# SPATIOTEMPORAL ANALYSES OF VISCERAL LEISHMANIASIS IN THE INDIAN SUBCONTINENT

by

**TIMOTHY MARK POLLINGTON**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy in**

**Mathematics of Systems**

**Mathematics for Real-World Systems (MathSys) Centre for Doctoral Training**

February 2022

# Contents

# List of Figures

**Impact of intensified control on VL: an ITSA (Chapter 5) 67**

**Appendix C Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic (supplementary to Chapter 3) 123**

**Appendix D Impact of intensified control on VL: an ITSA (Chapter 5) 129**

# List of Tables

Literature review of the 'tau' clustering statistic and the proposition of a new rate estimator (Chapter 2)

Impact of intensified control on VL: an ITSA (Chapter 5)

Appendix C Spatiotemporal clustering with variable exposure times: analysis using a new tau-rate estimator (supplementary to Chapter 4)

Appendix D Impact of intensified control on VL: an ITSA (supplementary to Chapter 5)

# Acknowledgments

I HAVE ATTEMPTED SCIENCE
PLEASE EXPLAIN
I FORMED AN IDEA AND THEN DISCOVERED I WAS WRONG.
THERE ARE NUMEROUS DIAGRAMS
I WAS WRONG IN NUMEROUS WAYS
I PRODUCED A DETAILED TRIBUTE TO MY WRONGNESS
THAT IS SCIENCE
NATHANWPYLE

wrongness, by Nathan.W. Pyle ©

## 0.1 Thanks—general to this thesis

My deepest respect for the three supervisors who mentored me throughout the PhD. **Déirdre Hollingsworth** (TDH) helped in the indirect funding of my PhD research through the Neglected Tropical Disease Modelling Consortium. She drew my attention to the *Newton Bhabha PhD programme* in 2017 and, with **Lloyd Chapman** (LACC), edited the grant application, which turned out successful, with the help of C. Veal from Research Support Services at the University of Warwick. Déirdre also kindly supported my visitor status at Big Data Institute (BDI) (University of Oxford) since 2018, which has widened my access to rare books & their high-performance cluster computer. I was the first PhD student of Lloyd, who was very generous with his time & conscientiousness in his supervision. In December 2016, he suggested I read the new Lessler et al. [118] paper on the tau statistic. Then a year later suggested the `surveillance` *R* package [130] to analyse the longitudinal dataset in Chapter 5. Both of these interventions led to productive research and formed the foundations of this thesis. Lloyd joined me in 2017 for the first two weeks of the Newton Bhabha PhD placement in Bihar, India, helping with my orientation & introduction to key research contacts in VL. While Déirdre & Lloyd have been there from the start, I am also grateful to **Mike Tildesley** (MJT) for accepting me under his wing since 2018, when Déirdre & Lloyd moved to the BDI & LSHTM, respectively. Mike gave me tutorial support and helped extend my PhD funding during the first COVID-19 lockdown. My enlightened supervisors have given me the freedom to choose a pragmatic research path where I have developed modelling skills within my mathematical capability.

Love to my Mum, Brother, Stu & Olly who gave me emotional & financial support. In answer to her question, the PhD is now complete!

I thank MathSys (Centre for Doctoral Training) for developing the MSc/PhD programme I enrolled in 2015. Of note are M. Richardson, C. Connaughton, H. Robson, D. Walker & J. Perkins for their pastoral & administrative assistance and, of course, fellow PhD students in my cohort. Since 2018 my new research home has been at the BDI, and I thank all of the Hollingsworth research group there (past & present) who made my stay welcome and enriched my postgraduate experience.

My appreciation to T. Hontvari (Legal Advisor for University of Warwick's Research & Impact Services), who oversaw the successful drafting of two data-sharing agreements between the University of Warwick & i) the CDC (many thanks to C. Bern & B. Marston who engaged CDC in this agreement, in association with icddr,b) and ii) RMRIMS, without which the research for Chapters 4 & 5 would not have been possible.

Last but not least, was the essential role of the funders of the MathSys programme (BMGF [grant № OPP1184344 via the NTD Modelling Consortium; EPSRC, MRC & University of Warwick [grant № EP/L015374/1]) which funded a 3-year stipend with additional funding during the COVID-19 pandemic. My supervisors were supported through BMGF [grant № OPP1184344 to LACC & TDH via the NTD Modelling Consortium, OPP1183986 to LACC via SPEAK India consortium]. I am also grateful to the funders for allowing me to change to part-time study and the commensurate extension to the PhD completion date for health matters. Views, opinions, assumptions or any other information within this thesis should not be attributed to these funders nor any person connected with them. No funders had a role in the study design, collection, analysis, interpretation of data, writing, or decision to submit any parts of this thesis.

## 0.2 Chapter-specific thanks

### 0.2.1 Chapter 2

I thank J. Lessler & H. Salje, who openly answered questions on their work. Also, S. Truelove replied to questions by email.

### 0.2.2 Chapter 3

H. Salje & J. Lessler helpfully shared their unpublished analysis code to reproduce the analysis in Lessler et al. [118], and I had useful discussions with S. Truelove, J. Giles & H. Salje at the Epidemics$^7$ poster presentation [162]. Mari Myllymäki answered `GET` *R* package support questions. Thank you to the editors & two anonymous reviewers of *Spatial Statistics* who made pertinent suggestions on the manuscript. A discussion with P.J. Diggle (PJD) was pivotal in forming proper methods for graphical hypothesis testing & methods for estimating the clustering range [57]. This influenced the improvements to the inference methods reported in Chapters 3 & 4; PJD also reviewed the second draft manuscript of [165]. The open-source software used is also credited[1].

### 0.2.3 Chapter 4

I am grateful to S. Morley, T. Lestang & D. Nadlinger of the Oxford code review network, who studied the code in July 2021 and gave speed-up recommendations [147]. Also, for the conversations with J. Toor & T. Crellen on elevated prevalence vs. risk for measuring public health outcomes.

The computational aspects of this chapter were supported by the Wellcome Trust Core Award grant № 203141/Z/16/Z & NIHR Oxford BRC. The views expressed are not necessarily those of the NHS, the NIHR or the Department of Health. I thank G. Ferrari of BMRC for his kind assistance when using the high performance computing cluster. The open-source software used is also credited[2].

[1] [106, 116, 129, 130, 136, 169, 177, 210, 212, 213]

[2] [23, 60, 66, 129, 132, 146, 169, 175, 177, 212, 213, 214]

### 0.2.4 Chapter 5

My deep gratitude to Rajendra Memorial Research Institute of Medical Sciences (RMRIMS) for hosting me during the Newton Bhabha PhD programme, funded by Newton Fund [grant № 69871277] and administered by the British Council in India. RMRIMS is an internationally-recognised research institute based in Patna, Bihar. Over the past decade, Bihar state has been the epicentre of VL transmission in the ISC. Working alongside world-leading experts in VL epidemiology provided me with invaluable insights. RMRIMS provided me with board & lodging between August & November 2017, but more importantly, a warm family of staff & students that made my stay a much cherished one. My sincere thanks to my mentor Dr N.A. Siddiqui-ji, who oriented me in current VL research, sent me into the field to see PKDL surveillance teams and became a dear friend who introduced me to their family. Thank you for trusting me and giving me the honour of analysing the impact of your 2015–2017 intervention.

The following thanks go to the people who contributed to the pilot study or data collection that resulted in the pilot & dataset, which I analysed. I am very grateful to the Ministry of Health & Family Welfare, Govt. of India; Directorate General of Health Services & the Directorate of NVBDCP. Especially the then Dr J. Prasad (Director General) & Dr A.C. Dhariwal (Director, NVBDCP) for their valuable support. Additionally, the indispensable contribution of Bihar State Health Authorities, the Additional Director-cum-State VBD Programme Office, Patna for recording, aggregating & kindly sharing the surveillance data. Additionally, I would like to thank A. Ranjan, Data Manager-cum-Project Manager (BMGF), the PhD students, technical staff & field staff who contributed to the study for their sincere support. I would also like to render my sincere thanks to health workers, Kala-azar[3] Technical Supervisors, Vector Borne Disease consultants, other paramedical staff of the health system network & community leaders of the study district. Lastly, I am very grateful to the thousands of Vaishali's rural villagers who willingly & enthusiastically participated in the pilot.

Regarding my analysis, I thank M Cameron & G. Medley for valuable discussions on estimating cases averted, S. Law for manuscript review, S. Meyer & J. Bracher for package advice on `surveillance` & `hhh4addon` [31, 130], and the editors of *Epidemics* journal & two anonymous reviewers for valuable input on the manuscript under second review. The

[3] Kala-azar is the hindi word for VL and in the ISC is synonymous with primary VL.

open-source software used is also credited[4].

The pilot study was supported by institutional funding (annual budget allocation) from ICMR, Dept. of Health Research, Ministry of Health & Family Welfare, Govt. of India on the instruction of the Directorate General of Health Services, Ministry of Health & Family Welfare, Govt. of India to VK, NAS, SK, VNRD, KP, RM & PD.

An anonymous manuscript review sent to *Epidemics* journal, fed back the need to be more relevant for policy. I subsequently improved the flow of the text that became Chapter 5 and made minor additions, to frame this analysis following the recent release of WHO's NTD 2021—2030 roadmap and why its results still matter.

[4] [24, 25, 76, 129, 151, 157, 177, 209, 211]

# Declarations

## 0.3 Inclusion of material from a prior thesis

This thesis is submitted to the University of Warwick to support my application for the degree of Doctor of Philosophy in Mathematics of Systems. It has been composed by myself and has not been submitted for any previous degree or professional qualification, apart from:

- the rationale for this research, as described in Chapters 1 & 6, draws heavily, and at times verbatim, from the unpublished PhD proposal drafted by myself and edited by supervisors Hollingsworth & Chapman, submitted to funders EPSRC & MRC in September 2016.

- near-verbatim descriptions of the Bangladesh dataset, internal consistency checks & missing data (§4.2.2), which were previously submitted in September 2016 for the Master of Science in Mathematics of Systems [161] preceding this PhD. However, for the data cleaning that was performed as described in §4.2.2, this was additional to that performed in the MSc study: the entire cleaning process was reviewed, code rewritten, and extra checks added, and so this work should be considered new & separate. This additional task took five months to complete.

## 0.4 Collaborative work

Research Chapters 2–5 are all in the process of, or have been, published. Therefore parts of these chapters contain verbatim representations of text & graphics in submitted material:

- Chapter 3 in Pollington et al. (2020). Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic. *Spatial Statistics.* DOI: 10.1016/j.spasta.2020.100438.

- Chapter 5 in Kumar (co-first), Siddiqui (co-first), Pollington (co-first) et al. Impact of intensified control on VL in a highly-endemic district of Bihar, India: an ITSA. *Re-submitted for second review with Epidemics journal.*

- Chapters 2 & 4 form a single manuscript in Pollington et al. Analysing spatiotemporal clustering with variable exposure times using the tau statistic: an application to VL. *In preparation.*

The work presented within this thesis was carried out by myself except in the cases outlined below:

- open-access measles dataset obtained from the `surveillance` *R* package [130] and originally collected via [138, 144, 153] for Chapter 3.

- field epidemiologists led by C. Bern, who collected & processed the Fulbaria village dataset for Chapter 4 from their original studies [18, 96, 170].

- monthly district-level incidence data supplied via RMRIMS for Chapter 5.

The following subsections describe collaborators' contributions to the submitted works and thus indirectly to this thesis using the CRediT framework [5]. The specific contributions of my three supervisors are denoted by TDH, LACC or MJT. They have also read & offered comments on the structure of Chapters 1 & 6.

### 0.4.1 Chapter 2: Literature review & Chapter 3: Developments in ...

*TMP*: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Literature review, Data curation, Writing - original draft & editing, Visualisation. *MJT*: Conceptualisation, Writing - review & editing, Supervision. *PJD*: Methodology, Validation, Writing - review & editing. *TDH*: Conceptualisation, Writing - review & editing, Supervision, Funding acquisition. *LACC*: Conceptualisation, Software, Validation, Data curation, Writing - review & editing, Supervision.

### 0.4.2 Chapter 4: Spatiotemporal clustering with variable exposure times: analysis using a new tau-rate estimator

*Conceptualisation*: TMP LACC TDH MJT; *Funding acquisition*: TDH MJT; *Methodology*: TMP PJD; *Validation*: TMP S. Morley T. Lestang D. Nadlinger; *Formal analysis, investigation, visualisation & writing - original draft*: TMP; *Data curation*: C. Bern D. Mondal B. Marston; *Supervision*: LACC TDH MJT; *Writing - review & editing*: All authors.

### 0.4.3 Chapter 5: Impact of intensified control on VL in a highly-endemic district of Bihar, India: an ITSA

My thanks to the researchers at the institute who described their recent study designs through a series of interviews to retrospectively inform the analysis plan. V. Kumar (VK), N.A. Siddiqui (NAS), S. Kesari (SK), V.N.R. Das (VNRD), K. Pandey (KP) & P. Das (PD) declared a competing interest as they were the permanent employees of RMRIMS, and R. Mandal a PhD student under its Dept. of Vector Biology. They initiated this institutional study on the instruction of the Directorate General of Health Services, Ministry of Health & Family Welfare, Govt. of India.

*Pilot conception & design*: VK R. Mandal (RM) SK PD; *Pilot implementation*: RM SK; *Secondary data collection*: VK NAS S. Das (SD); *Pilot supervision & resources*: PD KP VNRD SD; *Analysis conception & design*: TMP TDH LACC; *Data analysis & interpretation*: TMP NAS RM LACC TDH; *Analysis supervision & resources*: TDH LACC; *Manuscript drafting*: TMP LACC TDH NAS VK RM KP SK SD PD; *Literature search*: TMP LACC TDH NAS RM; *Figs. & tables*: TMP LACC TDH; *Appendices*: RM NAS VK; *Critical article revision*: LACC TDH TMP NAS RM PD SD VNRD KP; *Final approval for publication*: All authors.

## 0.5 Other research

During the PhD, I have co-authored several other papers that are listed below and are separate from this thesis—except for a paper on unit testing in infectious disease models [122], whose principles I have endeavoured to uphold in code to prevent error & improve confidence in the repeated results; and the Chapman et al. PNAS paper [44], which uses the dataset that I cleaned for Chapter 4:

- Lucas, Pollington et al. (2020). Responsible modelling: Unit testing for infectious disease. *Epidemics.* DOI: 10.1016/j.epidem.2020.100425

- Chapman, . . . , Pollington, et al. (2020). Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and PKDL. *PNAS.*
DOI: 10.1073/pnas.2002731117

- Crellen, . . . , Pollington et al. (2021) Dynamics of SARS-CoV-2 with waning immunity in the UK population. *Phil. Trans. Roy. Soc. B.* DOI: 10.1098/RSTB.2020.0274

- Davis, . . . , Pollington et al. (2021) Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. *Nat. Commun.* DOI: 10.1038/s41467-021-25531-5

- Clark, . . . , Pollington et al. (2021) How modelling can help steer the course set by the World Health Organization 2021–2030 roadmap on neglected tropical diseases. *Gates Open Res.* DOI: 10.12688/GATESOPENRES.13327.1

* * *

As per the expectations of MathSys [126], I believe this thesis to be of a publishable standard (Chapter 3 is published & Chapter 5 at second journal review stage) and contains the development & application of novel mathematics to understand a real-world challenge. I submit this thesis to the examiners Kat Rock (internal) & Henrik Salje (external) for consideration and thank them for agreeing to be involved in this process.

# Summary

The neglected tropical disease visceral leishmaniasis (VL) has greatly burdened vulnerable populations in the Indian subcontinent. The analyses in this thesis are motivated by observations from VL field epidemiology that cases cluster in space & time. Using a household-level dataset from a highly-endemic Bangladeshi village covering the years 2002–2010, I investigate spatiotemporal clustering using the tau statistic to estimate the magnitude & spatial range of clustering of cases to inform control interventions and to validate a recent mechanistic model result. Then, for Vaishali district, India, I employ a spatiotemporal statistical model to assess if an intensified intervention pilot during 2015–2017 was successful and how many cases may have been averted while accounting for district-level clustering of incidence.

To deliver high-quality insights, several novel advances in methodologies were made. A literature review of the tau statistic was performed that detailed its existing uses & methods of inference to assess the presence of spatiotemporal clustering and estimate the range of clustering around cases. This prompted corrections & improvements in inference methods leading to higher precision in clustering estimates than a previous baseline analysis on a measles dataset. A new rate estimator for the tau statistic was created to account for variable person-time at risk in the Bangladeshi study. Finally, customisations in the use of the `surveillance` & `hhh4addon` *R* packages were made to perform an interrupted time series analysis for the Vaishali study.

The findings of this thesis contribute to the current VL discourse by quantifying spatiotemporal clustering around cases, partially validating a recent result on clustering and giving a rigorous evaluation of a control pilot that may be required if incidence recrudesces. For spatiotemporal statistics, improvements in the tau statistic and the new applications of these *R* packages offer valuable examples in methodology & code for other infectious diseases. I summarise the findings of this thesis and list further research opportunities in VL, which I hope to explore as my career in infectious disease modelling progresses.

# List of abbreviations

ACD: active case detection

AIC: Akaike information criterion

ASHAs: accredited social health activists

BCa: bias-corrected & accelerated

BDI: Big Data Institute, University of Oxford

BMGF: Bill & Melinda Gates Foundation

CDC: Centers for Disease Control & Prevention, US

CI: confidence interval

DD: diagnosis-to-diagnosis

DHS: demographic health survey

EPHP: elimination as a public health problem

EPSRC: Engineering & Physical Sciences Research Council

GPS: global positioning system

HIV: human immunodeficiency virus

IDW: inverse distance-weighted

IEC: information, education & communication

IgG & IgM: immunoglobulin G & M antibodies

ILI: influenza-like illness

IP: incubation period

IQR: interquartile range

IRS: indoor residual spraying

ISC: Indian subcontinent

ITSA: interrupted time series analysis

LACC: Lloyd AC Chapman (supervisor)

LSHTM: London School of Hygiene & Tropical Medicine

MJT: Mike J Tildesley (supervisor)

MMPSB: modified marked point spatial bootstrap

MPCF: multitype pair correlation function

MPSB: marked point spatial bootstrap

MRC: Medical Research Council

MRCA: most recent common (genetic) ancestor

NTD: neglected tropical disease

NVBDCP: National Vector Borne Disease Control Programme

OD: onset-to-diagnosis

OT: onset-to-treatment distribution

PCD: passive case detection

PIT: probability integral transform

PKDL: post–kala-azar dermal leishmaniasis

PRIME-NTD: policy-relevant items for reporting models in epidemiology of neglected tropical diseases summary

PTAR: person-time at risk

RISB: resampled-index spatial bootstrap

RMRIMS: Rajendra Memorial Research Institute of Medical Sciences, Patna, Bihar, India

RPS: ranked probability score

SI: serial interval

SIR: susceptible-infectious-recovered compartmental model

SIS/SIRS: susceptible-infectious-susceptible-(recovered) model

TB: tuberculosis

TDH: T Déirdre Hollingsworth (supervisor)

TMP: Timothy M Pollington

URI: upper respiratory illness

UTM: universal transverse mercator

VL: visceral leishmaniasis

WHO: World Health Organization

§: main text section symbol

# CHAPTER 1

# Introduction

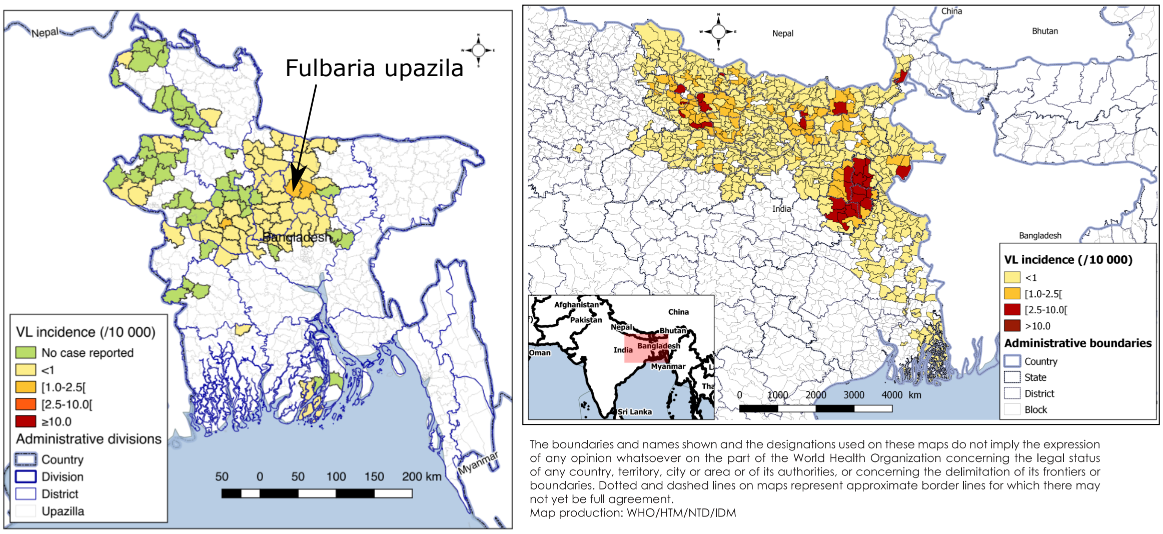
*An introduction to visceral leishmaniasis covering its recent history and pressing research questions is provided, which summarises the context in which this research started. Spatiotemporal features of the disease are particularly pertinent, and I provide examples of this. Finally, a thesis outline describes the evolution of the research.*

## 1.1 Visceral leishmaniasis disease

The disease of visceral leishmaniasis (VL) that persists in the Indian subcontinent (ISC) is caused by the *L. donovani* parasite, which is transmitted between humans [78] by the female *P. argentipes* sandfly [14]. The disease is usually fatal unless treated [89]. It is a chronic systemic infection causing recurrent fever, "fatigue, weakness and loss of appetite & weight", anaemia and "enlarged lymph nodes, spleen" and sometimes liver [28]. I use the term 'VL' to refer to the disease in its primary and most common manifestation, not its sequela *post–kala-azar dermal leishmaniasis* (PKDL), which although non-fatal is highly stigmatising—both stages are infectious [89]. PKDL is a chronic skin rash lasting from months to years while VL is a systemic illness [170]. In the ISC it is considered that PKDL infection would remain without treatment [215].

VL is a *neglected tropical disease* (NTD)—identified along with others because until recently it received insufficient global attention & funding compared with better-known global diseases like malaria, TB (tuberculosis) or HIV (human immunodeficiency virus). NTDs are often found in the tropics according to the environment niches that the pathogen (and sometimes vector) can survive in. They disproportionately affect the world's poorest populations through impacting individual's ability to work, live healthily & support family [40]. Despite large-scale elimination efforts, the continued burden that VL has on marginalised communities is clear, and the task to eliminate it is shared by all countries of the ISC. In 2015, 147 million people in Bangladesh, Bhutan, India, Nepal & Thailand were estimated to be at risk [202], with India & Bangladesh as the biggest contributors back then (Fig. 1.1). Through funding from the Bill & Melinda Gates Foundation (BMGF) & others, in recent times, modelling has helped to answer a range of pressing research questions for VL that have included elimination policy [114], elimination thresholds [140] & potential vaccine impact [115].

