis. **97**, 197–201 (2020)
55. World Health Organization, WHO supplies rapid antigen detection tests for COVID-19 response. https://www.who.int/srilanka/news/detail/09-11-2020-who-supplies-rapid-antigen-detection-tests-for-covid-19-response, 2020. Accessed: 10.12.2020
56. Jayasena, H., Chinthaka, W.: COVID-19 and developing countries: lessons learnt from the Sri Lankan experience. J. Royal Soc. Med. **113**(11), 464–465 (2020)
57. Kojaku, S., Hébert-Dufresne, L., Mones, E., Lehmann, S., Ahn, Y.-Y.: The effectiveness of backward contact tracing in networks. Nat. Phys. **17**, 652–658 (2021)
58. Griffin, S.: COVID-19: lack of test and trace data is frustrating government scrutiny. Br. Med. J. **369**, m2239 (2020)
59. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Society for Industrial and Applied Mathematics, Philadelphia (1994)
60. Ma, T., Wang, Shouhong: Bifurcation Theory and Applications. World Scientific Press, Singapore (2005)
61. Krasnosel'skii, M.A., Zabreiko, P.P.: Geometrical Methods of Nonlinear Analysis. Springer-Verlag, Berlin (1984)
62. Beardon, A.: Complex Analysis: The Argument Principle in Analysis and Topology. John Wiley and Sons, US (1979)
63. Wijaya, K.P., Páez Chávez, J., Pochampalli, R., Rockenfeller, R., Aldila, D., Götz, T., Soewono, E.: Food sharing and time budgeting in predator-prey interaction. Commun. Nonlinear Sci. Numer. Simul. **97**, 105757 (2021)
64. Al-Salman, A.M., Páez Chávez, J., Wijaya, K.P.: A modeling study of predator-prey interaction propounding honest signals and cues. Appl. Math. Modell. **89**(2), 1405–1417 (2021)
65. Karow, M., Kressner, D.: On a perturbation bound for invariant subspaces of matrices. SIAM J. Matrix Anal. Appl. **35**(2), 599–618 (2014)
66. Fraser, C.: Estimating individual and household reproduction numbers in an emerging epidemic. PloS One **2**(8), e758 (2007)
67. Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S.: A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. **178**(9), 1505–1512 (2013)
68. Thompson, R., Stockwin, J., van Gaalen, R.D., Polonsky, J., Kamvar, Z., Demarsh, P., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., et al.: Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics **29**, 100356 (2019)
69. The World Bank, Life expectancy at birth, total (years) - Sri Lanka. https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=LK, 2021. Accessed: 08.04.2021
70. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al.: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, china: a retrospective cohort study. Lancet **395**(10229), 1054–1062 (2020)
71. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science **368**(6490), 489–493 (2020)
72. Kalbfleisch, J.G.: Probability and Statistical Inference: Volume 1 Probability. Springer, New York (1985)

73. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, UK (2007)
74. Neale, M.C., Miller, M.B.: The use of likelihood-based confidence intervals in genetic models. Behav. Genet. **27**(2), 113–120 (1997)
75. Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics **25**(15), 1923–1929 (2009)
76. Dankowicz, H., Schilder, F.: Recipes for Continuation. Society for Industrial and Applied Mathematics, Computational Science and Engineering, Philadelphia (2013)
77. Doedel, E.J., Champneys, A.R., Fairgrieve, T.F., Kuznetsov, Y.A., Sandstede, B., Wang, X.-J.: *Auto97: Continuation and bifurcation software for ordinary differential equations (with HomCont)*. Computer Science, Concordia University, Montreal, Canada, (1997). Available at http://cmvl.cs.concordia.ca
78. Dhooge, A., Govaerts, W., Kuznetsov, Y.A.: MATCONT: a MATLAB package for numerical bifurcation analysis of ODEs. ACM Trans. Math. Softw. **29**(2), 141–164 (2003)
79. The World Bank, Economic and poverty impact of COVID-19. https://thedocs.worldbank.org/en/doc/15b8de0edd4f39cc7a82b7aff8430576-0310062021/original/SriLanka-DevUpd-Apr9.pdf, 2021. Accessed: 16.06.2021
80. Wijaya, K.P., Sutimin, T.. Götz., Soeowono, E.: On the existence of a nontrivial equilibrium in relation to the basic reproductive number. Int. J. Appl. Math. Comput. Sci. **27**(3), 623–636 (2017)

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

992. ABDULKAREEM, MARIAM. Environmental sustainability of geopolymer composites. 2021. Diss.

993. FAROQUE, ANISUR. Prior experience, entrepreneurial outcomes and decision making in internationalization. 2021. Diss.

994. URBANI, MICHELE. Maintenance policies optimization in the Industry 4.0 paradigm. 2021. Diss.

995. LAITINEN, VILLE. Laser powder bed fusion for the manufacture of Ni-Mn-Ga magnetic shape memory alloy actuators. 2021. Diss.