The disease was targeted by World Health Organisation (WHO) for 'elimination as a public health problem' (EPHP) (<1 case/10K people/yr) from the ISC by 2020 [201]. This was based on strong intergovernmental commitment, better diagnostic tools, enhanced surveillance, scaling-up sandfly control & faster access to medicines [201]. Falling cases in Bihar (India), one of the highest incidence regions in the ISC from 25,222 to 7,615 (2011–14), appeared to support the evidence base for the success of this strategy [201]. Since this PhD began, the 2020 target has nearly been reached,



The boundaries and names shown and the designations used on these maps do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.
Map production: WHO/HTM/NTD/IDM

Figure 1.1: **Disease distribution of new VL cases at upazila (Bangladesh, left map) or block (India, right map) level per 10,000 population in 2015.** Fulbaria upazila is indicated on the left map which is the location for the analysis in Chapter 4. Amended maps from WHO [203, 204].

with 46 blocks (subdistricts) above the 1 VL case (new & relapses)/10,000 population threshold in mid-December 2019 [207]. Commendably, the WHO roadmap now has three new targets with a 2030 deadline [206, 207]:

1. PKDL elimination (all PKDL cases detected & treated from recovered VL cases followed up for 3 years)

2. VL case fatality rate ($< 1\%$ nationally)

3. reaffirming the VL elimination target (all blocks in India at $< 1$ new/relapsing VL case per 10,000 population).

To shed more light on the transmission cycle that results in successful human-to-human transmission I detail the transmission stages (Fig. 1.2). In the ISC, the only known host reservoir that can maintain this lifecycle is humans, unlike other *Leishmania* species globally, where canines can also be hosts. After a successful bite of an infected human, the parasite develops in the midgut of the female sandfly and migrates to the anterior midgut & foregut, from where it is regurgitated into a new human during the sandfly's next blood meal [14]. The time from the bite of an infectious human to the infective bite of another human is an estimated five days [176]. Within the human host, they transform into a stage infective to sandflies, upon which the human becomes infectious.

The incubation period (IP) in the human before the onset of symptoms has been estimated to range "from 10 days to over a year" with typical cases found between 2–6 months [199]. However, in forming a serial interval distribution these range estimates were not useful as summary estimates based on empirical studies were needed. The IP has been estimated from a modelling study to have a mean of six months [42, 44]. The serial interval distribution for VL used throughout this thesis has a mean of 7 months: the composite distribution was based on summary estimates from an incubation period estimated from a transmission model on data from Bangladesh [42, 44] & infection period distribution estimated from a statistical model [99] on data from Bihar, India, respectively, as detailed in §5.2.3.1.1.

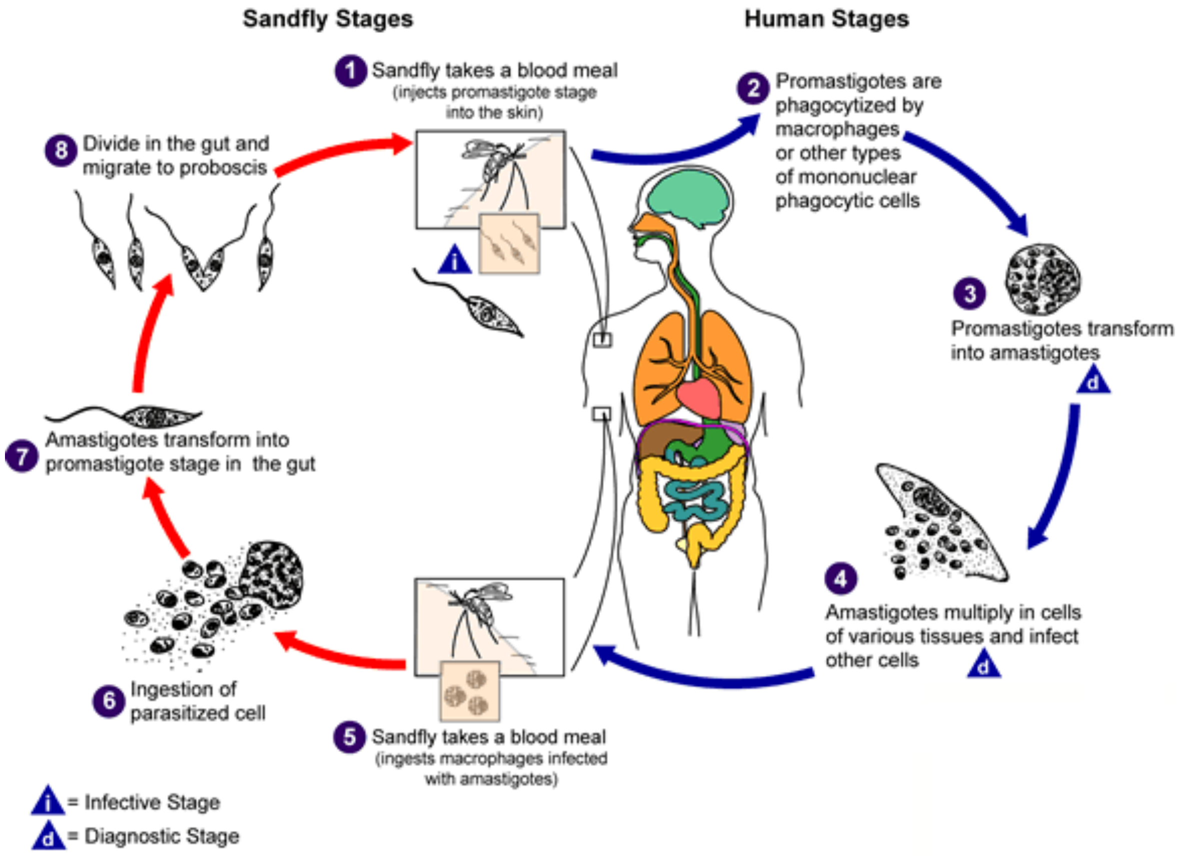

Sandfly Stages
Human Stages
1 Sandfly takes a blood meal (injects promastigote stage into the skin)
2 Promastigotes are phagocytized by macrophages or other types of mononuclear phagocytic cells
3 Promastigotes transform into amastigotes
4 Amastigotes multiply in cells of various tissues and infect other cells
5 Sandfly takes a blood meal (ingests macrophages infected with amastigotes)
6 Ingestion of parasitized cell
7 Amastigotes transform into promastigote stage in the gut
8 Divide in the gut and migrate to proboscis
i
d
i = Infective Stage
d = Diagnostic Stage

Figure 1.2: **Lifecycle of the parasite *L. donovani*** during successful human-to-human transmission via the *P. argentipes* sandfly vector. Amended image from CDC [41].

Seasonal patterns in VL incidence are common which are primarily forced by the annual cycle of sandfly abundance and time-varying human-sandfly contact behaviour to a lesser degree [199]. However, due to the variable incubation period the time of peak incidence is unlikely to match the time of peak sandfly-human transmission [199].

An asymptomatic stage of VL also exists which can serve as a reservoir of the parasite in the community. It can be confirmed using a serological test (the rK39 rapid diagnostic test) with an *absence* of the systemic symptoms associated with the infectious form (§4.2.1). Seroprevalence ranges from 7%–63% in the Indian subcontinent [34, 188] and accounts for 4–17 times the number of infections versus symptomatic cases [91]. As an asymptomatic case could lead to active disease, it may serve as a predictive marker to trigger follow-up surveillance to catch active cases earlier [193, 186]. The 95% confidence interval of the probability of infection to sandflies from asymptomatic cases has been estimated at (0, 2·3%) [190]. Given this information it was reasonable for Chapman et al. to assume a 2% probability *a priori* for their spatiotemporal model [44] which estimated that living in the same household as an asymptomatically-infected VL case had a monthly infection rate 49 times lower than residing with a symptomatic case.

It is widely accepted that there is a paucity of precise parameter estimates available to properly parametrise sandfly biology [37]. For instance, the sandfly life expectancy estimate is highly variable [176]. Although it does not make sense to currently model within-host dynamics, an alternative research path that would be productive is modelling the spatial features of the disease. Clustering in VL transmission has been observed at various spatial scales from household to state levels [17]. Nowadays, more detailed datasets are available that can provide district- or household-level locations. A review of VL models [176] found that only a single study modelled human VL and accounted for spatial features. However, this was only in a very crude sense by representing migration as a population-averaged flow into a compartmental model [68]. Since the review was published an individual-level mechanistic model has characterised infection risk as a function of distance from a case using a spatial-kernel approach with geolocation data for a Bangladeshi village [44]. A spatial scan statistic analysis has also been applied to village-level data for a subset of an Indian district [33].

While *L. donovani* species is responsible for VL in the ISC, the disease also occurs largely in Brazil & East Africa, however outside of the ISC human transmission plays a smaller role compared to zoonotic reservoirs [171], e.g. in Brazil dogs also contribute to the infectious reservoir [208]. *L. donovani* is also responsible for visceral disease in East Africa and *L. chagasi* in Brazil. The performance of VL rapid diagnostic tests as well as pathogen drug susceptibility (even in East Africa with the same species as the ISC) varies widely between these regions [50, 15]—PKDL can be left to self cure in Sudan whereas it is treated in the ISC [215].

There are also other forms of leishmaniasis disease—cutaneous & mucocutaneous, which while not as serious as the visceral form [208], cause considerable life-long suffering. The cutaneous form causes ulcerated or plaque-type skin lesions on exposed skin such as the face & arms [103] and often leave scarring post-treatment. The mucocutaneous form causes “destruction of mucous membranes of the nose, mouth and throat” [208], thus affecting the key functions of eating, swallowing, breathing & speaking. The subsequent community stigma based on a lack of understanding affects social participation and life chances in education, marriage & employment [103]. It is unsurprising that this harms psychological well-being and could lead to psychiatric illness [103]. PKDL cases follow the same fate but mercifully are saved scarring post-treatment if the lesions are not longstanding [215].

Active Case Detection (ACD) involves community volunteers or public-health staff actively locating themselves temporarily in the field rather than passively waiting for cases to self-refer at existing health care settings. There are four main types of ACD in the ISC which have a range of sensitivity and cost-effectiveness: i) *camps* (a temporary location in the centre of the village where cases self-refer, with splenic diagnosis facilities and previous Information, Education & Communication priming activities); ii) *house-to-house* search; iii) *incentive-based*; and iv) *index-case approach* (a special case of ii, using spatial proximity to index cases) [95]. In Bihar, India, efforts are made by both community health workers (accredited social health activists) and public health staff (from RMRIMS or CARE India for example) but splenic examination or rK39 tests are performed by medical staff on-site or back at the block-level public health facility.

In the index-case approach, case & non-case households are visited following the report of a recent index case—in India & Nepal during 2011 they defined ‘recent’ as onset or diagnosis within the last 12 months [95]. By detecting & treating the (infectious) VL or PKDL case, their illness is halted and along with it their contribution to onward transmission. The intervention is constrained spatially around the index case and within a certain time after its detection. A spatial limit is based on VL’s known spatial features (§1.3) and that secondary cases will likely be living closest to the index case (and most probably in the same household). Although the theoretical range of risk extends to the maximum flight range of a sandfly (up to 309m [158], §4.4.4), due to resource constraints and its untested efficacy at this range, the intervention radius is far smaller than this e.g. 50–75m [95]. Within the temporal limit of less than a transmission generation, one aims to find secondary cases before they can infect tertiary cases. However, due to the long & varied incubation period these secondary cases may not reach their symptomatic stage (which is thought to be the most infectious) on first visit, so repeat visits over several months are necessary. As VL incidence falls in India, the focus has shifted to PKDL: in 2017 RMRIMS were making field visits to outside community meeting spaces to solely find new PKDL cases.

## 1.2 Clustering & spatiotemporal infection processes

Knox defines *clustering* as “a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance” [104]. This could be *hotspot clustering* (also known as *local clustering*) which is “any area within the study region of significant elevated risk” [113]. Or alternatively, *global clustering*, which is a general tendency to cluster across a study. I avoid using the term ‘spatial/spatiotemporal dependence’ as it is not clearly defined in the literature.

Pathogen transmission is a dynamic process in time & space. Infectious diseases spread because a pathogen is transmitted by ‘contact’ with ‘parent’ cases (where ‘contact’ loosely includes the transmission from a parent case(s) via a vector, airborne transmission, fomites, environmental contamination *etc.*). It is therefore expected that observed cases are infected by a parent case(s) (infector(s)) when they are proximal in time *and* space. Throughout this thesis, all studies assume only one infector is required for a successful contact, even though for VL, the quality of infectious dose may alter disease outcome [79], and thus, successive infectious bites may increase the probability of transmission or development of disease. The additional distinction of a *spatiotemporal* infection process is because, usually, any case will only be infectious briefly relative to the study duration, thus leaving a temporal signal coinciding with their spatial presence. These spatiotemporal coincidences at the individual level can be explicitly captured in pair-relatedness variables [118, 182]; indeed, in Chapters 2–4, I shall employ the pairwise *tau statistic*. In contrast, non-infectious cases & their risk factors may only cluster spatially and less frequently spatiotemporally—however, the disease of pellagra serves as a cautionary example. It spatially clusters at the village & household levels (due to poverty & shared diet) and varies seasonally with the crop harvest, so it could feasibly produce an individual-level spatiotemporal signal, despite its non-infectious (dietary) cause [134].

## 1.3 Previous spatiotemporal analyses of VL

A rapid review that I did of human VL in India considered four recent papers [12, 21, 22, 152]. The cross-sectional study by Barnett et al. [12] explored VL incidence and the risk factors for being a case—two of which were spatial: distance to the nearest VL household & the number of VL households within 10m, both of which are proxies for first-order case intensity across the study. Two villages with a combined total of 2,203 people from 245 households were surveyed, including non-cases as well as cases, so that odds ratios could be computed to assess the risk factor effects. Barnett et al. [12] only considered the nearest VL household, ignoring contributions from other VL households close by, or non-VL (but possibly asymptomatic) households. Truncating at 10m would ignore direct transmission from households up to a few hundred metres away, based on the furthest distance a sandfly has been observed to fly [158]. The cowshed location, a common attractor of sandflies [17] & whose density is associated with protection [18], was not recorded and assumed to adjoin the house. They found weak statistical evidence (in each separate villages analysed) that the odds of being a VL case within 10m of a VL household (but excluding that household during three years) was 1·77 times higher (village 1) and 4·11 times higher (village 2) than at any other distance, after adjusting for other variables.

Perry et al. [152] produced VL prevalence estimates and identified risk factors. They used the local SaTScan™ spatial statistic with the Bernoulli test to allow for “small and large clusters to be detected", presumably using cases & non-cases but with a time-constant risk from the partial information provided. If spatial variation in risk varies over time (*e.g.* as an epidemic progresses, fewer susceptibles will surround an infected case, so a discrete Poisson probability model may have been more appropriate [125]. Within the two villages in Saran district, Bihar state, India that they surveyed, the two most likely clusters they found with strong statistical evidence had a relative risk (within them versus outside) of 138·3 & 11·9 and a radius of 54 & 100m, respectively. Note that the circular scan statistic has been criticised for identifying false positive discs if the true cluster is elliptically-shaped [192], which could lead to overestimation in their relative risk estimates.

The first of two Bhunia et al. [21] studies used village-level case data, sandfly collection data & remotely-sensed environmental covariates (Normalised Difference Vegetation Index, wetness & land use/cover). Entomological & environmental data is rare in Indian VL studies. They used a weighted-raster risk model, but the relative influence of each spatial covariate appears to have been manually selected. This produced a VL risk map: *i.e.* a raster of locally-estimated risk estimates. An additional limitation was their use of local polynomial interpolation to both estimate vector abundance & indoor climatic temperature, whereas spatial dependence in these values would suggest a geostatistical model, say, with kriging to model this dependence. The second Bhunia et al. [22] study detected hotspots within Vaishali district, India using an inverse distance-weighted (IDW) spatial kernel estimate to create a smoothed risk map for 2007–2011. The risk map is not accompanied by a map of the uncertainty in estimation. Their results support the hypothesis that clustering wanes over the years of the epidemic. However, it is over-simplistic to study the panel risk maps from 2007–2011 and infer disease spread/diffusion across just five time snapshots. Again I think a more sophisticated geostatistical model could replace this parametric IDW kernel, informed by the empirical variogram of the data.

Overall, from the four papers studied, there was a lack of a spatiotemporal analysis—*i.e.* analysis of variation in risk over space & time *together*. The spatial data resolution at the village level may be sufficient in high-endemicity settings where village-wide interventions are necessary. However, when elimination incidence levels are approached, it may lead to poorly targeted, wasteful control strategies intended for the whole village rather than ones specific to index cases & their immediate neighbours. In Chapter 4, I shall estimate the clustering endpoint distance around a typical VL case to address this issue.

## 1.4 Aims of this thesis & rationale

Considering that spatiotemporal clustering features in observational VL studies, I propose the following thesis aims:

1. understand how VL disease risk changes with distance from an infected case through estimation from spatiotemporal datasets
2. develop models to account for spatiotemporal features to answer pertinent questions in VL control

The primary research focus of this PhD is to perform VL analyses while accounting for the role of space in VL transmission. VL case detection is mainly passive, and control is based predominantly on blanket indoor residual spraying (IRS) of insecticide to reduce sandfly densities. However, evidence from household-level studies suggests that this strategy could be improved (*e.g.* by active case detection (ACD), contact tracing or reactive spraying around new cases), as there are long delays to diagnosis & treatment [29], and the risk of infection appears to be higher when living near or in the same house as an infected individual [18, 155]. Nevertheless, the current understanding of how risk varies with proximity to infected individuals is crude. Improved understanding of the spatial & temporal scales of VL transmission will inform more effective control strategies.

## 1.5 Approach

I took a pragmatic route to achieve these aims through Chapters 2–5 to provide policy-relevant estimates of clustering and develop new models to describe VL transmission. The essential ingredients for successful modelling are data, a suitable analysis framework and a relevant research question. In these analyses, I focus on two studies from Bangladesh & India, which have experienced some of the world's highest VL incidences in the last two decades. Regarding the meaning of 'upazila' (Bangladesh) & 'block' (India) spatial levels, both are one adminstrative level beneath the 'District' level as denoted on Fig. 1.1.

### 1.5.1 VL epidemic in a rural village in Bangladesh at the household level

I analysed a rich dataset collected by Prof. C. Bern & co-workers in a community cohort study covering a VL epidemic in Fulbaria upazila, Bangladesh in 2002–2010 [170]. It featured individual-level clinical & migration events and household locations.

The tau statistic was identified as a suitable statistic for descriptive spatiotemporal analysis of the data. This is based on its non-parametric form, ability to measure the magnitude of disease frequency change at different spatial scales as well as the range of spatial clustering, and the simulation studies showing its robustness to missing cases (§2.1 & 2.2). Therefore, in **Chapter 2**, I explored it further in a literature review. In the review, I assess previous papers that used this new statistic, critique them and offer improvements to make its implementation more consistent & rigorous. While the dataset was awaiting data-sharing agreement approval, I used an open-access measles dataset to test these improvements in **Chapter 3**.

Although the tau statistic is robust to case underreporting & spatial observation bias [118], there are unexplored aspects like the bootstrap sampling method & confidence interval type that may bias the estimated range of clustering. Using the previous analysis of Lessler et al. [118] as a baseline, in Chapter 3, I assess these aspects in terms of corrected graphical hypothesis testing of clustering & parameter estimation.

For the VL dataset, my first step was data cleaning which took several months due to the size & detail of the dataset. The research branched at this point into a) Lloyd Chapman utilising the dataset to estimate the VL risk profile around VL & PKDL cases using a spatial kernel transmission model [44] of which I was credited for data cleaning, editing the article & producing a map, and b) my application of the non-parametric tau statistic in **Chapter 4** to also measure the clustering scale for the same dataset.

### 1.5.2 An intensified district-level VL intervention during state-wide declines in incidence in Bihar, India

This project aimed to measure the impact the intensified intervention on Vaishali district during 2015–2017 had on VL case counts via an interrupted time series analysis (ITSA). During the collaboration, I worked with routine longitudinal governmental data which was unavailable for outside researchers. The dataset recorded monthly case counts for 33 districts in Bihar state from before the intervention (2012) until 2017. This topic addresses the urgent need for evidence to support new control policies to achieve VL elimination, alongside providing an example application of this modern modelling framework to measure programme impact. In **Chapter 5**, I further split these down into two analytical questions to ask i) if the intervention had had an effect while accounting for declining state-wide trends & transmission from neighbouring districts, and ii) if it did have an effect, how large it was—in terms of the number of VL cases averted. I customised a spatiotemporal statistical model framework, available through the well-established open-source `surveillance` & `hhh4addon` *R* packages, to answer these research questions.

The analysis of this non-randomised study design is a common challenge in practical public health epidemiology. To the best of my knowledge, it is one of the first ITSA applications of the *R* packages `hhh4addon` & `surveillance`'s hhh4 endemic-epidemic modelling of areal count time series. The extensive modelling notes in Appendix D & shared code [160] enables rapid re-use and assimilation into future research.

## CHAPTER 2

# Literature review of the ‘tau’ clustering statistic & proposing a new rate estimator

*In this first research chapter, I provide a detailed exposition of the ‘tau’ clustering statistic, the tau-distance & tau-time classes of its estimator and propose a new rate estimator for studies with variable person-time at risk. Once the tau statistic has been introduced, detailed results are presented of a forward literature search during its first eight years. This builds an understanding of the statistic from which suggestions are made for its development in the ‘tau’ research Chapters 2–4.*

## Abstract

The *tau statistic* is a recently-developed second-order correlation function for assessing the magnitude & range of global spatiotemporal clustering. It can be applied to epidemiological data containing geolocations of individual cases and temporal data on cases, such as the time of onset of disease symptoms (onset time). Different forms of the statistic (distance & time forms) can be used to assess spatial or temporal clustering given prior data on temporal/spatial relatedness between cases. The time form of the tau statistic can provide information on when the observed incidence rate is higher than average after an index case is detected, which is relevant for active surveillance. A new rate version of the statistic (the *tau-rate* estimator) is defined that accounts for variable person-time at risk (PTAR), ideal for studies with open populations. A forward literature search is performed on the original papers defining or reforming the statistic. This is the first review which explores its use & the aspects of its computation & presentation that could affect inferences drawn and bias estimates derived from it and inspires further analysis in Chapters 3 & 4.

Only half of the 16 included studies were considered to be using ‘proper’ tau statistics in line with the original papers that founded the tau statistic. However, their inclusion in the review still provided important insights into their analysis motivations. All papers that used graphical hypothesis testing & parameter estimation used incorrect methods. There is a lack of research on choosing the a) time-relatedness interval to relate case pairs or b) distance band set—both are required to calculate the statistic. Some studies demonstrated nuanced applications of the tau statistic in settings with unusual data or time relation variables, which enriched understanding of its possibilities.

## 2.1 Spatiotemporal clustering & the tau statistic, $\tau$

Understanding whether a disease process is clustered and estimating the magnitude & scale of this clustering in spatiotemporal terms is vital in modern epidemiology. The increasing availability of accurate geolocation data in recent years has enabled a better understanding of many diseases [187]. This can help inform decisions on infectious disease control to save limited public health or veterinary resources [90, 110]. However, clustering statistics in this domain typically disregard the *spatiotemporal* aspect, considering only the spatial dimension of data & metrics (or the spatial dimension at a series of fixed time points) (§2.2). The tau statistic [182] is more appropriate than most statistics for this task as it measures spatiotemporal rather than just spatial clustering, produces non-parametric estimates (without process assumptions) and, unlike the $K$ function [73], offers a relative magnitude in the difference of risk, rate or odds of disease (§2.3.1) versus the background level [118, 163]. The tau statistic should not be confused with 'Kendall's tau statistic/rank correlation coefficient' [26].

The distance-form of the non-parametric tau statistic $\tau^{(d)}$ evaluates a disease frequency measure (odds, prevalence or rate) within a certain annulus around an average case (Fig. 2.1) and compares it to a non-spatial 'background' measure (*i.e.* the same measure over any distance) [118, 163, 182]. Tau values signify either the presence of spatiotemporal clustering ($\tau > 1$), no clustering ($\tau = 1$) or inhibition ($\tau < 1$). It measures the general tendency of case or event pairs to cluster across a study spatially (*i.e.* a *global* statistic) while implicitly accounting for their potential to be transmission-related temporally using temporal information, making it a spatiotemporal statistic [118, 164, 182]. Occasionally, space & time are swapped to explicitly measure temporal clustering using the time-form of the (*tau-time*) tau statistic, with transmission relations based on spatial proximity (§2.4).

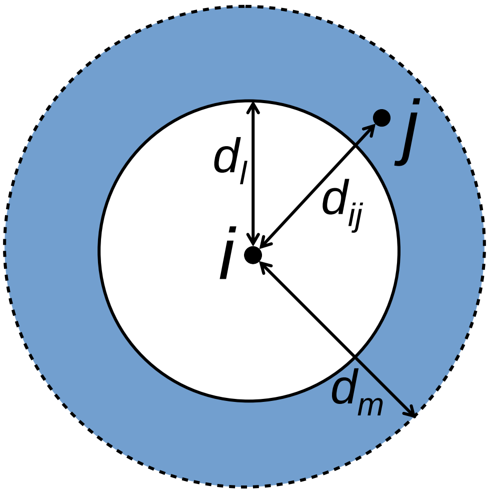


Figure 2.1: **A single distance band half-closed annulus of radii** $[d_l, d_m)$ around an average case $i$ with another case $j$ in it, separated by distance $d_{ij}$.

The $\tau$ statistic was first defined & applied in Salje et al. [182]. Later, Lessler et al. [118] described its context within spatial statistics & epidemiology, demonstrated robustness, formulated estimators for 'case-only' or 'case & non-case' data, and reformed nomenclature. Both, termed the *original papers*, have inspired a steady stream of research applying the $\tau$ or similar statistics.

## 2.2 Other statistics & tests for global clustering

The review in Chapter 2 focuses on the tau statistic [118, 182], but other statistics for assessing disease clustering have been informally reviewed to provide context. Ward summarises spatiotemporal methods for disease data [197] by those based on mechanistic modelling like the spatiotemporal kernel model compared against in §4.3.4 [44], or based on statistical modelling like the Matérn cluster process that describes a spatiotemporal point process; where statistics may be chosen for computational efficiency or the assumptions & sensitivities of the spatial distributions of the underlying population at risk [197]. Alternatively, empirical measures can estimate the global clustering of individual cases (first-order) or case pairs (second-order).

There are some similarities between the $\tau$ and earlier first-order spatial statistics, which focused on regions of excess risk $R(\mathbf{x}) = \lambda(\mathbf{x})/g(\mathbf{x})$ where the numerator represented the case intensity at location $\mathbf{x}$ in space $S$ & denominator the "background effect" [112]. Information on the scale of clustering can be obtained by changing tolerance bounds to detect where it is strongest [111].

Cuzick & Edwards' $k$-nearest neighbours test [51], Anderson & Titterington's Integrated Squared Difference function [7] & Tango's C [191] are tests for clustering that classify the data as cases or controls. Unfortunately, they only describe clustering in the spatial dimension. These three tests assume "two independent inhomogeneous Poisson processes with spatially-varying intensities: $m_1(\mathbf{x})$ for sampled cases & $m_2(\mathbf{x})$ for sampled controls" [191] randomly chosen from "individuals at risk in the study region" [191].

- *Cuzick & Edwards' k-nearest neighbours test* sums the number of case-case pairings within a specific range [51], which has similarities to the tau statistic.

$$T_k := \sum_i \sum_j a_{ij}\delta_i\delta_j, \text{ where } \delta_i = \mathbb{1}(i \text{ is a case}), \text{ and for locations } \mathbf{x}_j \text{ of}$$
$$\text{case } j, a_{ij} = \mathbb{1}(\mathbf{x}_j \in k\text{-nearest neighbours of } \mathbf{x}_i)$$

- *Anderson & Titterington's Integrated Squared Difference function ($\widehat{\text{ISD}}$)* smooths the difference of non-parametric kernel density-estimated relative risks in cases & controls ($\hat{m}_1, \hat{m}_2$) at point $\mathbf{x}$ across 2D space $S$ [7, 191].

$$\widehat{\text{ISD}} := \int_{\mathbf{x}\in S} \left(\hat{m}_1(\mathbf{x}) - \hat{m}_2(\mathbf{x})\right)^2 d\mathbf{x} \tag{2.1}$$

- *Tango's C* imposes a parametric kernel in the $\widehat{\text{ISD}}$ (Eqn. 2.1), *e.g.* a step function for hotspot clusters or exponential decay for clinal clusters [191].
- *Spatiotemporal K-function*, initially developed for stationary point processes [58], has strong connections to the tau statistic, as mentioned in the appendix of Salje et al. [182]. Epidemiologically, its stationarity will never adequately explain a disease process, and a constant intensity does not take into account population heterogeneity. Gabriel & Diggle's *inhomogeneous K function* extended it using a special class of inhomogeneous point processes [73] and is available through the `stpp` *R* package [74]. It requires a spatial case intensity estimate via kernel-based density estimation & temporal estimate from time series modelling [73], so the calculation can be lengthy.

As the incremental Knox test [3] & phi statistic [182] analyse spatiotemporal *interaction* rather than clustering, they are not studied within this thesis.

## 2.3 Tau-distance estimators, $\hat{\tau}^{(d)}$

The review (§2.5) uncovered different types of tau statistic. They are classified here by their estimator type—odds $\hat{\tau}_{\text{odds}}$ (§2.3.1 & Eqn. 2.2), or prevalence $\hat{\tau}_{\text{prev}}$ (§2.3.2 & Eqn. 2.3), and whether they act as the distance-form $\big($denoted $\hat{\tau}^{(d)}$ or $\hat{\tau}^{(d)}(d_l, d_l)\big)$ on distance annulus $[d_l, d_m)$ (§2.3.3 and Eqns. 2.2 & 2.3), or time-form $\big(\hat{\tau}^{(t)}$, $\hat{\tau}^{(t)}(t_1, t_2)\big)$ on time interval $[t_1, t_2)$ (Eqns. 2.7 & 2.8) about an average case, respectively. A new rate estimator is proposed $\hat{\tau}_{\text{rate}}$ (Eqns. 2.4–2.6) within the distance-form (§2.3.3). New versions are proposed of the odds & prevalence estimators of the time-form (Eqns. 2.7 & 2.8).

The odds & prevalence forms of the tau statistic provide a measure of the relative odds & prevalence respectively of temporally- (or spatially-) related cases within a certain distance (or time) band versus related cases over any distance (or non-negative time difference).

### 2.3.1 Odds ratio estimator, $\hat{\tau}^{(d)}_{\text{odds}}$

The most common tau estimator for the distance-form is the *odds estimator* $\hat{\tau}^{(d)}_{\text{odds}}$. It is the ratio between i) the estimated odds $\hat{\theta}^{(d)} \equiv \hat{\theta}(d_l, d_m)$ of finding any case $j$ that is 'related' (definition to be discussed below) to any other case $i$, within a half-closed annulus $[d_l, d_m)$, $(l, m \in \mathbb{Z}^+,\ l < m)$, around case $i$, to ii) the odds $\hat{\theta}(0, \infty)$ of finding any case $j$ related to any case $i$ over any distance separation $(d_{ij} \geq 0)$ [118] (Eqn. 2.2 & Fig. 2.1); the odds estimate $\hat{\theta}^{(d)}$ (Eqn. 2.2) is the ratio of the number of related case pairs $(z^{(d)}_{ij} = 1)$ within $[d_l, d_m)$ versus the number of unrelated case pairs $(z^{(d)}_{ij} = 0)$ within $[d_l, d_m)$. Unlike the prevalence & rate estimators, it is only applied to $n$ *cases*. The half-closed annulus is a correction to the original open interval (Lessler et al. [118]: Appendix 5); it was incorporated in December 2018 into the `IDSpatialStats` *R* package (which calculates the tau statistic) [116].

$$\hat{\tau}^{(d)}_{\text{odds}}(d_l, d_m) := \frac{\hat{\theta}(d_l, d_m)}{\hat{\theta}(0, \infty)},\ \ \hat{\theta}(d_l, d_m) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}\Big(z^{(d)}_{ij} = 1, d_l \leq d_{ij} < d_m\Big)}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}\Big(z^{(d)}_{ij} = 0, d_l \leq d_{ij} < d_m\Big)},$$
$$\text{where } z^{(d)}_{ij} = \begin{cases} 1, \text{if } t_j - t_i \in [T_1, T_2] \\ 0, \text{if } t_j - t_i \notin [T_1, T_2], \text{where } t_i, t_j = \text{onset times of disease symptoms,} \end{cases} \tag{2.2}$$
$$\text{and } \mathbb{1}(\cdot) \text{ is the indicator function: equal to 1 when its arguments are all true \&}$$
$$\text{0 otherwise. } T_1, T_2 \text{ represents the range of the temporal relatedness definition.}$$

The relatedness of a case pair $z^{(d)}_{ij}$ is commonly determined using temporal information (such as the difference in the onset times of both cases $t_j - t_i$) [163]. The *serial interval* (SI) is the period between the onset times of symptoms in the infector $t_i$ & their infectee $t_j$. Typically cases are defined as being temporally related when their onset times are within a single mean SI of each other, i.e. $T_1 = 0$ & $T_2 =$ mean SI. For both distance- & time-forms, relatedness can also include serological $(z^{(d)}_{ij} = 1$ for same-serotype pair) or genotype information $\big($*e.g.* $z^{(d)}_{ij} = 1$ if $i, j$ share their most recent common ancestor (MRCA) within some time of their onsets [181]$\big)$ [118] (Table 2.1).

There are inherent challenges to only using a temporal relatedness indicator based on a date of symptom onset: it may be unavailable (§2.5.5); the disease state may have been misclassified (§2.5.6); the temporal range representing a single transmission chain is flexible (§2.5.5). Infections may go undetected because of an asymptomatic disease state, diagnostic failure, failure of patient to respond to symptoms or barriers to healthcare access. The use of additional relatedness indicators is known to reduce the underestimate of tau within the region of spatiotemporal clustering (§2.5.10 &

[118]). Lessler et al. [118] have shown that clustering of undetected infections through spatially-biased observation does not affect the results of the tau statistic. However, it is reasonable to assume that an asymptomatic state could bias the tau statistic as cases are likely to remain in the state for different periods of time than those in the symptomatic state.

The main computation of Eqn. 2.2 is effectively a double sum over 'relatedness' indicator functions $\mathbb{1}(\cdot)$ for case pairs. $\hat{\tau}(d_l, d_m)$ is then evaluated over a *distance band set* $\underline{\Delta}$. Sometimes an expanding disc is described by setting $d_l = 0$, relabelling $d_m = d$ to give $\hat{\tau}(d)$ instead. Although $\hat{\tau}$ is strictly evaluated for a given distance band $[d_l, d_m)$ when a $\tau$-distance graph is drawn, a value of $\hat{\tau}(d)$ can be obtained through linearly interpolating between the distance band midpoints. Unlike the $\tau$ vs. distance graphs considered by the review (as described in §2.5), plotting each point estimate $\hat{\tau}(d_l, d_m)$ at the midpoint $\frac{1}{2}(d_l + d_m)$ is deprecated to the endpoint $d_m$, to reduce a common reader error causing an incorrect read-off of the wrong clustering endpoint distance [165].

$\hat{\tau}^{(d)}_{\text{odds}}$ is similar to the conventional odds ratio in epidemiology, as it is a 'ratio of odds' yet note how the numerator's distance condition $(d_l \leq d_{ij} < d_m)$ is a subset of the denominator $(d_{ij} \geq 0)$, whereas traditionally an odds ratio contrasts two mutually-exclusive conditions. It is the $\hat{\theta}^{(d)}$ or $\hat{\pi}^{(d)}$ estimator functions rather than quotient function $\hat{\tau}^{(d)}$ which is "equivalent to ratios of ... multitype pair correlation functions" (MPCF) *c.f.* Lessler et al. [118]: the $\tau$'s functional form cannot be an MPCF because the numerator's distance band $[d_l, d_m)$ is nested within the denominator's $[0, \infty)$ (Eqns. 2.2–2.4).

The tau statistic $\tau$ is a *second-order* correlation function because the (potential transmission, denoted by a broken arrow '$\dashrightarrow$') time-directed difference $\{t_j - t_i : t_j \geq t_i\}$ in symptom onset of *pair* $i \dashrightarrow j$ are considered, not just individual *case* $i$. These measures are particularly appropriate for investigating the infection process *between* individuals since it is typical to assume that one parent case $i$ infects '$\rightarrow$' one susceptible offspring $j$ (where a solid arrow '$\rightarrow$' denotes a definite transmission). For self-immunising diseases and assuming a single contact is a sufficient infectious dose then a pair will share at most one transmission. Although the chronology & identity of pairs is unknown, $\tau$ copes with this by considering those spending transmission-competent $i$ '$\dashrightarrow$'$j$ occasions together.

### 2.3.2 Relative prevalence estimator, $\hat{\tau}^{(d)}_{\text{prev}}$

Additional non-case location data allows one to compute the *prevalence estimator* (distance-form) $\hat{\pi}^{(d)} \equiv \hat{\pi}(d_l, d_m)$ of related case pairs within a certain annulus versus any case or non-case pairing, and thus $\hat{\pi}^{(d)}$ approximates a risk of onset [118]. The tau statistic then becomes the relative prevalence of related case pairs within an annulus $\hat{\pi}(d_l, d_m)$, versus anyone (*case* or *non-case*) at any distance from an average case $i$, $\hat{\pi}(0, \infty)$, applied to $N$ *people*.

$$\hat{\tau}^{(d)}_{\text{prev}}(d_l, d_m) := \frac{\hat{\pi}(d_l, d_m)}{\hat{\pi}(0, \infty)}, \quad \hat{\pi}(d_l, d_m) = \frac{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{1}\left(z^{(d)}_{ij} = 1, d_l \leq d_{ij} < d_m\right)}{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{1}(d_l \leq d_{ij} < d_m)}, \tag{2.3}$$

and $z^{(d)}_{ij}$ is defined as in Eqn. 2.2.

### 2.3.3 A new rate ratio estimator, $\hat{\tau}^{(d)}_{\text{rate}}$

The distance forms of the odds & prevalence measures (Eqns. 2.2 & 2.3) assume a fixed time-in-study for all, thus ignoring migration, births & deaths. This may lead to inaccuracies in their ratio (*i.e.* the tau statistic) when analysing long studies as it ignores that everyone is exposed to infection risk for different durations—epidemiologists typically take account of this through *person-time at risk*

(PTAR). Furthermore, the actual burden of disease would be underestimated for diseases that confer little immunity and occur as multiple events $\big($*e.g.* cholera where different strains can co-circulate infection from either yet does not confer cross-protection [2, 89]$\big)$, assuming only one event per person was counted.

Therefore a new *rate form of the tau statistic*, $\hat{\tau}^{(d)}_{\text{rate}}$ is introduced which accounts for these aspects using 'case & non-case' data with time-varying geolocations, study entry & exit times and onset & recovery times. $\hat{\tau}^{(d)}_{\text{rate}}$ is defined as the ratio of the incidence rate $\lambda$ for individuals within distance band $[d_l, d_m)$ to that of individuals at any distance $[0, \infty)$ (Eqn. 2.4).

$$\hat{\tau}^{(d)}_{\text{rate}}(d_l, d_m) := \frac{\hat{\lambda}(d_l, d_m)}{\hat{\lambda}(0, \infty)}. \tag{2.4}$$

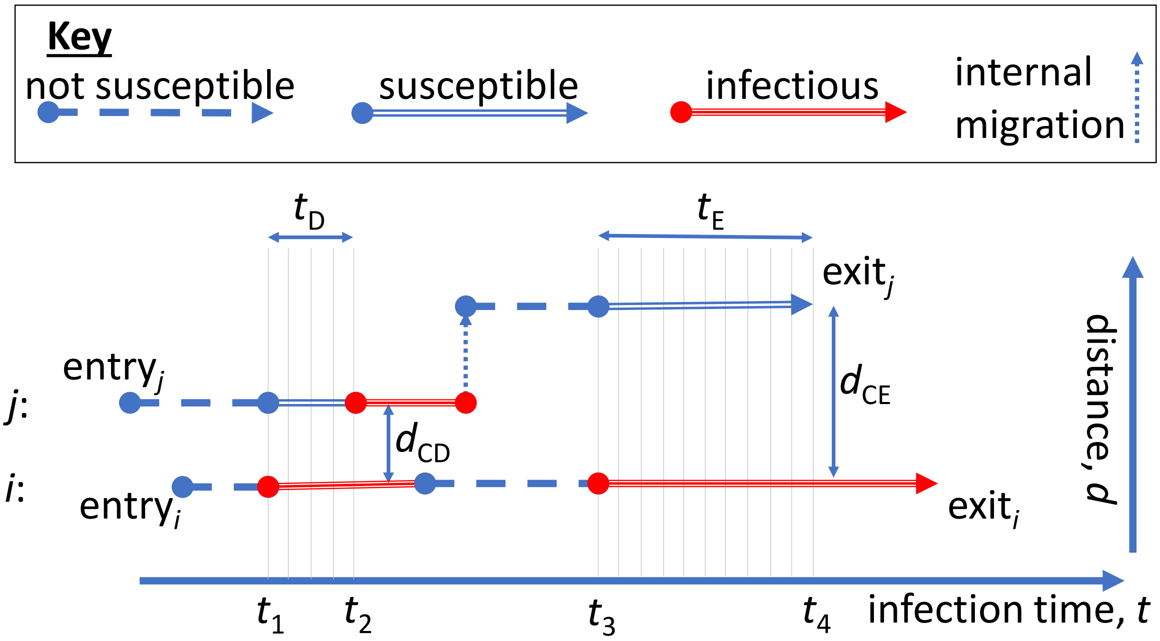


Figure 2.2: **Pair-relatedness for tau-rate distance-form estimator $\tau^{(d)}_{\text{rate}}$ in Eqn. 2.4.** A generic example describing PTAR calculation (during transmission-competent occassions $t_D + t_E$) of individual $j$, due to $i$, by accounting for $j$'s changing location w.r.t. $i$, $j$'s susceptibility to infection & $i$'s infectious period. $j$ enters via birth/immigration at some location D and stays within $d_{CD}$ distance separation of $i$'s eventual entry at some location C. $j$ is considered exposed for 4 time units to the risk of $i$ during $[t_1, t_2]$ since $j$ loses maternal/previous immunity at $t_1$, becoming susceptible. This potentially-related event pair for $i \dashrightarrow j$ is only counted if $d_{CD}$ is within the distance band $[d_l, d_m)$ under calculation. At $t_2$, $j$ becomes infected (not necessarily from $i$) and is no longer at risk. $j$ moves to distance $d_{CE}$ from $i$ but is exposed again to risk from $i$ for 10 time units during $[t_3, t_4]$ when $i$ is infectious again. This new event pair is counted only if $d_{CE}$ is still within $[d_l, d_m)$. The non-susceptible periods after infection for both $i$ & $j$ illustrated here could represent the duration of pathogen-killing chemotherapy or time-limited protection conferred by this generic disease—for VL, infection-acquired immunity would last for a considerable time.

$\lambda$ is traditionally defined as the number of new events divided by PTAR [167]. The formulation for $\hat{\lambda}$ was composed as follows. For $N$ people (composed of $n_c$ unique total cases and thus in generality allowing multiple cases per person, although it is limited to one here) with $K$ events during the study, $\hat{\lambda}$ is estimated by summing their $K$ event pairings in the numerator and $PTAR$ in the denominator (Eqn. 2.5). In this multiple-event paradigm, there may be multiple disjoint overlaps of the susceptibility of $j$ & infectiousness of $i$ that contributes to $j$'s total time-at-risk (Fig. 2.2). Individual $i$ can have from zero to multiple disease events, labelled $a_i$. For a single event ($a_1$ say), there are one-to-many *pair events* that represent a potential transmission link to an event $b$ of $j$ (*i.e.* $a_1 \dashrightarrow b_1, a_1 \dashrightarrow b_2, \ldots$). The *relatedness* of a case event $z_{ab}^{(d)}$ can be determined using temporal information $\big($*e.g.* close onset times of disease symptoms $t_a, t_b$ within some time interval $[T_1, T_2]\big)$; typically $T_1 = 0$ & $T_2 =$ mean SI. To be counted in the numerator of Eqn. 2.5 those onset events need to be within a time difference $\big($*e.g.* $t_b - t_a \in [0, \text{mean SI}]\big)$ for $z_{ab}^{(d)} = 1$ *and* within $[d_l, d_m)$. However, the denominator of Eqn. 2.5 describes the total pair-time at risk and sums the time that the infectious$_i \dashrightarrow$ susceptible$_j$ pair spatiotemporally coincides, when $i$ is infectious during [inf. start$_i$, inf. end$_i$] (*e.g.* between symptom onset $t_i$ & treatment time) and $j$'s susceptibility during [susc. start$_j$, susc. end$_j$].

$$\hat{\lambda}(d_l, d_m) = \frac{\sum_{a=1}^{K}\sum_{b=1,k_l\neq k_m}^{K} \mathbb{1}\left(z_{ab}^{(d)} = 1, d_l \le d_{ab} < d_m\right)}{\sum_{i=1}^{n_c}\sum_{j=1,j\neq i}^{N}\sum_{t=1}^{t_{\text{end}}} \mathbb{1}\left(Z_{ij}^{(d)}(t) = 1, d_l \le d_{ij}(t) < d_m\right)},$$
$$\text{where } z_{ab}^{(d)} = \begin{cases} 1, & \text{if } t_b - t_a \in [T_1, T_2] \\ 0, & \text{if } t_b - t_a \notin [T_1, T_2], \end{cases} \tag{2.5}$$

$$\text{with } Z_{ij}^{(d)}(t) = \mathbb{1}\Big(\big([\text{inf. start}_i, \text{inf. end}_i] \cap$$
$$[\text{susc. start}_j, \text{susc. end}_j] \cap [t]\big) \neq \{\phi\}\Big), \tag{2.6}$$

where $k_a$ denotes which individual the event $a$ belongs *i.e.* $k_l \neq k_m$ means self-comparisons are prohibited when $l$ & $m$ events belong to the same person [118]. The units of $\hat{\lambda}$ are time$^{-1}$ or, more specifically, people–pair-time at risk$^{-1}$. The rate estimator differs from the odds or prevalence tau estimators $\hat{\theta}^{(d)}$ & $\hat{\pi}^{(d)}$ (Eqns. 2.2 & 2.3), which are *person-oriented* in both their numerator & denominator as they sum over cases, or cases & non-cases, respectively.

As the infector & their respective infectee(s) are usually present asynchronously, the relatedness term $Z_{ij}$ needs to be time-directed so the term `abs`$(t_j - t_i)$ as featured in code for [164] would not work, and it would need to be replaced with $t_j - t_i$ and the double sum $\Sigma_i \Sigma_j$ would need to evaluate over *all* values of $i \in [1, n]$ & $j \in [1, n]$. Whereas in Pollington et al. [165] the odds & prevalence estimators could choose a faster-to-compute upper-triangular 'half-summation' where $i \in [1, n]$ & $\big(j \in [1, n]\big) \wedge \big(j < i\big)$.

## 2.4 Tau-time estimators, $\hat{\tau}^{(t)}$

The relatedness of cases $z_{ij}^{(t)}$ is now defined through *spatial* proximity (Eqn. 2.7 & Fig. 2.3). $\tau(t_1, t_2)$ is still considered a tau statistic, as spatiotemporal information is retained just calculated differently. As $\tau^{(t)}$ uses time bands, *only cases* (implicitly with temporal data) can be considered. The diagnostic/indicative plot becomes a $\tau^{(t)}$ vs. time graph $\big($*e.g.* Fig. 4.11$\big)$. It is plotted from a connected line of point estimates $\big\{\hat{\tau}(t_l, t_m) : [t_l, t_m) \in \mathcal{T}\big\}$ for the *time band set* $\mathcal{T} \coloneqq \big\{[t_l, t_m) : l, m \in \mathbb{Z}^+, m = l+1, l \le C\big\}$ over $C$ total time bands, similar to the distance band set $\mathcal{D}$ definition in Pollington et al. [165]. Any actual $i \rightarrow j$ transmission but with negative SIs are ignored, which would lead to bias for infections with a long and/or variable latent period & pre-symptomatic infectious period.