996. PITKÄOJA, ANTTI. Analysis of sorption-enhanced gasification for production of synthetic biofuels from solid biomass. 2021. Diss.

997. MASHLAKOV, ALEKSEI. Flexibility aggregation of local energy systems—interconnecting, forecasting, and scheduling. 2021. Diss.

998. NIKITIN, ALEKSEI. Microwave processes in thin-film multiferroic heterostructures and magnonic crystals. 2021. Diss.

999. VIITALA, MIRKA. The heterogeneous nature of microplastics and the subsequent impacts on reported microplastic concentrations. 2021. Diss.

1000. ASEMOKHA, AGNES. Understanding business model change in international entrepreneurial firms. 2021. Diss.

1001. MUSTO, JIRI. Improving the quality of user-generated content. 2021. Diss.

1002. INKERI, EERO. Modelling of component dynamics and system integration in power-to-gas process. 2021. Diss.

1003. GARIFULLIN, AZAT. Deep Bayesian approach to eye fundus image segmentation. 2021. Diss.

1004. ELFVING, JERE. Direct capture of CO2 from air using amine-functionalized resin - Effect of humidity in modelling and evaluation of process concepts. 2021. Diss.

1005. KOMLEV, ANTON. Magnetism of metal-free graphene-based materials. 2021. Diss.

1006. RISSANEN, MATTI. EcoGame and Ecosystem Profiler: solutions for business ecosystem management. 2021. Diss.

1007. VANHAMÄKI, SUSANNA. Implementation of circular economy in regional strategies. 2021. Diss.

1008. LEHTINEN, VESA. Organisaation emergentti itseohjautuvuus, case sinfoniaorkesteri: ”Miksi orkesteri soittaa hyvin, vaikka sitä johdettaisiin huonosti?”. 2022. Diss.

1009. KÄHKÖNEN, TIINA. Employee trust repair in the context of organizational change – identification and measurement of active trust repair practices. 2022. Diss.

1010. AHONEN, AILA. Challenges in sport entrepreneurship: cases in team sport business. 2022. Diss.

1011. LEVIKARI, SAKU. Acoustic emission testing of multilayer ceramic capacitors. 2022. Diss.

1012. ZAHEER, MINHAJ. Evaluation of open-source FEM software performance in analysing converter-fed induction machine losses. 2022. Diss.

1013. HAAPANIEMI, JOUNI. Power-based electricity distribution tariffs providing an incentive to enhance the capacity effectiveness of electricity distribution grids. 2022. Diss.

1014. BUAH, ERIC. Artificial intelligence technology acceptance framework for energy systems analysis. 2022. Diss.

1015. GIVIROVSKIY, GEORGY. In situ hydrogen production in power-to-food applications. 2022. Diss.

1016. SOMMARSTRÖM, KAARINA. Teachers' practices of entrepreneurship education in cooperation with companies. 2022. Diss.

1017. KAN, YELENA. Coherent anti-stokes raman scattering spectromicroscopy in biomedical and climate research. 2022. Diss.

1018. MÄNDMAA, SIRLI. Financial literacy in perspective – evidence from Estonian and Finnish students. 2022. Diss.

1019. QORRI, ARDIAN. Measuring and managing sustainable development in supply chains. 2022. Diss.

1020. MARTIKAINEN, SUVI-JONNA. Meaningful work and eudaimonia: contributing to social sustainability in the workplace. 2022. Diss.

1021. MANNINEN, KAISA. Conducting sustainability target-driven business. 2022. Diss.

1022. LI, CHANGYAN. Design, development, and multi-objective optimization of robotic systems in a fusion reactor. 2022. Diss.

1023. CHOUDHURY, TUHIN. Simulation-based methods for fault estimation and parameter identification of rotating machines. 2022. Diss.

1024. DUKEOV, IGOR. On antecedents of organizational innovation: How the organizational learning, age and size of a firm impact its organizational innovation. 2022. Diss.

1025. BREIER, MATTHIAS. Business model innovation as crisis response strategy. 2022. Diss.

1026. FADEEV, EGOR. Magnetotransport properties of nanocomposites close to the percolation threshold. 2022. Diss.

1027. KEPSU, DARIA. Technology analysis of magnetically supported rotors applied to a centrifugal compressor of a high-temperature heat pump. 2022. Diss.

1028. CHAUHAN, VARDAAN. Optimizing design and process parameters for recycled thermoplastic natural fiber composites in automotive applications. 2022. Diss.

1029. RAM, MANISH. Socioeconomic impacts of cost optimised and climate compliant energy transitions across the world. 2022. Diss.

ISBN 978-952-335-832-4
ISBN 978-952-335-833-1 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta 2022