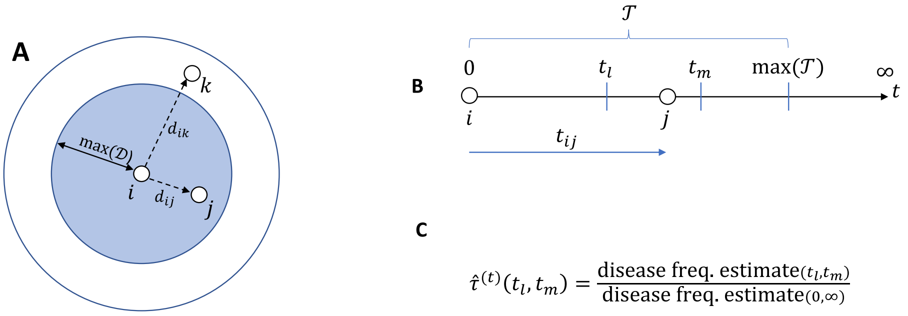

A
k
$d_{ik}$
$\max(\mathcal{D})$
i
$d_{ij}$
j
B
$\mathcal{T}$
0
$t_l$
$t_m$
$\max(\mathcal{T})$
∞
t
i
j
$t_{ij}$
C
$\hat{\tau}^{(t)}(t_l, t_m) = \frac{\text{disease freq. estimate}_{(t_l,t_m)}}{\text{disease freq. estimate}_{(0,\infty)}}$

Figure 2.3: **The key components of the tau-time statistic.** It is a pairwise calculation around an average case $i$ surrounded by other cases, which is only evaluated for spatially-proximal pairs within a fixed range $\max(\mathcal{D})$ and thus $z_{ij}^{(t)} = 1$, else if more spatially distal then $z_{ik}^{(t)} = 0$ (**A**), and furthermore (**B**) conditional on their onset time difference $t_{ij} = t_j - t_i$ being within a specific time band $[t_l, t_m)$. The tau-time statistic is the ratio of a disease frequency measure (odds or prevalence), evaluated over a specific time band $[t_l, t_m)$ (from a series of time bands within the time band set $\mathcal{T}$) versus any time separation $[0, \infty)$, as shown in **C**.

### 2.4.1 Odds ratio estimator, $\hat{\tau}^{(t)}_{\text{odds}}$

The *time-form of the odds estimator* $\hat{\theta}(t_1, t_2)$ is the odds of disease in distance band $[d_l, d_m)$ for cases whose onset difference $t_{ij}$ is within $[t_1, t_2)$ versus those cases separated by any non-negative time difference $t_{ij} \geq 0$.

$$\hat{\tau}^{(t)}_{\text{odds}}(t_1, t_2) := \frac{\hat{\theta}(t_1, t_2)}{\hat{\theta}(0, \infty)} \equiv \frac{\hat{\theta}(t_1, t_2)}{\hat{\theta}(t_1 = 0, t_2 = \infty)}.$$

A symmetric switch is made between space $d$ & time $t$ variables:

$$\hat{\theta}(t_1, t_2) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}\left(z_{ij}^{(t)} = 1, t_1 \leq t_{ij} < t_2\right)}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}\left(z_{ij}^{(t)} = 0, t_1 \leq t_{ij} < t_2\right)}, \tag{2.7}$$

$$\text{where } z_{ij}^{(t)} = \begin{cases} 1, & \text{if } d_{ij} \in [D_1, D_2) \\ 0, & \text{otherwise.} \end{cases}$$

and $z_{ij}^{(t)}$ is evaluated over fixed parameters $D_1, D_2$ which we set to $D_1 = 0, D_2 = \hat{D}$ as described in 4.2.4.

### 2.4.2 Relative prevalence estimator, $\hat{\tau}^{(t)}_{\text{prev}}$

Unlike $\hat{\tau}^{(t)}_{\text{rate}}$, $\hat{\tau}^{(t)}_{\text{prev}}$ can still be formulated; like $\hat{\tau}^{(t)}_{\text{odds}}$, it operates on the smaller case-only dataset.

$$\begin{aligned}
&\hat{\tau}^{(t)}_{\text{prev}}(t_1,t_2) := \frac{\hat{\pi}(t_1,t_2)}{\hat{\pi}(0,\infty)},\\
&\hat{\pi}(t_1,t_2) = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{1}\left(z^{(t)}_{ij}=1, t_1 \leq t_{ij} < t_2\right)}{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{1}(t_1 \leq t_{ij} < t_2)},\\
&\text{where } z^{(t)}_{ij} \text{ is defined as in Eqn. 2.7.}
\end{aligned} \tag{2.8}$$

### 2.4.3 Why a rate estimator cannot exist for the time-form, $\tau^{(t)}$

The odds & prevalence estimators for the time-form $\hat{\tau}^{(t)}_{\text{odds}}$ & $\hat{\tau}^{(t)}_{\text{prev}}$ can be obtained by switching spatial $d$ & temporal $t$ variables (*e.g.* in Eqn. 2.7). However, it is not possible to construct a rate estimator for the tau-time statistic: to remain a rate estimator in Eqn. 2.5, the denominator of $\hat{\lambda}(t_1,t_2)$ needs to continue to sum PTAR. In the distance-form, the denominator sums PTAR conditional on distance band separation & infectious-susceptible states coinciding (Eqns. 2.5–2.6). However, as a symmetric corollary, the time-form of the denominator should sum PTAR conditional on time band separation while still requiring infectious-susceptible states coincide, *e.g.* $\sum_{i=1}^{n_c}\sum_{j=1,j\neq i}^{N}\sum_{t=1}^{t_{\text{end}}}\mathbb{1}(Z_{ij}(t)=1, t_1 \leq t_{ij}(t) < t_2)$; yet the time band separation requires case-case pairs, thus missing the real metric of interest—PTAR experienced by susceptible people. This is not in keeping with rate statistics that encompass all individuals in their denominator. Additionally, the double time-conditioning appears to miss out space and does not make sense for a spatiotemporal statistic. Practically, conditioning on both time conditions would also result in pairs being counted in the denominator within only a narrow time spectrum.

### 2.4.4 An incorrect time-form odds estimator?

The previous time-form in Azman et al. [8] was formulated as:

$$\begin{aligned}
\hat{\theta}(t_1,t_2) &= \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{1}(z^{(d)}_{ij}=1, d_l \leq d_{ij} < d_m)}{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{1}(z^{(d)}_{ij}=0, d_l \leq d_{ij} < d_m)} \equiv \hat{\theta}(d_l,d_m)\\
\text{where } z^{(d)}_{ij} &= \begin{cases} 1, & \text{if } t_{ij} \in [t_1,t_2) \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{2.9}$$

which is algebraically identical to the distance-form for the odds estimator (Eqn. 2.2), apart from the definition for $z^{(d)}_{ij}$, which becomes variable as $t_1, t_2$ change roles from fixed parameters (previously $T_1, T_2$) to exploratory variables. From the literature review, the first version of the time-form in Azman et al. [8] as an odds estimator $\hat{\tau}^{(t)}_{\text{odds}}$ (Eqn. 2.9) was found to have logical inconsistencies relative to that proposed here (Eqn. 2.7) based on a space-time variable switch: Azman et al. may have used this formulation, as mirrored in `IDSpatialStats` *R* package [116], since no tau-time functions were then available—a pull request is planned to remedy this [166]. The 'corrected' tau-time statistic here will be compared to Azman et al.'s version in §4.3.

## 2.5 Literature review

This review aimed to understand the use & implementation of the tau statistic $\tau$ & related statistics relevant to statistical inference since the tau statistic's creation in the two original papers [118, 182]; the $\tau$ definition is restricted to that consistently described in both of the papers only. Forming this narrow mathematical distinction of the tau statistic is not for abstract fancy. Without it, future theory will struggle to establish a research base for lack of standardisation & external validity, to the detriment of its users.

### 2.5.1 Search strategy, selection & data extraction

A forward literature search was performed on 10 January 2019 of the two original papers, including articles (full text or abstract), conference abstracts, books, preprints, theses & dissertations in any language. Google Scholar was used to find articles (referred to as 'set B') that cite set A (either original paper [118, 182]) excluding duplicates. Articles that cited set B were also considered, called set C: because set C may have referred to the closest paper of inspiration from set B rather than set A. The code to calculate both $\hat{\tau}_{\text{odds}}$ & $\hat{\tau}_{\text{prev}}$ estimators is available in the `IDSpatialStats`—so active forks were checked from GitHub repositories of `IDSpatialStats` [116, 117]. Google.com was also searched for webpages & blogs about the "tau statistic", with disambiguation exclusions. This review was also announced to some of the previous paper authors (Salje, Lessler, Truelove & Cummings) to inquire about upcoming work in their research groups. Only those who actively used the statistic for analyses were accepted; mere citations to mention a previous clustering result were disregarded.

The remaining works were read fully, and their corresponding authors contacted to clarify missing information; furthermore, following the preprint's release [163], a 'right-to-reply' was offered to them on 1 December 2019 to this commentary of their work. The metadata was extracted to summarise, find similarities & ensure reviewing consistency. It covered disease type; country & setting; study type (cohort/cross-sectional/*etc.*); sampling method for the data; calculation method of the tau statistic; and how they presented results of the tau statistic in text & graphics. For works available online prior to journal publication, it was their later journal version that was recorded for the bibliography.

The sixty-one papers that mentioned the two original papers but did not use the statistic were ignored. There was no active code, webpages, nor blogs about the statistic apart from `IDSpatialStats`. All were peer-reviewed articles/reports except one recent preprint [173]; another reading was repeated on November 2021 for [173] which had become a publication [172] in the intervening time. All peer-reviewed works were from respected journals with a minimum recent impact factor of 2·8. In January 2019, 16 papers were found, including the two original papers that claimed to use $\tau$ in their analyses (Table A.1). Salje et al. [182] saw 15 separate works following it and Lessler et al. [118] saw 10 papers follow (Table A.1:col. 1), totalling 16 papers[1]. A detailed tabulation of metadata extracted for the 16 papers is available in Appendix A.1. The timeline of publication year of these works was 2014 ($n = 2$)[2]; 2015 (*1*); 2016 (*5*); 2017 (*1*); 2018 (*5*) & 2019 (*1*) (Table A.1:col. 1). There were seven that cited both the two original papers [8, 70, 94, 172, 180, 183, 194] and a further seven that cited Salje et al. [182] only [20, 83, 84, 119, 178, 181, 189]. All papers had multiple authors and always involved Salje or Lessler. As the review did not have a direct health-related outcome, it was not eligible for PROSPERO registration.

[1] [8, 20, 70, 83, 84, 94, 118, 119, 172, 178, 180, 181, 182, 183, 189, 194]

[2] $(n)$ = number of works

### 2.5.2 Disease spectrum & study location

These papers covered seven human diseases—chikungunya, cholera, HIV, influenza/influenza-like illnesses/upper respiratory illnesses, measles, pneumonia, and dengue which made the most appearances ($n = 8$) (Table A.1:col. 2). Analyses cover all populated continents except South America & Oceania. Settings included a region of substantial landmass [194], where there were spatial restrictions nearby or through their populations due to rivers [8, 20, 70, 94, 172, 178, 182, 181], major roads [189] or interior walkways [119] (Table A.1:col. 3).

### 2.5.3 Methods of analysis

Only eight were considered $\tau$ papers (Table A.1:col. 4) because the others i) lacked a temporal component in their calculation [194] thus reducing them to a spatial statistic, ii) or instead were a phi statistic $\phi$ concerned with spatiotemporal interaction [182] [20, 84, 172], risk ratio [119, 178, 181, 194] or odds ratios [189]. They are still included to cover a broad spectrum of use cases, based on the authors' belief & intention that it was a tau analysis, which is still valuable to this review.

The $\tau$ has been well tested in a range of infectious diseases exploring person-to-person & vector-transmitted diseases, with short-to-medium SIs and different markers of case relatedness (Table A.1:cols. 2 & 4). The study settings have ranged from urban, peri-urban to rural settings at different population densities (Table A.1:col. 3).

The distance-form $\tau^{(d)}$ was used in assessing spatiotemporal clustering, and estimating the clustering endpoint distance $\hat{D}$. It could also be used for indicative use when beginning a descriptive analysis, as prior information for statistical modelling, or when validating an existing model result as is done in §4.3.4. Two novel alternatives were: a) calibrating an Approximate Bayesian Computation model by adding $\tau$ as a summary statistic to capture global clustering [70]; & b) providing an empirical stopping criterion once a random labelling algorithm had reached a global clustering threshold [194] (Table A.1:col. 5). The same dataset with different distance band set $\mathcal{D}$ can produce non-unique $\tau$ estimates (detailed in §2.5.8 & observed in §3.3.5), highlighting a problem for internal validity and requiring further investigation.

### 2.5.4 How the reviewed papers incorrectly perform statistical inference

It is common for authors to perform a *graphical hypothesis test* to assess the evidence against no spatiotemporal clustering nor inhibition by using visual inspection of a $\tau^{(d)}$ vs. distance graph. As detailed in Pollington et al. [165], all papers incorrectly estimated this range while simultaneously establishing the significance of clustering. Mixing graphical hypothesis testing with parameter estimation is incorrect as the former can only give a binary answer of accept or reject whereas the latter is continuous. Nearly all authors determine the range when the lower bound of the CI touches 95%. Azman et al. [8] take account of the uncertainty in the range of spatiotemporal clustering by requiring that the lower confidence bound has crossed unity over two consecutive distance bands or the median distance when bootstrap samples fall below 1·2. However, this is arbitrary as there is no theoretically-informed correction factor.

Most papers constructed bootstrapped estimates around the point estimate to form a *central envelope* with a particular upper & lower bound according to a series of pointwise CIs; they chose the clustering endpoint distance as where the lower bound of the central envelope first touches $\tau = 1$ [165] (Fig. 3.5a). Salje et al. [182] randomly permuted the time marks $t$ across all cases $\big(\text{with points } (x, y, t)\big)$ to simulate a process with no spatiotemporal clustering nor inhibition. An envelope was constructed

| **Location +** | **tau studies** | **non-tau studies** |
|---|---|---|
| case time* | 3 | 4 |
| case time & serotype | 4 | 0 |
| case time, serotype & MRCA** time | 0 | 1 |
| serostatus or none | 1 | 3 |

Table 2.1: **Frequency of epidemiological variables used in the papers' statistics** to describe transmission-relatedness of pairs, $z_{ij}^{(d)}$. While just 'Location' is represented by the spatial arguments of the estimator functions *i.e.* the $d_l \leq d_{ij} < d_m$ term in Eqn. 2.2, the non-header rows of column 1 represent different $z_{ij}^{(d)}$ formulations. *presentation, admission or symptom onset time of the case. **Most Recent Common (genetic) Ancestor.

about these simulations that straddled $\tau = 1$ to form a *null envelope* to simulate $H_0 : \tau = 1$; where the point estimate first touches the upper bound marks the endpoint (Fig. 3.5b). Again the upper & lower bounds are defined by a series of pointwise CIs. Pointwise CIs are standard to describe the uncertainty in $\hat{\tau}$ however, many authors [8, 83, 94, 118, 182] incorrectly use them for hypothesis testing to assert "statistically significant" [83, 182] results: it is incorrect to scan the graph and search at multiple points along $d$ where the central or null envelope bound of $\tau$ is first crossed and then declare that as the clustering endpoint distance. Since multiple pointwise CIs are compared with $\tau = 1$, this inspection amounts to a series of multiple hypothesis tests that inflates the chance that a true null hypothesis is rejected (type I error) [10].

To correct for these, the following were performed separately a) using global envelope tests [137] for the hypothesis test of no clustering $H_0 : (\tau = 1)$ (§3.2.3) & then b) the horizontal set of points where the bootstrapped simulations $\underline{\hat{\tau}}^*$ first intersect $\tau = 1$ to estimate the range of clustering $\hat{D}$ (§3.2.4) (as kindly suggested by Diggle on 22 October 2019 [57]). The latter also provides a measure of precision for the clustering range, unavailable under the previous methods.

### 2.5.5 Defining pair-relatedness

'Location & time' are the common variables (seen in 5 papers) used to identify probably-related transmission pairs (Tables 2.1 & A.1:col. 4). Uniquely, Grabowski et al. [83] require no temporal, geno- nor serotype information to link pairs—instead they use an implicit temporal relation from 'pre-study/prevalent' to 'during-study/incident' cases. Since a prevalent case is defined as having HIV before the study, and incident cases are those detected during the 19-month study, a temporal relation from prevalent to incident cases can be formed. This may be a useful workaround if explicit onset data is unavailable for a study. They also challenge the assumption that (newer) incident HIV cases better identify probable transmission pairs (for a stronger tau signal) by using prevalent cases instead. The likely explanation for 'prevalent -› incident' case pairings showing higher relative risk within households than 'incident -› incident' pairs is both the low infection risk of HIV-1 via heterosexual vaginal sex in low-income country settings $\big($0·08% per event [30]$\big)$ and the relatively short study, so there would have only been time to observe 'prevalent -› incident' case pair associations.

All authors use case or virus pairs to represent the transmission chain, except Grantz et al. [84], who use death pairings; but this limits what can be inferred about transmission: the distribution of deaths is the convolution of the transmission process (of interest to us) with the infection-to-death process, where the latter would be confounded by local poverty & access to healthcare. However, deaths may be the only practical variable from the initial assessment of an outbreak of an unknown cause.

The binary treatment of temporal relatedness in $\tau$ is crude and could be improved using a weighted treatment around a full SI distribution. Giles et al. [77] also elude to incorporating "uncertainty due to pathogen generation time". If a full SI distribution is unavailable, there may still be smarter ways to employ available summary statistics (mean, median or range) than the commonly-used [0, mean SI] window.

For temporal relation, in order to represent primary transmission pairs, *i.e.* the single, direct transmission event from $i \twoheadrightarrow j$, it is common to choose the time-relatedness interval length for $z_{ij}^{(d)}$ (Table A.1:col. 4) as the single (mean) SI. Table 2.2 compares the time intervals chosen against published SIs. As well as the length $(T_2 - T_1)$ of the interval, the start & endpoints $T_1, T_2$ are of interest too—papers commonly use $[T_1 = 0, T_2 = \text{mean SI}]$. Azman et al. [8] make a nuanced $[T_1, T_2]$ selection for interpretative purposes. Initially, they chose [0, 5d (days)]—sensible as cholera (their disease of focus) can have an IP as short as a few hours [89]. However, they switch to [1, 5d] to show the elevated prevalence in cases they could reduce, as it is unrealistic to respond to $i$'s report of onset to mitigate the same-day onset of $j$. Studies seldom reference the source of this interval—Azman et al. [8] is an exception. Additionally, there is a warning of the poor reliability of published IP parameters—half of the respiratory viral studies reviewed did not cite their source [174]. It is not yet known how the time-relatedness interval choice $[T_1, T_2]$ $\big(\text{where } z_{ij}^{(d)} = \mathbb{1}\big((t_j - t_i) \in [T_1, T_2]\big)\big)$ biases the tau statistic through inclusion of extraneous co-primary (an offspring of the parent of the index case) or secondary cases. The transmission contamination from co-primaries $i^* \twoheadrightarrow j^*$ (where $i^*$ shares a parent infector with $i$) or early secondary transmission $j \twoheadrightarrow k$, could bias the spatiotemporal signal of primary transmission $i \twoheadrightarrow j$ [165]. These effects could all diffuse temporal clustering and weaken $\tau_{\text{rate}}^{(d)}$.

| Disease | SI chosen | Published value |
|---|---|---|
| Cholera | [0, 5d]($n$ =*2*)<br>[0, 4d](*1*)<br>[1, 4d](*1*)<br>[0, 5d],...,[25d, 30d](*1*) | median 5d, range 1–11d [198, 9] |
| Dengue | Same month [0, 0mo](*1*)<br>[1, 3mo](*1*)<br>[3, 4–30mo](*1*) | mean 15–17d [4] |
| Measles | [0, 2wk](*1*) | mean 11·7d [196], 14·9d [48] |

Table 2.2: **SIs featuring in the reviewed articles compared with published values.** Papers choosing variable times [183] or model-informed times [180] were excluded. Paper frequencies in round brackets. d = days; mo = months; wk = weeks.

The temporal resolution in days, weeks or months was ultimately constrained by the reporting system. Hoang Quoc et al. [94] used data with a temporal resolution similar to the length of the SI (Table 2.2), which is not ideal: as it could miss additional transmission pairs ($i \twoheadrightarrow j$, then $j \twoheadrightarrow k$) as conceivably within the mean 15–17 day IP for dengue [4], a case $i$ may infect its primary case $j$, then $j$ infect $k$, yet at monthly resolution only $i \twoheadrightarrow k$ would be counted as a pair. For those which explicitly reported it, temporal resolution was as follows: cholera 1 day ($n$ =*2*); dengue 1 day (*1*) or 1 month (*2*); measles 1 day (*1*).

Genetic pathogen diversity is a key ingredient in resolving a transmission chain [38] which may lead some to think that $\tau$ analyses using genetic relatedness are restricted to fast-mutating viruses only. However, this concern is focused on models which reconstruct entire transmission chains. For a statistic similar to the tau [181], its coarse binary classifier, including genotypic markers of dengue virus, was sufficient for medium-scale analysis. Genotypic markers for relatedness could face limitations if used in isolation. The pathogen needs to be diverse enough to discern different transmission chains and mutate frequently enough to act as an additional temporal marker if other time data is low resolution. In low mutation rate/low-diversity scenarios, distinct genotypes may only occur through human migration rather than pathogen mutation, which could bias clustering patterns if migrants have different disease risk factors. Understanding the minimal pathogen genetic characteristics required for a tau analysis could encourage its take-up in genetic surveillance.

### 2.5.6 Misclassification

For diseases rapidly progressing from symptom onset to death, like cholera, people who cannot easily access public healthcare facilities may die before receiving those cheap & simple treatments. There may be misclassification if the case definition $\big($*i.e.* acute watery diarrhoea at any age [89]$\big)$ shares signs or symptoms with other infections, *e.g.* E.coli or shigella—if infection control prevents these too, then potential reduction in $\tau^{(d)}$ at close distances will be overestimated. Spatial misclassification may also arise if infectious contact occurs via alternative processes. A study of military recruits [119] considered bed location in sleeping quarters as the spatial unit. A crowded mess hall could also efficiently spread respiratory pathogens as a hypothetical alternative. If the latter was true, then using bed location would introduce error and weaken the clustering signal. This underlines the importance of follow-up field epidemiology regarding the potential *places* of infection and how data-driven approaches using $\tau$ could be vulnerable to flawed conclusions.

For an outbreak investigation of an unknown pathogen, $\tau$ as a global statistic could help evaluate infectious disease hypotheses, as the mean SI is obtainable from the inter-peak time differences of the empirical epidemic curve. However, early on, when case numbers are low, data scarce & case definitions broadly-defined, developing hypotheses using local spatiotemporal statistics is more appropriate for cluster detection to instigate epidemiological investigation [187].

### 2.5.7 Coverage of the tau estimators

The odds ratio estimator for case-only data (Eqn. 2.2) was the commonest because studies typically collect the geolocation of only cases; the distance-form appears in three of the 16 papers reviewed [8, 70, 94] with the lesser-known time-form in two [8, 180]. The prevalence estimator (Eqn. 2.3) appeared in three papers [83, 182, 183]. Despite odds ratios & risk ratios not being mathematically equivalent and some papers using the term 'risk' generically for all disease measures, at low prevalences of $\sim$1%, they are equivalent [49].

The rate estimator $\tau_{\text{rate}}$ defined in §2.3.3 is yet to be used. There was one application of a rate-style risk ratio that varied with distance [119]; this made sense as the epidemiological unit was respiratory illness events—something a person could repeatedly have. However, they did not explicitly account for variable PTAR, presumably because all participants stayed for the whole study period. For instance, for $\pi(d)$ (the numerator of $\tau_{\text{prev}}$), Levy et al. used the probability of finding sick pairs within distance $d$ out of all sick pairs, rather than the probability that pairs found within $d$ are sick, while their denominator for $\tau$ was the proportion of pairs within $d$ rather than the proportion of sick pairs with $d$ compared to all pairs. Similarly, others make $\tau$ the ratio between seroconverted and all individuals [178, 181] or cases & non-cases [189] rather than between the risk/odds of finding a case within a distance versus at all distances.

### 2.5.8 Distance band set choice, $\mathcal{D}$

The tau statistic may, in theory, be definable at a specific single distance lag & time lag from a case to describe an instantaneous relative proxy of risk. However, it could never be estimated for a real dataset as one only has a finite collection of points in spacetime. Apart from household transmission ($d = 0$), a given space-time lag combination, if sufficiently narrow, would return no pairs in the denominator of an estimator function and thus an undefined value. One therefore has to settle for coarse distance bands. This thought experiment calls into doubt if a 'true' tau statistic exists & is unique. Unless a mathematical proof can show asymptotic convergence to a particular limit for infinite data points as the distance band's width tends to zero. It is also telling that even if the transmission tree is known $\big($so that $z_{ij}^{(d)}$ represents *definite* transmission pairs, not probable ones$\big)$, the estimate is still dependent on the distance bands chosen.

The distance from an average case $i$ can be represented by a half-closed annulus with distance band $[d_l, d_m)$ or as an open disc $[d_l = 0, d_m)$. The choice depends on the purpose of the analysis. An annulus will give a more precise estimate closer to some 'instantaneous' $\tau$, but conversely, as narrower distance bands contain fewer pairs, $\tau$ will become more variable and lead to $\tau$ graphs that are spikier with indiscernible trends. Alternatively, an open disc conveys the excess odds/prevalence/rate up to a said distance $d_2$ for use by policymakers: it mirrors the ACD operational radius, *i.e.* up to a fixed distance from an index case $[0, d_2)$, rather than an impractical annulus shape. If an expanding disc is chosen, one sets $d_l = 0$, relabels $d = d_m$ and writes $\tau(d_l, d_m)$ as $\tau(d)$ instead. However, open discs smooth out any intermediary spatiotemporal structure like village-to-village. Smoothing can be accentuated by allowing distance band overlap [8, 118, 194]. As $d_2$ increases, annuli will cover more pairs, so the series of point estimates cannot be said to represent independent pairings as distal distance bands will contain those that also formed proximal pairings. Additionally, the overlapping will artificially reduce the estimator's variance with greater distance which is detrimental to the performance of global envelope tests [10] $\big($essential for graphical hypothesis testing in [165]$\big)$. Setting bins with equal numbers of pairs may solve this. The same problem in the choice of the time band set $\mathcal{T}$ has been mostly ignored here; for that choice we choose time bands that match the time resolution of the data (*e.g.* months in Chapter 4).

### 2.5.9 Study length, region size & data quality

The most common study type was the cohort, ranging from 3 months to 5 years with median of 15·5 months (Table A.1:col. 6); two studies were cross-sectional. Most case definitions were of a clinical standard beyond those typical for surveillance (Table A.1:col. 4). The spatial resolution of data was constrained by GPS receivers, *i.e.* ~10m (Table A.1:col. 6). When relying on self-reported street addresses for geolocation, a follow-up household visit estimated large spatial errors of 110m–1km [94, 182]. Furthermore, precision could be lost when cases were aggregated at a higher spatial level because of gridded population data [70] or too few cases during the study [20].

### 2.5.10 Statistical characteristics

Only one temporal, serotype or genotype relatedness metric is needed to infer related pairs. However, through epidemic simulations, Lessler et al. have found that more metrics will better identify true transmission pairs so that the range of clustering will less resemble the area of *elevated prevalence* and more the area of *elevated risk*, thus reducing the range of clustering and increasing the magnitude of $\tau$ in this region [118]. They have also shown that the tau statistic is robust to population spatial heterogeneities, correctly identifying no clustering in a spatially-clustered population, unlike the pair correlation function. For a simulation using a serotype or genotype relatedness indicator in addition to onset time relation, the tau statistic consistently estimates the range of clustering when only a random 1% of cases are observed or if there is spatial observation bias, *e.g.* around a surveillance outpost. This is because it is "robust to heterogeneities in sampling probability over a study area, as the probability of sampling will similarly affect both the numerator and the denominator" [118]. It is currently unknown how low the incidence can drop before a spatiotemporal signal cannot be measured.

### 2.5.11 Graphical presentation

Most tau statistic papers use a $\tau$-versus-distance graph $\big($or a $\tau$-versus-time, seen in [8]$\big)$ to show the magnitude of $\tau$ varying with distance or time. As mentioned in §2.3.1, the convention to plot $\tau(d_l, d_m)$ at the midpoint of the distance band for the diagnostic/indicative plot, *i.e.* $d = \frac{1}{2}(d_l + d_m)$ like in Lessler et al. [118], may be misinterpreted if not explained in the caption. The ideal default would plot at the end of the distance band $d_2$ instead unless the graph is used to estimate $\hat{D}$ or $\hat{T}$ then the midpoint makes sense. The number of bootstrap samples chosen had a wide range: 100 (*2*), 500 (*6*), 1,000 (*4*), 10,000 (*1*) or unknown (*2*). Pollington et al. [165] uses 2,500 simulations for global hypothesis tests or to get the distribution of the endpoint of clustering, or 100 bootstraps (or 2000 in Chapter 4) for pointwise CIs. The general use of two continuous lines to represent the upper & lower parts of a series of pointwise CIs is misleading. Therefore, plotting point estimates, each with an error bar $\big($*e.g.* Salje et al. [183]$\big)$ can direct the reader to consider each in turn.

For within-household transmission, the spatial aspect of the infection process is no longer modelled as household members have no spatial freedom where they live as their house is represented as a point. It may therefore be misleading to plot a line joining $[d_1 = 0, \tau(0, 0)] \leftrightarrow [d_2, \tau(d_1 = 0, d_2)]$ unless the first distance band includes non-zero distances *i.e.* $d_1 \geq 0$.

Plotting the tau axis on a log scale can aid the identification of the curve's structure. However, a log-distance axis scale may affect accurate $\hat{D}$ or $\hat{T}$ determination. Some plot more than three tau lines on the same graph [8, 178], making it difficult to discern error bars—aligned panel plots are a better alternative. The graph should cover the full extent of both bounds of the confidence interval (CI). The axes' lines should meet at the origin so that the reader can easily read off values. The horizontal line for $\tau = 1$ is always helpful. The figure's caption or legend should note the tau estimator, distance- or time-form, envelope type, number of bootstrap samples [185], distance or time band set & definition for pair-relatedness: since the graph's shape is dependent on these values.

Using panel plots for different distance bands, the time-form $\tau(t_1, t_2)$ can map out a "dynamic risk zone" [8] however the 2D space vs. time tau colour plot in Salje et al. would provide a more compact representation ([182]:Fig. 3)—each pixel represents the tau estimate for a given distance & time lag. For diagnostic purposes, this would be appropriate for a disease of an unknown aetiology where a diagnostic plot for initial explanatory analysis is required because the SI is approximate owing to limited samples. As well as the spatiotemporal signal of primary transmission, it can reveal seasonality (through repeated regular patterns in the temporal axis) & the immunising effect of each serotype [182]. However, like spatiotemporal variograms, the number of pairs separated by long spatial or temporal lags reduces, requiring caution near the plot's extremities.

## 2.6 Recommendations for further quantitative research

One could compare rate & prevalence estimators to see when they differ according to changes in average time-in-study. There may also be increased uncertainty in $\tau(t)$ point estimates for time band sets that extend to this average time-in-study. It is also unclear how best to choose the distance band set $\mathcal{D}$ to reduce both bias & variance in the tau statistic and whether equidistant or 'equi-number' bins should compose them.

For $\hat{D}$ calculation, the first intercept of $\tau(d)$ with $\tau = 1$ is required. However, in endeavouring to get a 'good' estimate of $\hat{D}$ as defined by the interpolated intercept of a decreasing connected tau series first intersecting $\tau = 1$, how does the configuration of the distance band set matter—expanding disc $[0, d_m]$ or non-overlapping band $[d_l, d_m)$?

Further investigation into the use of the tau statistic as a (global) spatial summary statistic for Approximate Bayesian Computation (ABC) [70] could reveal to what extent the tau statistic can help with computation accuracy & efficiency and how it should be weighted relative to other summary statistics within the ABC algorithm.

Although Lessler et al. [118] has shown the tau statistic to be robust to utilising just 1% of the original data, what is the minimum *number* of cases for the tau statistic to perform reliably? This is particularly apt for data from a near-elimination setting. Lessler et al. [118] have also shown the robustness of the tau statistic to clustering of the underlying population. However, is the tau statistic prone, like other spatial statistics, to *population shift bias* (where the population changes over space & time) [123]?

Diseases with an effective reproductive number, ("the average number of people someone infected at time $t$ can infect over their infectious lifespan" [72]) $R_e(t) > 1$, will overestimate the clustering range while underestimating the magnitude of the $\tau$ in the true region of clustering [118]. For epidemic settings in Fulbaria & Bihar as described in Chapters 4 & 5, we would expect $R_e(t)$ to cycle around $R_e(t) = 1$. For Chapter 4, it is hoped that the tau analysis over the nine years would average out and lessen the underestimate/overestimate to $\tau$ & $\hat{D}$, respectively, that is expected to occur during $R_e(t) > 1$. Through simulation studies and an $R_e(t)$ profile for the study period, further research could develop a corrected tau statistic, especially for epidemic settings. Additionally, differences in health status or treatment-seeking/healthcare could change the disease's latent period or infectious period, respectively. Would this require a reappraisal of the time-relatedness interval over the course of the study?

Immunity from disease exposure had a sizeable biasing effect on the estimation of the mean transmission distance of their simulated epidemics [179]. It would be sensible to systematically assess tau statistic performance for immunising (SIR-style), waning (SIRS-style) & non-immunising (SIS-style) diseases on different estimators—for VL the immunising & waning models would be relevant where the susceptible (S) compartment represents those who have had infection (I) and R are those who have recovered from disease, mostly following drug treatment. Diggle has suggested (personal comm.) to validate the tau statistic against reference point processes (*e.g.* homogeneous Poisson point, Cox or Poisson cluster). It is also unclear to me from reading the literature, the subtle difference between spatiotemporal clustering & interaction—how these two phenomena are measured and their interpretation & implications for infectious disease dynamics.

## 2.7 Discussion

Clustering analysis can characterise infection dynamics and inform ongoing disease control & academic research. The tau statistic $\tau$ has been applied for this purpose to disease datasets containing the location of cases (& sometimes non-cases) and variables linking probable transmission pairs by temporal, serological or genotypic attributes over a variety of settings. Knowledge about $\tau$ has thus far been concentrated in the medium of academic journals and limited to papers written by authors of the original papers—Salje & Lessler. However, this practical statistic could be useful to many infectious disease modellers & epidemiologists, particularly with the open-access `IDSpatialStats` *R* package. To boost adoption, as part of outreach activities, *R Markdown* tutorials of the tau statistic are planned for open-access training hubs like RECONlearn.org.

All papers used incorrect methods for graphical hypothesis testing & parameter estimation. A research gap was identified in choosing the time-relatedness interval to relate case pairs or define the distance band set. Some applications of the tau statistic used nuanced data or time relation variables, which enriches future analysis options. There is still a gap in systematically comparing $\tau$'s properties with other modern statistics [165]—namely, the spatiotemporal $K$ function [73]. Some of the inconsistencies in how the tau statistic or its derived estimates have been defined & interpreted since its inception are now the focus of improvements in the next chapter.

CHAPTER 3

# Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic

*Motivated by the literature review in Chapter 2 & that crucial conversation with P.J. Diggle, a fundamental flaw was corrected in tau analysis, which conflated graphical hypothesis tests with parameter estimation of the clustering endpoint distance. This chapter tests that improvement using an open-access dataset of a widely-studied historical measles outbreak with the odds ratio estimator* $\tau^{(d)}_{\text{odds}}$*. Code from Lessler et al.'s earlier analysis provided a baseline comparison. Point estimation methods are found to heavily bias disease clustering range estimates and spatial bootstrapping schema impact their precision.*

## Abstract

Different factors are tested that could affect graphical hypothesis tests of clustering or bias clustering range estimates based on the statistic by comparison with a baseline analysis of an open-access measles dataset. From re-analysing this data, the spatial bootstrap sampling method used to construct the CI for the tau estimate & CI type is found that can bias clustering range estimates. The bias-corrected and accelerated (BCa) CI is suggested as essential for asymmetric sample bootstrap distributions of tau estimates.

Statistical evidence is found against no spatiotemporal clustering & no inhibition, $p$-value $\in [0, 0{\cdot}022]$ (global envelope test). A tau-specific modification of the Loh & Stein spatial bootstrap sampling method is developed, which gives bootstrap tau estimates with 24% lower sampling error and a 110% higher estimated clustering endpoint than previously published (61·0m vs. 29m) and an equivalent increase in the clustering area of elevated disease odds by 342%. This difference could have important consequences for control. Correct practice of hypothesis testing of no clustering and clustering range estimation of the tau statistic are illustrated in the Graphical abstract (Fig. 3.1). Properly implementing this helpful statistic is advocated to reduce inaccuracies in control policy decisions made during disease clustering analysis.

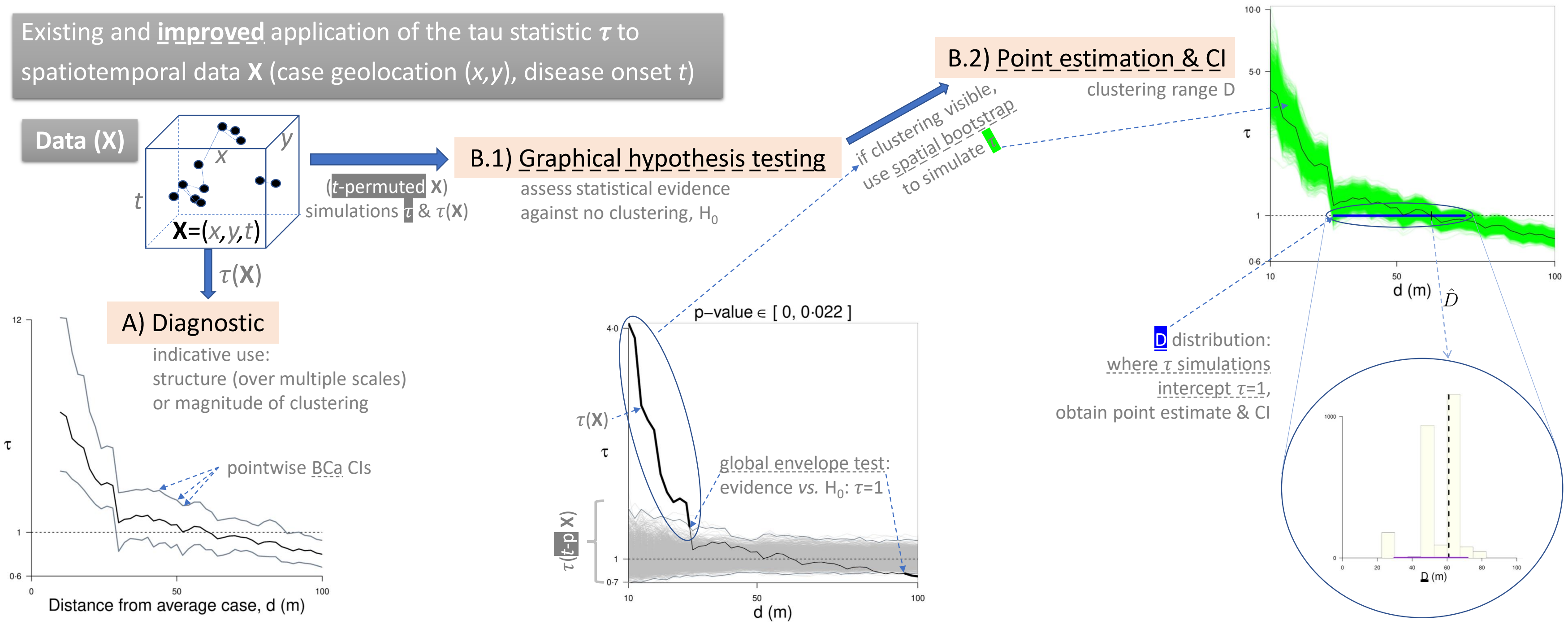


Figure 3.1: **Graphical abstract: Application of the tau statistic to spatiotemporal data. Second version involving $\hat{T}$ shown in Fig. 4.1**. Starting in the top-left with data $\boldsymbol{X} = (x, y, t)$ consisting of cases/non-cases with geolocations $x, y$ and cases with onset $t$. One can **A)** apply the distance form of the tau statistic to data $\tau^{(d)}(\boldsymbol{X})$ to produce a diagnostic/indicative plot (bottom-left plot) to explore spatiotemporal structure over multiple scales or the magnitude of clustering. The confidence envelope is composed of pointwise BCa CIs. Alternatively, spatiotemporal clustering can be assessed through **B.1)** a graphical hypothesis test by plotting $\tau^{(d)}(\boldsymbol{X})$ together with global envelopes constructed on null simulations of time-permuted data $\tau^{(d)}\big((x, y, t\text{-permuted})\big)$ (middle-bottom plot). Where $\tau^{(d)}(\boldsymbol{X})$ exceeds the global envelope provides evidence against the null hypothesis $H_0$ of no spatiotemporal clustering nor inhibition. Conditional on clustering being established in B.1, one progresses to **B.2)** estimation of $\hat{D}$ the clustering endpoint distance, to guide policymakers on the spatial range of the elevated burden of disease around cases. The point estimate for $\hat{D}$ is when $\tau^{(d)}$ first intercepts $\tau = 1$ (top-right plot). To obtain an estimate of its precision, one uses spatial bootstrap estimates of $\tau^{(d)}$ and where they first intercept $\tau = 1$ gives the distribution of $\underline{D}$ from which CIs can be constructed.

## 3.1 Motivation

Last chapter's review of the tau statistic's use found that its present implementation inflated type I errors (incorrectly rejecting a true null hypothesis) when testing for clustering and may have biased estimates of the range of clustering [163]. This motivates an investigation into these aspects by analysing a well-studied open-access measles dataset containing household geolocations & symptom onset times of cases (§3.2.1). It represents a *spatially discrete process* since infection is only recorded & can only occur at discrete household locations, so the (statistical) support is not spatially continuous [59].

An ordered approach is adopted: one first tests for clustering (§3.2.3) and then, conditional on finding evidence against 'no clustering' (nor inhibition), the *clustering range* is estimated (§3.2.4). The first precision estimate for the clustering range is also provided (Fig. 3.8). It is hoped that these improved methods will encourage the proper application of this burgeoning statistic.

$* \ \tau \ *$

In the following sections (§3.2–3.3), a descriptive analysis of the data is provided before systematically testing several aspects of the tau statistic's implementation and their impact on the estimated clustering range, $\hat{D}$. Throughout this chapter we solely use the distance-form odds estimator of the statistic $\tau^{(d)}_{\text{odds}}$, to enable comparison with the original Lessler et al. analysis; tau nomenclature in formulae remains generic in the main text (*e.g.* $\tau$) as these inference steps apply to all estimators, and odds-specific (*e.g.* $\tau_{\text{odds}}$) when discussing actual results.

## 3.2 Methods

### 3.2.1 The dataset & baseline analysis

An infectious disease dataset is analysed of measles from case households in Hagelloch, Germany, in 1861 [130, 138, 144, 153]. The epidemic over a small ∼280m × 240m area lasted nearly three months, and five distinct generations can be discerned from the epidemic curve (Fig. 3.2). Out of the 197 under-14-year-olds, 185 became infected, along with three teenagers, leaving 377 remaining teenagers & adults uninfected [138]. Figure 3.3 indicates a weak signal of direct transmission between cases, as cases with onsets close together in time (shown by similar colours) tended to be spatially near to each other. The minimum inter-household separation was 7·9m which is feasible for terraced small dwellings. It is unknown what source Oesterle [144] used for the residential coordinates of cases—the choice of locum (front door of the household facing the street or building centroid) if inconsistent would contribute to spatial sampling error, let alone the unknown place of transmission which may not have been the home.

In setting the temporal relatedness at $[0, \text{mean SI of measles}]$ we aim to pick up the primary transmission chains between infectors & infectees whose symptom onset dates we have from the data. As section §2.5.5 explains, this selection is imperfect as unrelated transmission chains will also be included. Furthermore, although reasonable, the choice of $[0, \text{mean SI}]$ has not yet been demonstrated in simulations regarding its sensitivity & specificity for picking up the primary transmission chains only.

Computations were run in *R* using RStudio [169, 177] (Appendix B.1.1). Lessler et al.'s (unpublished) analysis has been reproduced to act as a baseline result (Fig. 3.4). Using *their* interpretation of

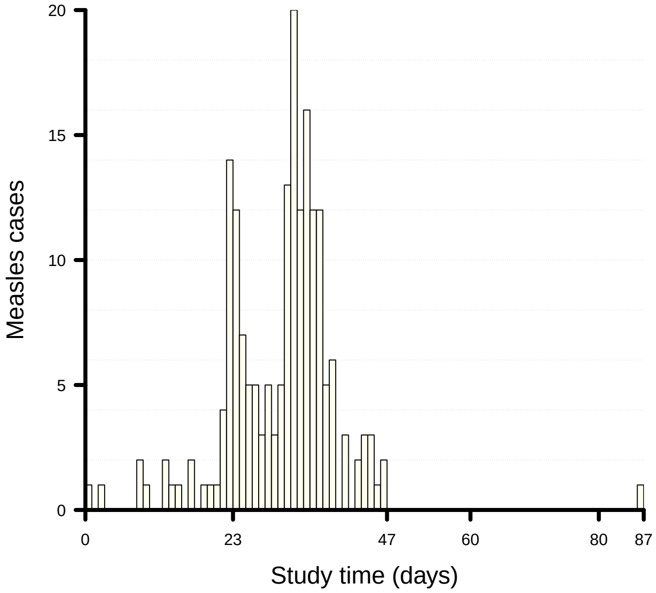


Figure 3.2: **Epidemic curve of the 188 measles cases in Hagelloch in 1861**.

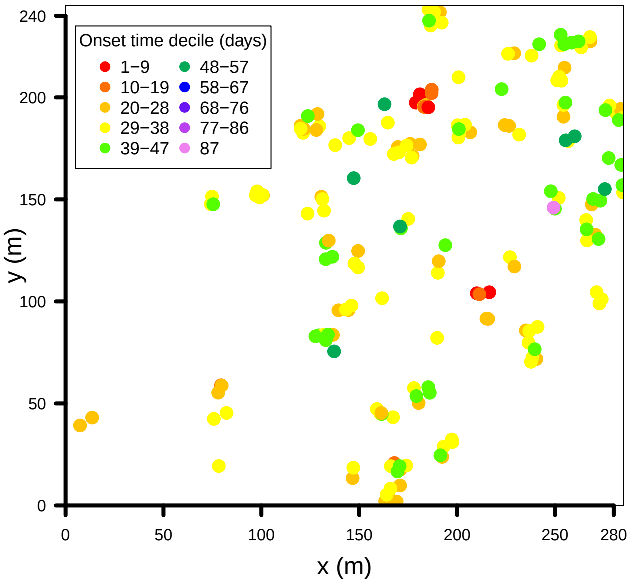


Figure 3.3: **Spacetime points of cases' locations with onset times as colour marks**. Cases jittered up to 5m separately in $x$ & $y$ dimensions using the Uniform distribution to show multiple case households. There is some indication of cases in nearby households (∼50m apart) having a similar onset date, which may indicate direct transmission up to this distance.

Figure 3.4, spatiotemporal clustering is reported up to 30m [118]. The analysis code in *R Markdown* is available from `github.com/t-pollington/developments_tau_statistic`.

The tau-distance graph in Fig. 3.4 also reaches tau values below 1. There is not an identifiable reason for this: it could be a spatiotemporal inhibition process over these spatial ranges or the natural necessity that spanning from 0 to the maximum pairwise distance of the data would explore values below 1 as well as above it. This is because the tau distance is a statistic normalised by its denominator, i.e. $\theta(0, \infty)$ for the odds estimator. $\theta(0, \infty)$ acts as the average baseline across all distance pairs. Necessarily, if there are distance bands where $\theta(d_1, d_2)$ is higher than $\theta(0, \infty)$ thus giving $\tau(d_1, d_2) > 1$, there must be distance pairs to which it is lower—not all distance pairs can be above the average due to the definition of an average. Note in practice tau-distance would not be extended to this maximum due to the lack of pairs to evaluate and the considerable imprecision in tau as a result.

### 3.2.2 The approach to hypothesis testing & parameter estimation

An *envelope* is loosely defined as a series of piecewise linear (*syn.* connected-line) functions in the Cartesian plane, with some bound applied above & below. *Central/null envelopes* describe the line function, *i.e.* whether it originates from simulations of a bootstrapped point estimate or time-permuted null distribution, respectively, whereas *global envelope* or *pointwise CI* (*syn.* confidence band) refer to the way function lines are bounded. A global envelope is a CI for a series of line functions but does not represent a single distance band of one tau point estimate $\hat{\tau}(d_l, d_m)$ (*i.e.* a pointwise CI), but rather the entire distance band set $\underline{\mathcal{D}}$. At say a 95% significance level, in 95% of outcomes of constructing a global envelope, the random envelope would contain the true value of $\tau(d_l, d_m), \forall\ [d_l, d_m) \in \underline{\mathcal{D}}$ [10].

The graphical hypothesis test (§3.2.3) & parameter estimation (§3.2.4) methods (Fig. 3.1) offer corrections to the implementations of the tau statistic or similar statistics reviewed in Chapter 3: [8, 20, 83, 84, 94, 118, 119, 173, 178, 180, 181, 182, 183, 189, 194]), which incorrectly used an envelope about the point estimate constructed from pointwise CIs to estimate the clustering endpoint $\hat{D}$ as the distance at which the lower bound of the first pointwise percentile CI above $\tau = 1$, touches $\tau = 1$ (Fig. 3.5a) [163], or where the connected point estimate line first intercepted the upper bound of the null envelope (Fig. 3.5b). Either error amounts to multiple hypothesis testing and inflates type I errors.

### 3.2.3 Graphical hypothesis test of no clustering

Instead, a *global envelope* is constructed around the null hypothesis distribution ($H_0$: $\tau = 1$, no spatiotemporal clustering nor inhibition) [135]. This is generated by randomly permuting the time marks $t_i$ of the spatiotemporal data points $X_i$ = (x-coordinate$_i$, y-coordinate$_i$, onset time$_i$) to scramble any spatiotemporal clustering present and simulate what $\hat{\tau}$ would be under $H_0$. It is assessed whether a subset of distance bands $\underline{\delta}$ of $\underline{\mathcal{D}}$ exists (as contiguous or disjoint regions) where the tau point estimate $\hat{\tau}(d)$ is ever above/below the upper/lower bound, respectively, of this (global) null envelope. This null envelope is of extreme rank type ("defined as the minimum of pointwise ranks") with 95% significance level & extreme rank length $p$-value interval (note: a range, not a single $p$-value) [136]; as constructed by the `GET` *R* package [136](Fig. 3.1). The test is two-tailed, which is necessary as only once the graph is plotted is the presence of clustering or inhibition known (alternative hypothesis $H_1 : \tau \neq 1$). 2,500 'time-mark permuted' tau simulations are computed for an optimal test [137].

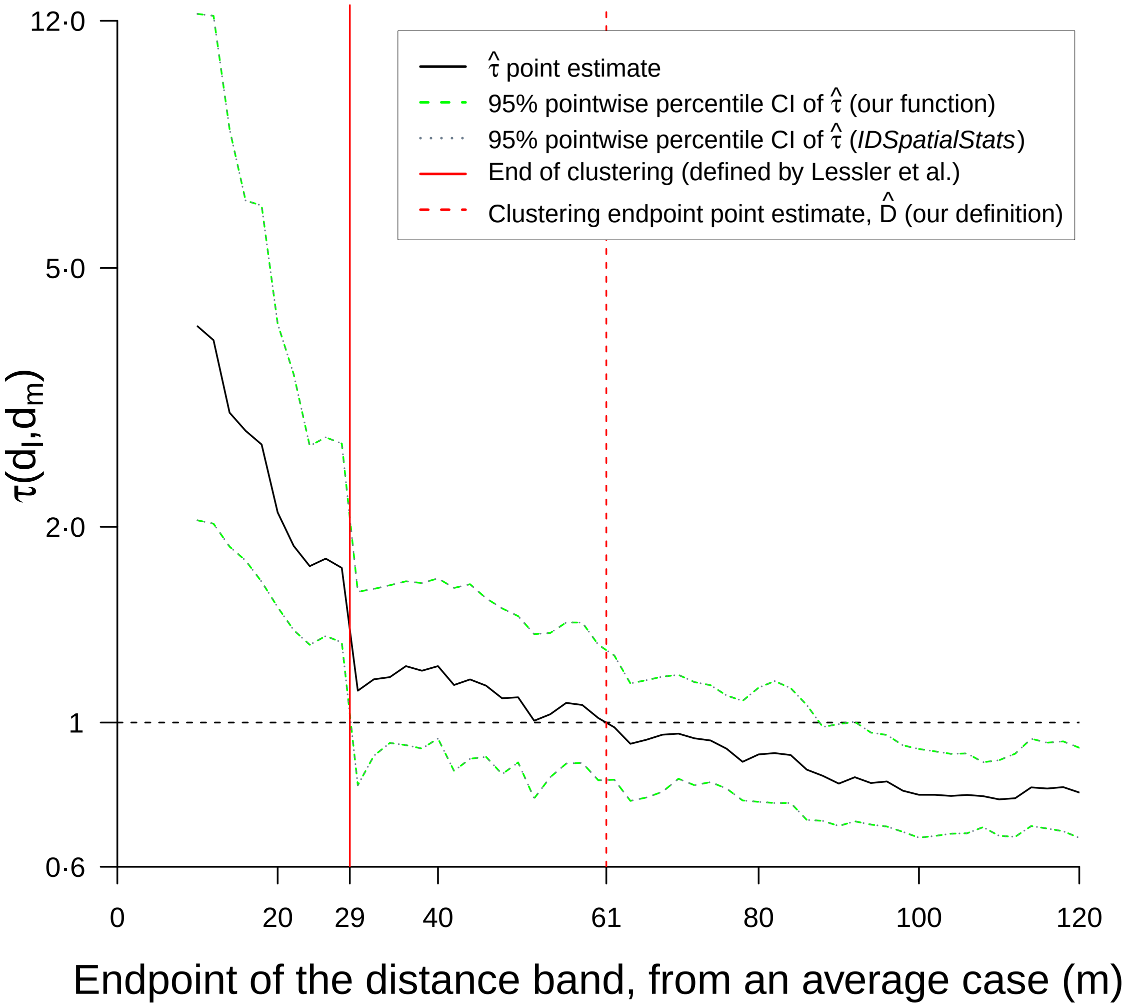


Figure 3.4: **Baseline result: a reproduction of a previous analysis using** $\tau^{(d)}_{\text{odds}}$ [118, Fig. 4C]. Note that that the end of the clustering range reported by Lessler et al. is where the lower bound of the envelope intersects $\tau = 1$ ($\hat{D}_{\text{base}} = 29\text{m}$) (this convention, however, is *not* endorsed). Regardless, as the horizontal axis is the midpoint of the distance band (*i.e.* $(d_1 + d_2)/2$), [0, 30m) is the actual clustering range that would be interpreted using their convention, as confirmed by Lessler (personal comm.). The near-perfect superimposition of their envelope and that defined here validates the implementation of tau functions herein compared to their `IDSpatialStats` *R* package. 100 bootstraps per pointwise CI. Measles cases are considered temporally related within [0,14d]. Distance band set from Lessler et al.: $\{[0,10), [0,12), [0,14), \ldots, [0,50), [2,52), [4,54), \ldots, [74,124\text{m})\}$

### 3.2.4 Parameter estimation of the clustering range, $\hat{D}_{\text{odds}}$

If hypothesis testing establishes the evidence against no spatiotemporal clustering within a subset of distance bands $\underline{\delta}$ (§3.2.3), it is then sensible to estimate the *endpoint of spatiotemporal clustering* $\hat{D}$ for the clustering range $[d_1 = 0 \text{ (assumed)}, d_m = \hat{D})$ where the point estimate intercepts $\tau = 1$,

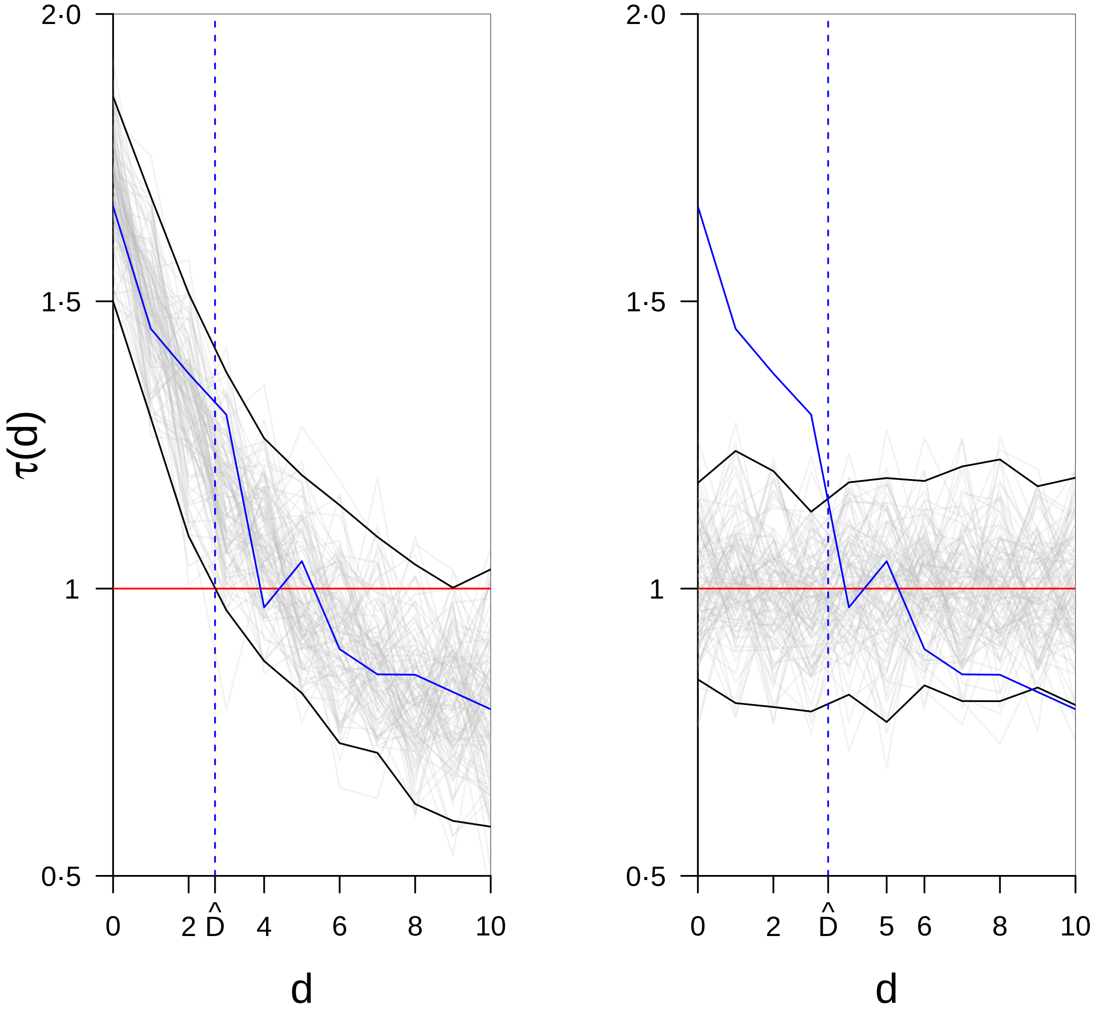


Figure 3.5: **Illustrated example: Naïve methods conflating graphical hypothesis testing & point estimation of $\hat{D}$** (see §3.2.2)—choosing one envelope type as 'central' (a) or 'null' (b), then simultaneously testing the hypothesis of clustering and estimating the range of clustering parameter $\hat{D}$ [163]. The single red line $\tau = 1$ represents no spatiotemporal clustering nor inhibition. Grey lines indicate a) negative exponential lines with Normal noise to characterise a series of spatial bootstrap estimates $\hat{\tau}^*$ of a typical tau function, or b) a line at $\tau = 1$ with Normal noise to represent simulations of $\tau = 1$ for null envelope construction; black lines mark out the envelope bounds. The solid blue line characterises an empirical tau point estimate $\hat{\tau}(d)$. Instead, the method is split into separate hypothesis testing and parameter estimation steps in §3.2.3 & §3.2.4, respectively.

*i.e.* $\hat{D} := \{d : \hat{\tau}(d) = 1\}$. The startpoint of spatiotemporal inhibition is calculable (Appendix B.1.3) but is not of interest here as the main motivation is for parameters that have practical relevance to disease control. Due to discrete distance bands, one linearly interpolates between the midpoint of distance band $[d_l, d_m)$ of the last $\hat{\tau}$ above one, and that of the next $\hat{\tau}$ below one $[d_{l+1}, d_{m+1})$, to obtain $\hat{D}$.

To calculate the uncertainty of $\hat{D}$, one uses bootstrapped tau estimates $\underline{\hat{\tau}}^*$. For each bootstrapped simulation (that represents a connected line of simulated tau estimates for increasing $d$, *i.e.* $\{\hat{\tau}^*(d_l, d_m) : [d_l, d_m) \in \underline{\mathcal{D}}\}$), one records those that originate from above $\tau = 1$ and then intersect $\tau = 1$ at some greater distance $D$, *i.e.* those for which there exists $D$ satisfying $\hat{\tau}^*(D) = 1$. $N = 2{,}500$ samples are used, which is more than sufficient for a typical bootstrap sample [67]. This horizontal set of values $\underline{D}$ is then taken and used to obtain a CI to describe the uncertainty in $\hat{D}$ (Fig. 3.1). The research now focuses on spatial bootstrap methods (§3.2.4.1), CI construction (§3.2.4.2) & distance band sets (§3.2.4.3).

#### 3.2.4.1 Spatial bootstrap sampling methods for $\hat{\tau}$

To construct a central envelope of $\hat{\tau}$ to obtain $\hat{D}$, one needs to generate a non-parametric spatial bootstrap distribution of tau estimates, $\underline{\hat{\tau}}^*$. Through bootstrap theory, the sampling distribution $\underline{\hat{\tau}}^*$ may serve as a proxy for the actual distribution of $\hat{\tau}$ on the data; and further, the envelopes constructed from $\underline{\hat{\tau}}^*$ may approximate the envelope of $\hat{\tau}$ on the data [63]. Three spatial bootstrap methods are compared; all are non-parametric because they randomly resample the data without imposing a distribution [120].

**3.2.4.1.1 Resampled-index spatial bootstrap (RISB)** This first method starts with the spatiotemporal data $\mathbf{X} = (X_i)_{i=1,\dots,n}$ where $X_i = (\text{x-coordinate}_i,\ \text{y-coordinate}_i,\ \text{onset time}_i)$. Using the Uniform distribution, one resamples with replacement the data's indices $\underline{i} = (1, \dots, n)$ $n$ times (equal to the number of cases), to produce a new empirical *spatial bootstrap* sample of indices $\underline{i}^* = (i^*_k)_{k=1,\dots,n}$ & data $\mathbf{X}^* = \left(X_{\underline{i}^*}\right)$ ($\underline{i}$ & $\underline{i}^*$ have the same length, but $\underline{i}^*$ is bound to contain duplicated indices due to sampling with replacement). The tau-odds estimator is computed on each bootstrap sample $\mathbf{X}^*$ to get $N$ bootstrapped $\tau$ estimates $\underline{\hat{\tau}}^* = (\hat{\tau}^*_1, \dots, \hat{\tau}^*_N)$; the same approach could be applied to other $\tau$ estimators. Loh critiques this "naïve" sampling with replacement of the points $X_{\underline{i}}$ of a spatial dataset to produce a spatial bootstrap sample, because "the spatial dependence structure has to be preserved as much as possible" [120] ... "to reflect properties of the original process" [121]. Lessler et al. & others used this method and additionally for any $p$, $q$ resampled indices ($p \neq q$), dropped $(i^*_p, j^*_q)$ pairs where they represented the same point ($i^*_p = j^*_q$) to avoid 'self comparisons' [118].

**3.2.4.1.2 Loh & Stein marked point spatial bootstrap (MPSB) applied to the tau-odds ratio estimator (not recommended)** Loh & Stein's MPSB is a fast, non-parametric method to obtain a bootstrap distribution of a second-order correlation function [121]. For a clustered process simulated by a Matérn process, the CIs constructed using it had higher empirical coverage than other methods and were computed faster [121].

For the RISB (§3.2.4.1), each bootstrap estimate $\hat{\tau}^*$ is computed from resampled (and smaller) spatiotemporal data $\mathbf{X}^*$ containing duplicated points from duplicate indices in $\underline{i}^*$. However, the MPSB instead takes a spatial bootstrap sample of the locally-evaluated $\tau$-functions $\underline{\tau}_i$ (Eqn. 3.1) corresponding to each $i^* \in \underline{i}^*$ across all points $\underline{j}, j \neq i^*$, so each local $\tau_i$ covers all points in $\mathbf{X}^*$ unlike the RISB:

$$\begin{aligned}\hat{\tau}_i(d_l, d_m) &:= \frac{\hat{\theta}_i(d_l, d_m)}{\hat{\theta}_i(0, \infty)} \\ \text{where } \hat{\theta}_i(d_l, d_m) &= \frac{\sum_{j=1, j\neq i}^{n} \mathbb{1}(z_{ij} = 1, d_l \leq d_{ij} < d_m)}{\sum_{j=1, j\neq i}^{n} \mathbb{1}(z_{ij} = 0, d_l \leq d_{ij} < d_m)}\end{aligned} \tag{3.1}$$

The local $\hat{\tau}_i$ functions (Eqn.3.1) computed for the MPSB are similar to applying a spatial bootstrap to the $K$-function [10], which like $\tau$, is a second-order correlation function. However, this is *not* recommended for literal interpretation of Loh & Stein's method of averaging localised $\tau$-functions for the tau statistic, as the MMPSB method explains (§3.2.4.1 & Appendix B.1.4) but is provided here for completeness (Eqn. 3.2).

$$\tau^*_{\text{MPSB}}(d_l, d_m) = \frac{1}{n} \sum_{i^*} \frac{\theta_{i^*}(d_l, d_m)}{\theta_{i^*}(0, \infty)} = \frac{1}{n} \sum_{i^*} \frac{\left(\frac{m_{i*}(d_l, d_m, k=1)}{m_{i*}(d_l, d_m, k=0)}\right)}{\left(\frac{m_{i*}(k=1)}{m_{i*}(k=0)}\right)} \tag{3.2}$$

**3.2.4.1.3 Modified marked point spatial bootstrap (MMPSB)** This third & final method differs slightly from Loh & Stein's MPSB—rather than spatial bootstrapping the local $\tau$-functions (Eqn. 3.2), going deeper one computes the number of related or unrelated *local mark functions* $m_i(k)$, according to their Boolean time-relatedness $k \in \{0, 1\}$.

The number of time-related cases (#related) within a distance $[d_l, d_m)$ around a case $i^*$ chosen in the spatial bootstrap sample is:

$$\#\text{related}(d_l, d_m, k = 1, i^*) \equiv m_{i^*}(d_l, d_m, k = 1) = \sum_{j \in \underline{j}, j \neq i^*} \mathbb{1}(d_l \leq d_{i^*j} < d_m, z_{i^*j} = 1) \tag{3.3}$$

and then an average is taken over the $n$ cases in the spatial bootstrap sample of indices $\underline{i}^*$:

$$\overline{\#\text{related}^*(d_l, d_m)} \equiv m^*(k = 1) = \frac{1}{n} \sum_{i^* \in \underline{i}^*} \sum_{j \in \underline{j}, j \neq i^*} \mathbb{1}(d_l \leq d_{i^*j} < d_m, z_{i^*j} = 1), \tag{3.4}$$

and similar steps for time-unrelated cases yield:

$$\overline{\#\text{unrelated}^*(d_l, d_m)} \equiv m^*(k = 0) = \frac{1}{n} \sum_{i^* \in \underline{i}^*} \sum_{j \in \underline{j}, j \neq i^*} \mathbb{1}(d_l \leq d_{i^*j} < d_m, z_{i^*j} = 0), \tag{3.5}$$

and finally the odds & odds ratio estimator can be calculated as before:

$$\theta^*(d_l, d_m) = \frac{\overline{\#\text{related}^*(d_l, d_m)}}{\overline{\#\text{unrelated}^*(d_l, d_m)}} = \frac{\sum_{i^* \in \underline{i}^*} \sum_{j \in \underline{j}, j \neq i^*} \mathbb{1}(d_l \leq d_{i^*j} < d_m, z_{i^*j} = 1)}{\sum_{i^* \in \underline{i}^*} \sum_{j \in \underline{j}, j \neq i^*} \mathbb{1}(d_l \leq d_{i^*j} < d_m, z_{i^*j} = 0)} \tag{3.6}$$

$$\tau^*_{\text{MMPSB}}(d_l, d_m) = \frac{\theta^*(d_l, d_m)}{\theta^*(0, \infty)} \tag{3.7}$$

For all estimator functions (*i.e.* $\theta$, $\pi$ or $\lambda$) this is equivalent to changing the double summation in each numerator & denominator from $\sum_{i^*} \sum_{j^*}$ under RISB to $\sum_{i^*} \sum_{j}$ under MMPSB. In the case of rate estimators the denominator summation is $\sum_{i^*} \sum_{j}$ and the numerator $\sum_{a^*} \sum_{b}$.

#### 3.2.4.2 Confidence interval (CI) construction

Applying a percentile CI to the sample bootstrap distribution $\underline{D}$ (previously defined in §3.2.4) assumes it is symmetric, which is not the case, especially at short distances (Fig. 3.9) [39]. BCa CIs can cope with asymmetrical distributions better than percentile CIs. For non-parametric problems, Carpenter & Bithell [39] consistently found Efron's BCa method best due to its low theoretical coverage errors for approximating the exact CI. BCa had "second-order correct coverage" errors under some assumptions, while a percentile CI was first-order correct at best [64]. The BCa algorithm transforms a distribution of bootstrap calculations by normalisation to stabilise its variance so that a CI can be constructed, then back-transforms it [64].

#### 3.2.4.3 Distance band sets

The tau statistic is non-unique as it depends on the distance band set chosen [163], so the potential variation in $\tau$ estimates from this choice is of interest. From analysing cases' pairwise distances an arbitrary non-overlapping distance band set is proposed, *i.e.* $\underline{\mathcal{D}} = \big\{$[0,7), [7,15), [15,20), [20,25), [25,30), …, [195,200m)$\big\}$ as a comparison to Lessler et al.'s overlapping set $\big\{$[0,10), [0,12), [0,14), …, [0,50), [2,52), [4,54), …, [74,124m)$\big\}$, and test these using $N$ = 2,500 samples under the MMPSB method.

## 3.3 Results & Discussion

### 3.3.1 Graphical hypothesis tests: global envelopes vs. pointwise CIs

There is moderately strong evidence against the hypothesis of no spatiotemporal clustering & no inhibition $\big(p\text{-value} \in [0, 0{\cdot}022]\big)$ based on constructing the global envelope around $\tau = 1$ under the null hypothesis (Fig. 3.6), and thus it is concluded that the data $\mathbf{X}$ is inconsistent with the null model ($H_0 : \tau = 1$). So one turns to the alternative hypothesis that there is clustering or inhibition. Figure 3.6 suggests clustering at short distances & inhibition at long distances. Unfortunately, these results cannot be compared with those of previous papers (see §3.2.2) since they used an incorrect pointwise CI approach to assess clustering, for which a $p$-value is unavailable.

### 3.3.2 Impact on the estimated clustering endpoint, $\hat{D}$

The estimated clustering endpoint is $\hat{D}_{\text{odds}} = 61{\cdot}0$m with a 95% percentile CI of (29·0, 83·0m) over 100 bootstrapped simulations using RISB sampling (Fig. 3.7), or (29·2, 83·5m) over 2,500 simulations (using 100% of simulations, see Appendix B.1.2); more bootstrapped simulations do not appear to affect the sampling error.

The point estimate $\hat{D}_{\text{odds}} = 61{\cdot}0$m is 110% higher than the baseline clustering range ($\hat{D}_{\text{base}}$ =29m). Previous estimates derived via the improper method of finding the distance at which the lower bound of the central envelope (around $\hat{\tau}^{(d)}_{\text{odds}}$) touches $\tau = 1$ (Fig. 3.5a) underestimated this range. The plateauing shape of $\hat{\tau}^{(d)}_{\text{odds}}(d)$ before it reaches $\tau = 1$ contributes to the increased imprecision in the estimate of $\hat{D}_{\text{odds}}$. This highlights the utility of a human assessing the graph rather than rigidly using a $\tau = 1$ threshold, as it is likely that disease control over, say, a 60m radius around an average case would see the biggest gains over its first 30m with diminishing returns at wider radii (Fig. 3.8).

The 110% increase in the radial parameter $\hat{D}_{\text{odds}}$ (§3.3.2) from using the corrected parameter estimation algorithm (§3.2.4) is important for public health interventions, but more so as their time &

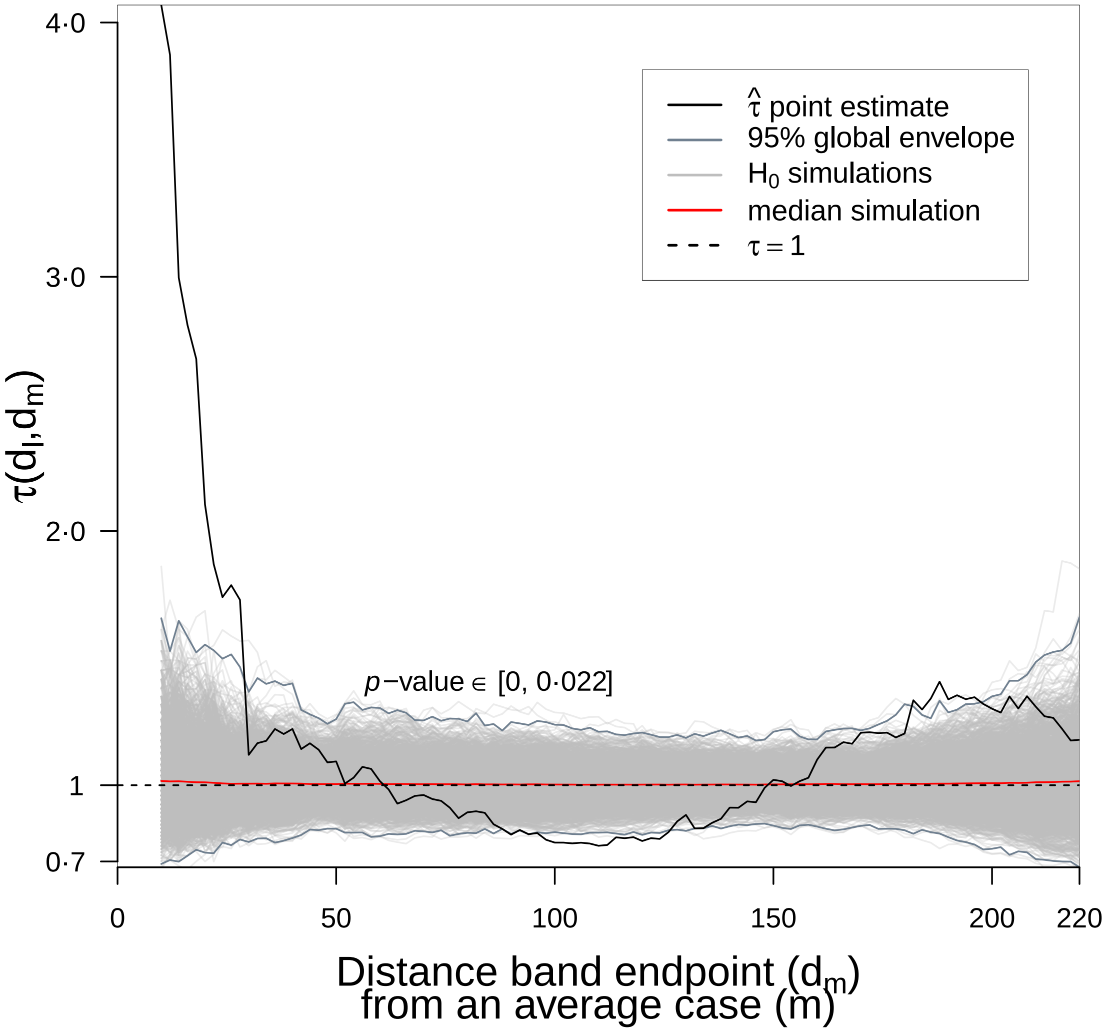


Figure 3.6: **Global envelope test**, 'extreme rank' type, two-sided at 95% significance level using 2,500 simulations of the null hypothesis ($H_0$: no spatiotemporal clustering & no inhibition, *i.e.* $\tau = 1$) for $\tau^{(d)}_{\text{odds}}$. Measles cases are considered temporally related within [0,14d]. Note there is a region where $\hat{\tau}$ just exits the global envelope lower bound (suggesting inhibition at long distances) and the obvious departure above the upper bound (suggesting clustering at close distances). There is the confidence that $H_0$ is being simulated properly because the median simulation stays close to $\tau = 1$ throughout. Distance band set $:= \{[0, 10), [0, 12), [0, 14), \ldots, [0, 50), [2, 52), [4, 54), \ldots, [74, 124\text{m})\}$.

cost is more closely proportional to area, and the areal increase is 342% $\big($since $\pi(\hat{D}^2 - \hat{D}^2_{\text{base}})/\pi\hat{D}^2_{\text{base}} = 3{\cdot}42$, assuming $d_l = 0\big)$.

### 3.3.3 Spatial bootstrap sampling: MMPSB vs. RISB

Using the MMPSB schema (§3.2.4.1) yields a narrower envelope than the RISB, leading to a 95% BCa CI for $\hat{D}_{\text{odds}}$ of (29·8, 71·8m) (Fig. 3.8); both CIs used 100% of simulations.

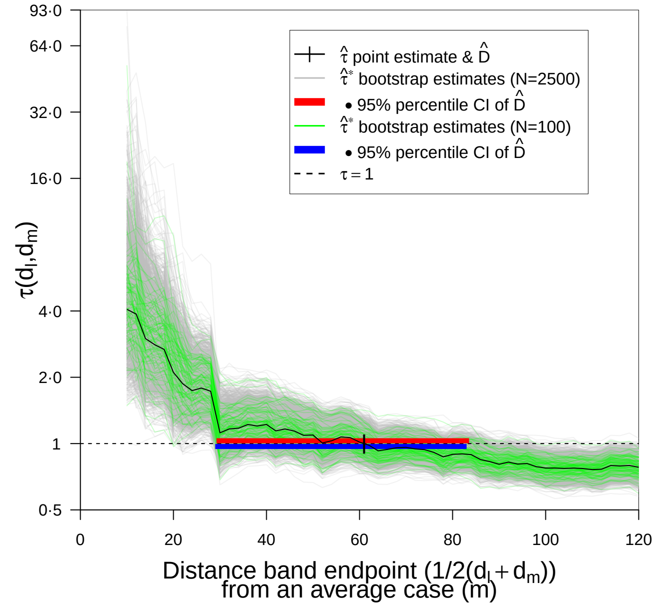


Figure 3.7: **Effect of the number of samples on $\hat{D}_{\text{odds}}$ precision when using RISB sampling**. Both CIs used 100% of simulations. $\hat{D}_{\text{odds}} = 61{\cdot}0$m; $N = 100$: 95% BCa CI (29·0, 83·0m); $N = 2500$: CI (29·2, 83·5m). Distance band set as Figure 3.6. Measles cases are considered temporally related within [0,14d].

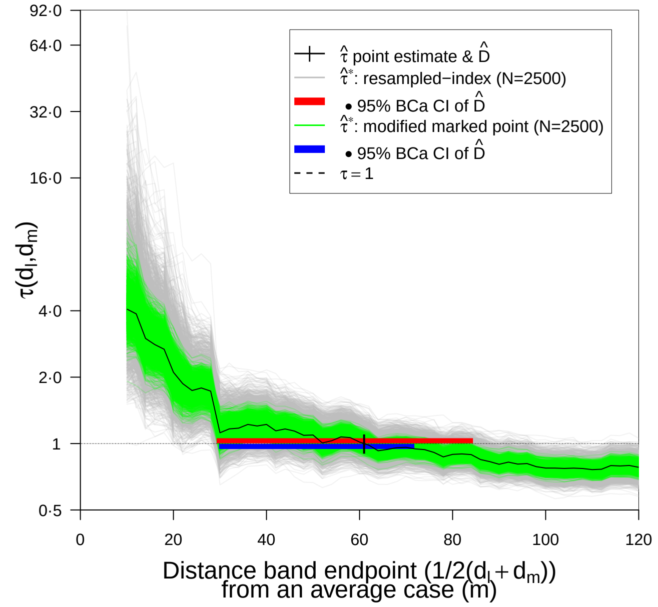


Figure 3.8: **Effect of the spatial bootstrap sampling method on $\hat{D}_{\text{odds}}$ precision**. RISB 95% BCa CI (29·3, 84·4m); MMPSB CI (29·8, 71·8m); both CIs used 100% of simulations. Distance band set as Figure 3.6, $N$ =2500. Measles cases are considered temporally related within [0,14d].

If the tau point estimate had been shallower near the $\tau = 1$ intercept, then the range of spatiotemporal clustering would be far more extensive and the benefit of MMPSB more apparent. It is expected that RISB will underestimate this range, given why MMPSB is better: it outperforms because RISB loses more pair information from resampling indices and avoiding self-comparisons. This was checked empirically for the measles data: the tau point estimate was computed on $188 \times 187 = 35{,}156$ pairs. On average from 1,000 simulations, the RISB sampled from 119 unique people, leading to $119 \times 118 = 14{,}042$ unique pairs evaluated or $\sim$ 39·9% of the original pairs. Of course, many additional duplicate pairs are used in the RISB but one is only interested in unique pair information that is retained. The MMPSB only has 119 unique mark functions, but each is compared with the other 187 cases, leading to 63·3% of pairs being retained.

### 3.3.4 CI type: Bias-corrected & accelerated vs. percentile

Histograms of the asymmetric distribution of $\underline{D}_{\text{odds}} = \{D_i : \hat{\tau}^{(d)}_{\text{odds}}{}^*(D_i) = 1, i = 1, \ldots, N\}$ by the number of bootstrapped samples indicate for both $N = 100$ or 2,500 samples that a percentile CI gives a less precise estimate; both CIs used 100% of simulations (Fig. 3.9). The BCa method provides slightly narrower CIs than the original percentile CIs (Fig. 3.9). The RISB appears to introduce positive skew (mean > median) in $\underline{D}_{\text{odds}}$, whereas MMPSB with sufficient samples ($N = 2500$) introduces a slight negative skew. MMPSB reduces the bias ($\bar{\underline{D}} - \hat{D}$) between mean/median estimates of $\underline{D}_{\text{odds}}$ & the point estimate $\hat{D}_{\text{odds}}$ from ~10m to ~5m.

### 3.3.5 Distance bands

Overlapping distance band sets appear to produce $\hat{D}_{\text{odds}}$ estimates with higher variance (95% BCa CI (29·8, 71·8m)) than non-overlapping sets (CI (18·4, 28·6m)) (Fig. 3.10), but a clearer & smoother trend in tau with increasing distance (both CIs used 100% of simulations). The non-overlapping $\underline{\mathcal{D}}$ also struggles to contain $\hat{D}_{\text{odds}}$ (Fig. 3.10) because the simulations are more erratic about $\tau = 1$, the distribution of $\underline{D}_{\text{odds}}$ is strongly bi-modal, which even the BCa technique struggles with. The increased volatility of $\hat{\tau}$ also results in multiple intercepts with $\tau = 1$, but for usability, a single range of clustering is preferred, given by the overlapping $\underline{\mathcal{D}}$.

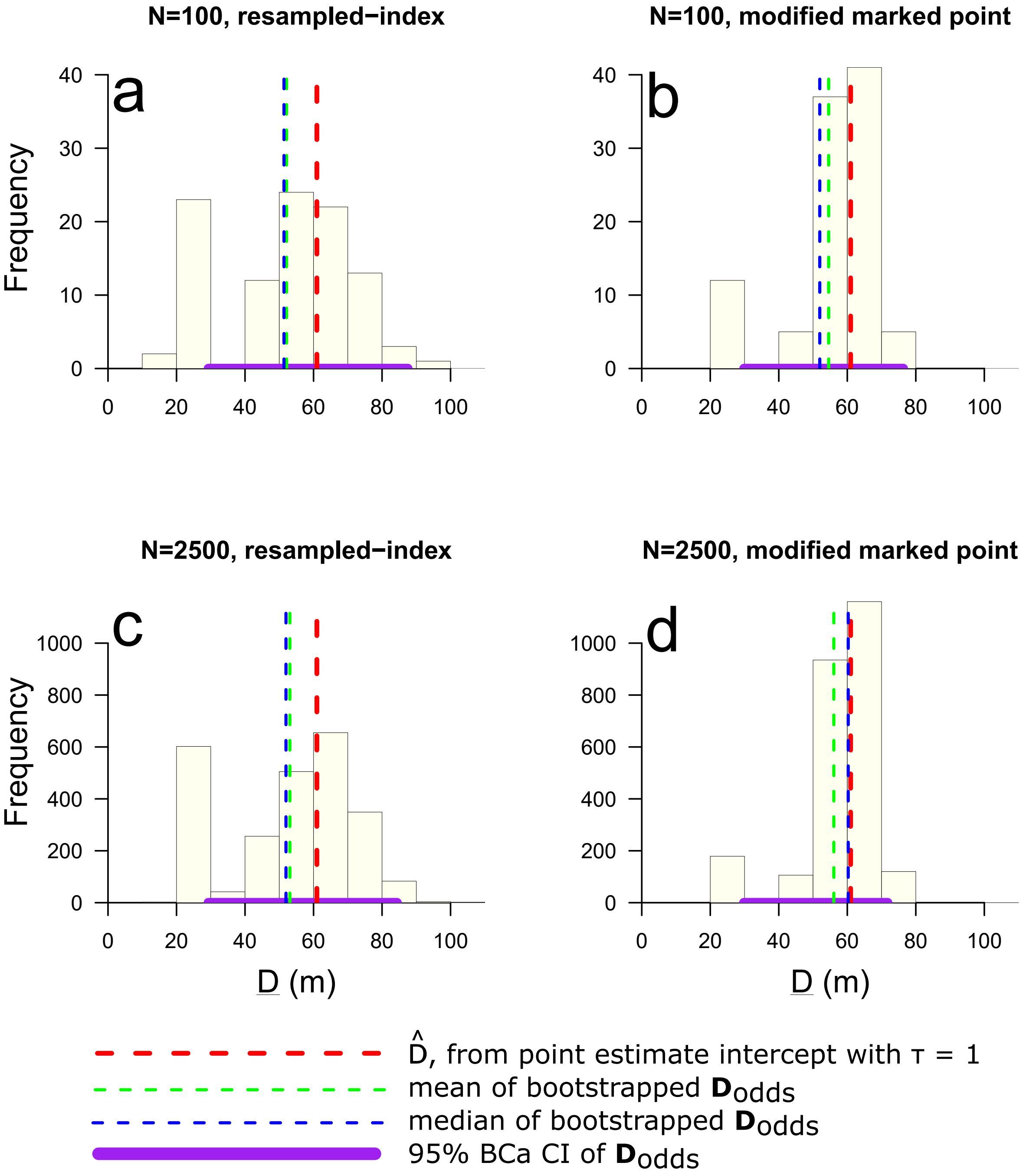


Figure 3.9: **Distribution of the endpoint clustering distance,** $\underline{D}_{\text{odds}}$, the set of samples from the sampling distribution of values of $\hat{D}_{\text{odds}}$, *i.e.* $\underline{D}_{\text{odds}} = \{\hat{D}_i : \hat{\tau}_i^*(\hat{D}_i) = 1, i = 1, \ldots, N\}$, by number of bootstrapped samples N=100 (top row) or N=2500 (bottom) and by spatial bootstrap sampling method RISB (left column) or MMPSB (right). Vertical dotted lines indicate the $\hat{\tau}$ point estimate (red), mean (green) & median (blue) of the bootstrapped tau estimates. For the RISB, both have positive skew as the mean estimate is greater than the median estimate, whereas for the MMPSB, both have a negative skew. All spatial bootstrap estimations have a negative bias concerning mean or median summary measures versus the point estimate, of approximately ∼10m for the RISB & approximately ∼5m for the MMPSB. The data points used to construct the BCa CIs (purple line on the horizontal axis) from the $\hat{D}_{\text{odds}}$ estimates in (a) are copied from Figure 3.7 (N=100 simulations) while those for (c) & (d) are from Figure 3.8, while (b) has been freshly calculated. All four CIs used 100% of simulations. Distance band set as Figure 3.6. Measles cases are considered temporally related within [0,14d].

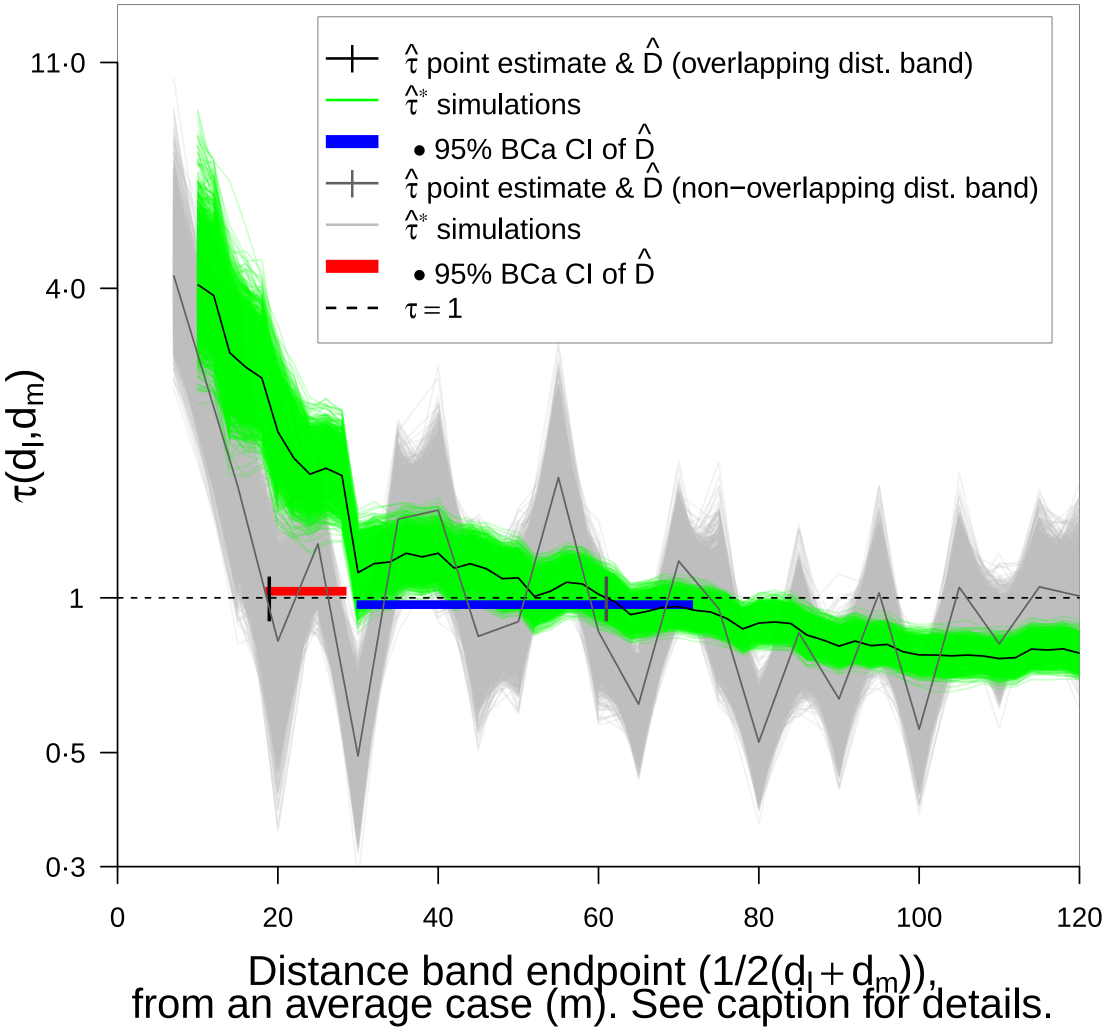


Figure 3.10: **Effect of the distance band set on $\hat{D}_{\text{odds}}$ precision using MMPSB sampling.**
Overlapping set ([118]) := $\{[0,10),[0,12),[0,14),\ldots,[0,50),[2,52),[4,54),\ldots,[74,124\text{m})\}$ & non-overlapping := $\{[0,7),[7,15),[15,20),[20,25),[25,30),\ldots,[195,200\text{m})\}$.
Non-overlapping sets yield a more erratic point estimate $\hat{\tau}$ yet tighter 95% BCa CI (18·4, 28·6m) versus (29·8, 71·8m) however, on further investigation, the distribution of $\underline{D}_{\text{odds}}$ is heavily bimodal; both CIs used 100% of 2,500 bootstrapped simulations. Measles cases are considered temporally related within [0,14d].

## 3.4 Conclusion & recommendations for improved use

It has been shown that the way clustering ranges are calculated using the tau-odds estimator can lead to biased estimates. However, using MMPSB & BCa CIs to calculate the clustering range for this measles dataset resulted in bias reductions equivalent to increasing the clustering area of elevated odds by 342%. These improvements will appear in future versions of the `IDSpatialStats` package. The results of §3.3 support the following recommendations:

- the MMPSB should be used to simulate $\hat{\tau}$ instead of the RISB method, which could lead to underestimating the clustering range.
- BCa, rather than percentile CIs should be used as they give better coverage when the bootstrap distribution of tau simulations $\underline{\hat{\tau}}^*$ is non-symmetric.

The distance band set choice $[d_l, d_m) \in \underline{\mathcal{D}}$ affects the smoothness of the point estimate $\hat{D}_{\text{odds}}$ and its precision. A better understanding of choosing distance bands for a given purpose is now needed. It is unclear how second-order correlation functions like the tau statistic & Ripley's $K$ function [73], founded initially in spatiotemporal point processes with continuous support in $\mathbb{R}^2$, behave for this data. An area of further research for the tau statistic could be the application of minimax theory to help in the choice of distance bands, e.g. Tsybakov [195] lays out a series of methods for investigating the convergence properties, optimality & adaptive estimation for non-parametric estimators.

Finally, the number of bootstrap samples required for graphical hypothesis testing & estimation purposes is unknown; it is believed that related research by Davidson & MacKinnon [56] could inform a heuristic algorithm.

$* \ \tau \ *$

The adoption of the statistical protocol described is encouraged (Fig. 3.1) to properly test for clustering and, if appropriate, estimate its range. Control programmes are being informed by the tau statistic, and applying these bias-reduction methods will improve its accuracy and future health policy decisions. In addition to modellers or epidemiologists working on real-time outbreaks or post-study analysis, it is hoped statisticians are inspired to apply this statistic to spatiotemporal branching processes in new fields. Using these improved inferential methods, the tau rate estimator from Chapter 2 is now employed to study a spatiotemporal VL dataset.

# CHAPTER 4

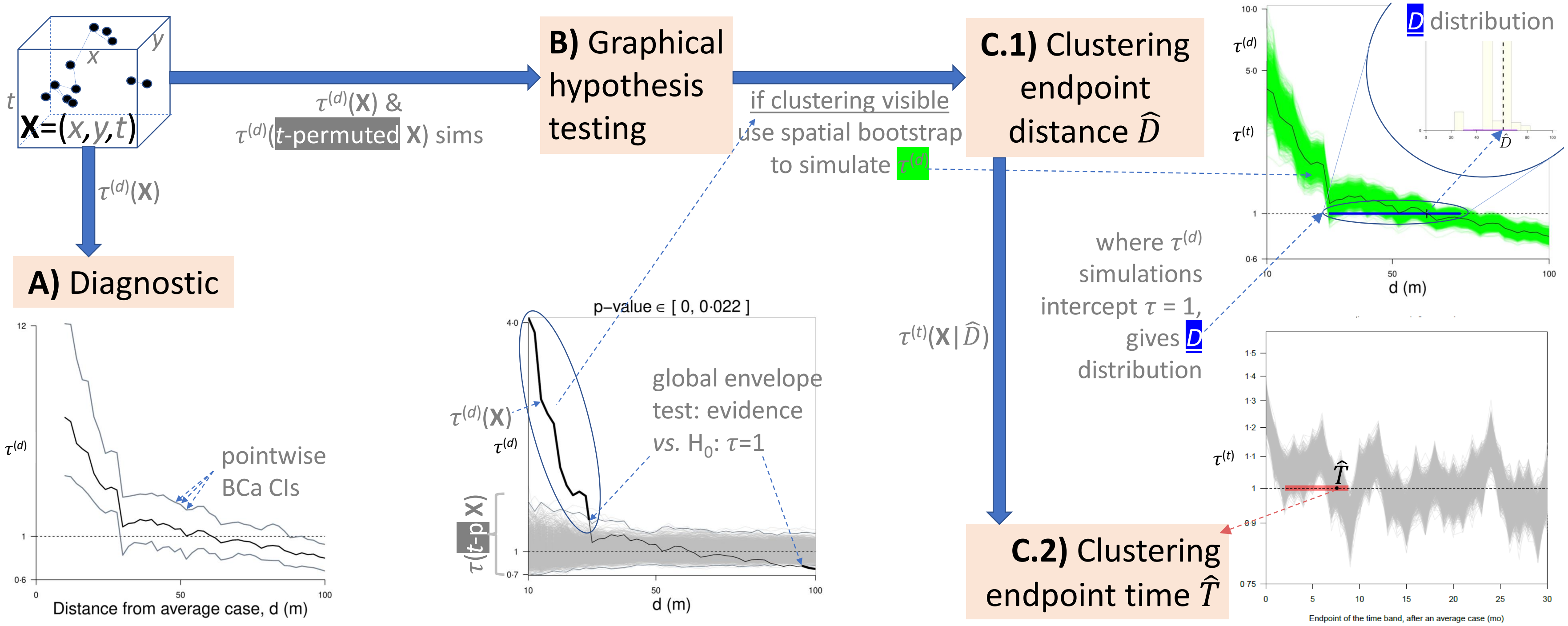
# Spatiotemporal clustering with variable exposure times: analysis using a new tau-rate estimator

*This is the final research chapter on the tau statistic and is a culmination of best practice gleaned from others' use of it in Chapter 2 & the improvements to its stages of inference in Chapter 3. The new rate estimator is applied, as developed in Chapter 2, to a VL dataset whose participants have different person-time at risk in the study due to varying time-in-study & changing immunity. There is also the opportunity to compare its estimates against a previous transmission model estimate on the same dataset & learn the (possible) optimal times to perform active surveillance following an index case.*

## Abstract

Using the newly-defined rate version of the statistic (the *tau-rate* estimator), it is applied to a VL disease dataset from Fulbaria, Bangladesh to estimate the clustering range of VL & compare the rate form with the existing tau-odds/prevalence estimators.

Statistical evidence is found to reject the null hypothesis of no spatiotemporal clustering nor inhibition, $p$-value $\in [0, 0{\cdot}02]$ over the 6km range of analysis. The clustering endpoint distance about an average index case based on the tau-rate estimator is 542m (BCa 95% CI 448–1,414m), in broad agreement with a previous model-based estimate (407m) and closer than estimates from the tau-odds (88m) & prevalence (1,779m) estimators. Therefore, the rate estimator appears to provide a more accurate clustering endpoint estimate for this dataset with considerable amounts of migration. Furthermore, within this ∼550m disc one finds that the observed rate would be up to 27% higher than average in months 0 & 1 after the index case's onset month. The tau statistic can thus provide useful information, albeit with some caveats, for targeting control interventions in space & time around cases to save resources.



X=(x,y,t)
x
y
t
τ(d)(X) &
τ(d)(t-permuted X) sims
τ(d)(X)
A) Diagnostic
B) Graphical hypothesis testing
if clustering visible
use spatial bootstrap
to simulate τ(d)
C.1) Clustering endpoint distance D̂
D̂ distribution
where τ(d) simulations intercept τ = 1, gives D̂ distribution
τ(t)(X | D̂)
C.2) Clustering endpoint time T̂
pointwise BCa CIs
Distance from average case, d (m)
τ(d)
p–value ∈ [ 0, 0·022 ]
global envelope test: evidence vs. H0: τ=1
τ(t-p X)
d (m)
τ(t)
T̂
Endpoint of the time band, after an average case (mo)

Figure 4.1: **Graphical abstract: Application of the tau statistic to spatiotemporal data, full version**. Caption as in Fig. 3.1, plus '**C.2)**' estimates the clustering endpoint time $\hat{T}$ to guide active surveillance on times after index case detection when cases are higher than normal (bottom-right plot). It is similar to $\hat{D}$ in C.1, but the tau-time estimate conditions on a disc of radius $\hat{D}$ around an average case $\tau^{(t)}(\boldsymbol{X}|\hat{D})$.

## 4.1 Motivation

In this chapter, a new (incidence) rate form of the statistic (the *tau-rate* estimator) is proposed. Designed for variable PTAR datasets (§2.3.3), it is applied, along with other forms of the statistic, to a VL dataset from Bangladesh [96] to estimate the clustering range of VL & explore the performance of the different tau estimators. The tau statistic has not yet been applied to diseases like VL, whose highly variable IP [89] could increase the uncertainty in the clustering range estimated to unacceptable levels.

The development of tau-time estimators (§2.4), literature review of the tau statistic (§2.5) and in parallel improvement to graphical hypothesis testing and point & interval estimation for $\tau$ in Chapter 3, has led to the novel analysis in this chapter. Modifications to the statistic to account for PKDL & disease-specific assumptions are also made §4.2.1.

## 4.2 Methods

### 4.2.1 VL epidemiology in Bangladesh & assumptions made

In this particular study, those who had reported VL before the study were defined by “a febrile illness with weight loss and/or abdominal swelling” [170]; whereas current suspected VL cases were defined as “2 weeks of fever plus skin darkening, weight loss, splenomegaly, and/or hepatomegaly”. Previous PKDL cases were defined as “a macular, papular or nodular rash lasting at least 1 month” which was diagnosed by a clinician and “treated with sodium stibogluconate with resolution”; current suspected PKDL cases with a rash lasting at least a month were examined by a clinician. All suspected VL & PKDL cases were screened by a clinician for these signs & symptoms and confirmed using an rK39 rapid diagnostic test. We ignore the asymptomatic stage of infection (§1.1) as it was not measured by the study.

There is substantial underreporting in Bangladesh as 136,500 VL cases were estimated in 2006 compared to 5,067 officially reported [101]. The study district of Mymensingh is thought to represent 50% of all VL cases nationally [6], while Fulbaria upazila of population 448,467 in 2011 [145], saw the most cases in 2008–13 [45]; the left panel of Fig. 1.1 indicates the upazila-level incidence rate in 2015 compared to the rest of the country. A smaller proportion of individuals usually progress to PKDL without having had VL, as did 8% in this dataset. Treated PKDL patients are assumed to be non-infectious straight after treatment, while untreated cases who self-resolve are non-infectious at the time of resolution.

The rate estimator can accept more than one event per person; however, herein it is assumed that only a single event occurs. For those who had VL twice or PKDL twice, it is assumed that it was not from two separate re-infections but rather that they never cleared the first infection successfully; this has been observed in co-infected HIV-VL patients [148]. Thus the double sum in Eqn. 2.5 simplifies from events $K$ to cases $n$. Xenodiagnosis involves using a non-infected sandfly to feed on (the skin of) VL & PKDL cases, to assess the relative infectivity of the human to the vector, by counting the parasites that had developed upon sandfly dissection. Xenodiagnosis studies [44, 96, 133] estimate that PKDL cases are 64% as infective as VL cases, and this is accounted for in all tau estimators as detailed in §4.2.3.

VL is assumed to be an ‘SIR’ disease for the 7·5 years of the data; either way (‘SIR’ or ‘SIRS’), an assumption is required but is not without uncertainty given that a previous review showed highly variable seroreversion rates, ranging $\sim 5 \times 10^{-3}$–5/yr [43].

### 4.2.2 Dataset, data cleaning & imputation

The data under analysis was collected in a community cohort study during a VL epidemic from 2002 to 2010 in Fulbaria upazila, Mymensingh district, Bangladesh [96] (Fig. 4.2). Self-reported VL case histories, PKDL case histories, treatment histories, migration & household geolocations were recorded retrospectively for 2002–6 (except for a subset that had been studied earlier and received annual house-to-house surveys in 2002–2004[16]) & prospectively for 2007–10. Relying on peoples' memories is likely to have introduced recall bias during the retrospective analysis years. Participants are likely to have forgotten when they first experienced onset and only remember more recent times. This is more of an issue for VL as field staff have reported in their case interviews in Bihar that fever symptoms are cyclical (CARE India, personal comm. & [28]). Furthermore VL cases in these rural areas primarily come from hard-working agricultural workers who have an incredible tolerance to the harsh working conditions and commonly dismiss early-onset fevers.

Following data cleaning, data was available on 24,759 individuals from 5,110 households unevenly spread over a 12km×12km area (Fig. 4.3 inset) [44, 161]. The primary investigator asserted that all households had been sampled in the study region (Bern, personal communication). Of course their study recruitment was based on informed consent, so it is possible that some households may have declined. Papers emanating [170, 96] from the study did not measure the proportion of households surveyed versus a standard reference like census records. Given the monthly surveillance and the outcome of death for undetected cases, it would have been difficult to hide unknown cases from the investigators. The SI distribution of VL (mean 7mo) was estimated as the onset-to-infection time (taken as half the empirical onset-to-treatment (OT) distribution) plus a published estimate of the IP [42, 44]. 'Empirical' in this context means obtained from observational data as opposed from a model-based estimate. Likewise, the SI of PKDL (the time from PKDL onset in an infector to VL onset in an infectee) (mean 15mo) used the PKDL OT based on those who got PKDL treatment or self-resolved, plus the VL IP.

Comparing Chowdhury et al.'s 2008–10 data from central- & district-level case reports for this region [45] with that calculated here indicates this study covered ∼4%(113/2,552) of cases. The households were naturally clustered as 19 'paras': hamlets on higher ground surrounded by arable land [27] (Fig. 4.3 inset). For each para, houses were sampled in a nearest-neighbour sequence that generated an inhomogeneous spatial process mirroring the linear road network & embankments on which houses were built. As the odds & prevalence estimators do not account for the PTAR, internal migrators (who only represent 3% of all people in the time-censored dataset) will be represented twice across two concurrent rows as infectee $j$ or then infector $i$—the alternative was complete row removal of both entries which was rejected.

Households belonged to two spatially-disjoint north-west & south-east regions (Fig. 4.3 inset). So, the double summation in both the numerator & denominator of the estimator functions was split into two—compatible with $\tau$ being a global statistic.

Commonly, simulations of the null hypothesis of no spatiotemporal clustering nor inhibition can be produced by applying tau functions to time-scrambled data [165]. Extending this to a rate statistic $\tau^{(d)}_{\text{rate}}$ with infectiousness & susceptibility time variables, for a 'person (row) × variable (column)' structure, the time variables are randomly swapped row-wise keeping each row's time variables immutable for temporal consistency.

Institutional review boards at icddr,b (protocol № 2001-021 & 2007-003), CDC (3230 & 5065) & University of Warwick's Biomedical & Scientific Research Ethics Committee (REGO-2019-2344) all approved this analysis. The tau functions were coded in the *C* language as adaptations from `IDSpatialStats` v0.3.7 and interfaced with *R* code written using `Rcpp` [62], providing significant

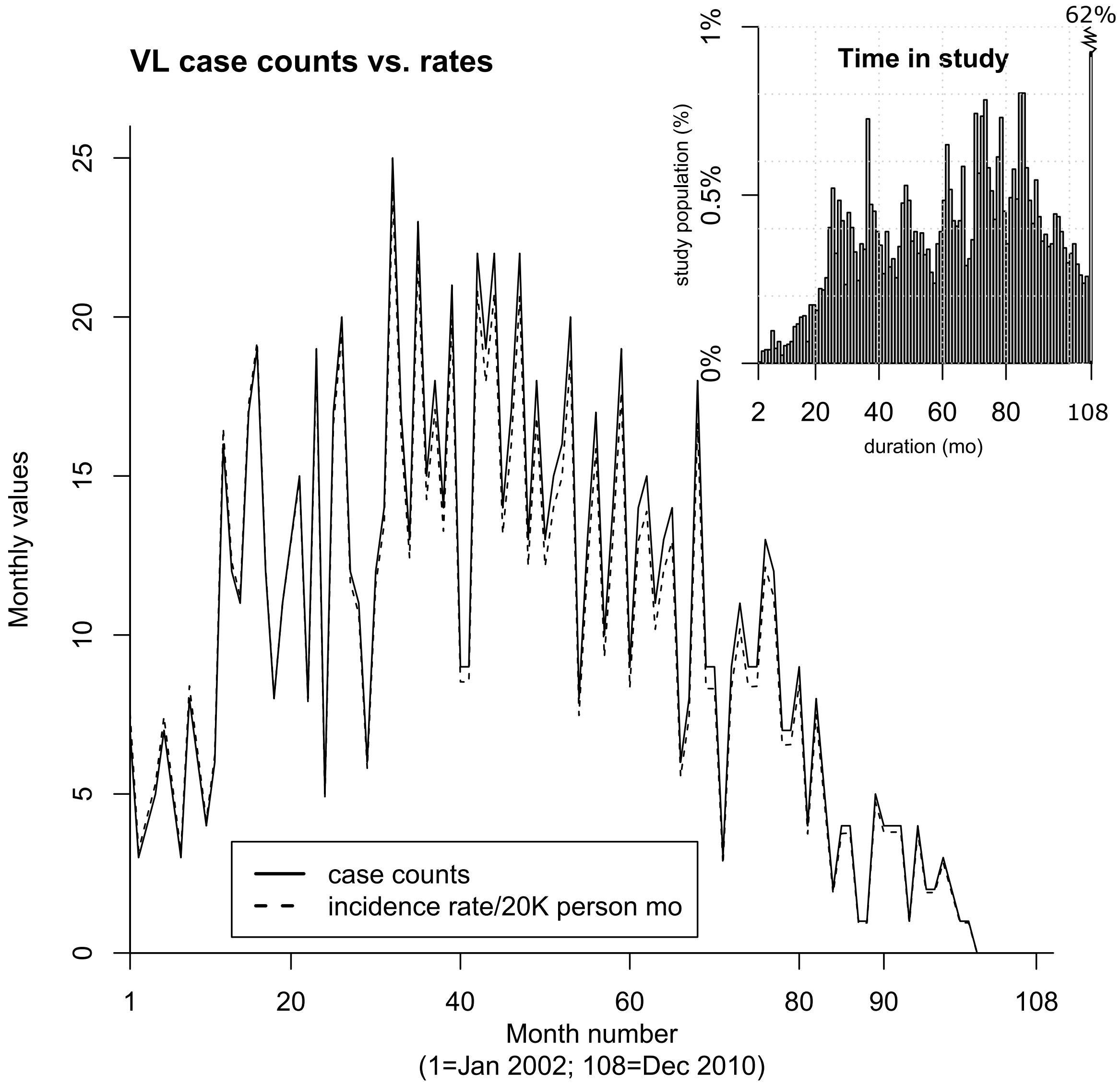


Figure 4.2: **Comparing the time series of month of symptom onset for case counts versus the incidence rate of newly detected VL cases on the original dataset**. In the main figure the '20K person mo' represents the total person-months at risk for the study population, rescaled by 20,000. Inset shows the distribution of study durations; note the y-axis scale and that a sizeable minority (38%) of the study population had study durations less than 107 months.

speed improvements [159]. Unit tests helped verify that the results were within expected ranges or had a certain structure [122]. Analysis code in *R Markdown* is available at `github.com/t-pollington/taurate`.

Close inspection of the data revealed inconsistencies, which required extensive cleaning, using *STATA 14*. Errors were found in the time ordering of entry, exit & case events which were resolved by referring to the raw data. Following marriage, a new house would often be built close to the house of one of their families for the newly-weds to live, yet those who moved into new households (known from their updated household IDs) did not have updated GPS coordinates; these were updated by

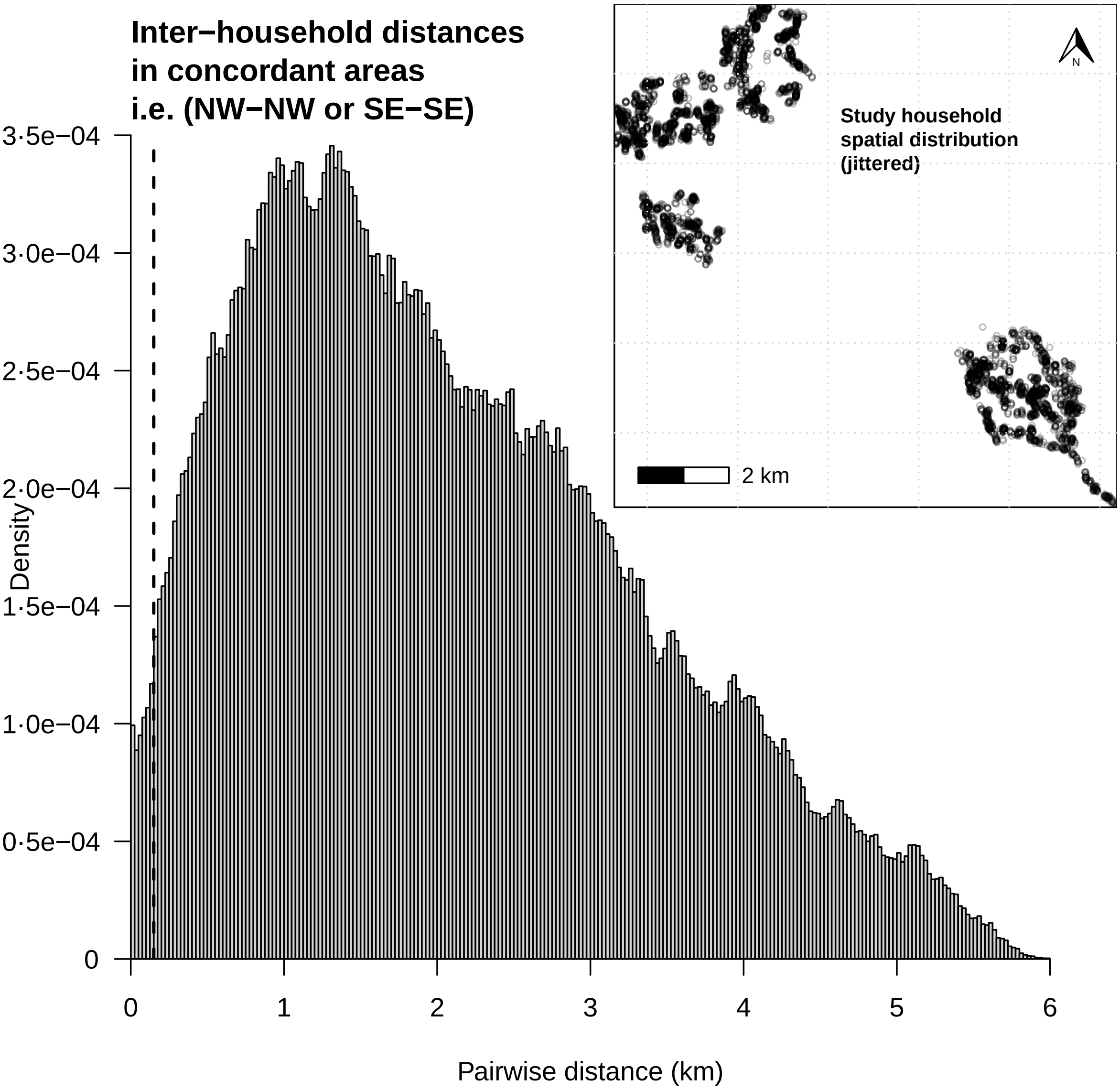


Figure 4.3: **Household pairwise distances in the time-censored dataset**. Inset: study households in Fulbaria village, Bangladesh; UTM projection [69]. The vertical dashed line marks a distance of 150m between households, beyond which the tau variance starts to reduce, as in Figure 4.5. Household locations jittered to protect their actual coordinates.

moving the new house in a random cardinal direction by 1×$10^{-4}$ degrees (∼10m) away from the older house.

The original dataset contained 25,512 observations. Most individuals were represented by a single observation unless they internally migrated when two were used. The two main attributes with missing month data but known year were VL onset with 246 (1·0%) missing and date of birth with 21,886 (85·8%) missing—these missing months were imputed uniformly, i.e. each month had a $^1/_{12}$ chance of being chosen. There were 382 people with missing VL treatment dates who had had a case of VL before the study started; as they had all reported a specific drug treatment, they were assumed treated before Bern et al.'s 2002–2010 epidemiological study started and hence immune to further infection.

A histogram of study entry month for participants entering & born after 2002 and who had not migrated showed an abnormally high proportion of January entries (35·1% 1,564/4,456): due to the unknown entry months being coded as January. However, correcting a January record uniformly to any month was naïve for some groups. For example, of those who had non-January entry months, 74·5%(2,859/3,837) entered on their birth month; the birth month was assumed correct as it had an expected seasonal distribution. For those born & entering on the same year and after 2002, they were assigned as entering in January (as originally reported) with 74·5% probability if they had been reported as born in January (as this was the probability seen in non-January enterers who had the same month of birth); or else to the other 11 months uniformly with probability $^1/_{11}$; b) and if the reported January enterer had a non-January birth month, then their entry month would be imputed uniformly (with probability $^1/_{12}$) to any month; in addition, those who had VL in the entry year had their imputation constrained in the range [January, VL symptom onset mo].

One needs to censor completely (*i.e.* complete row removal from the dataset, no temporal trimming) those cases with infectiousness falling over the entry period (month 1), *i.e.* with a symptom onset pre-entry & treatment during the study. Firstly, this is because including them would make their $t_j - t_i$ onset time difference with an infectee look artificially smaller, and, if they had initially been temporally unrelated, it could make them look related with a shorter time difference. Secondly, the location of $i$ & $j$ is an unknown pre-study and avoidance of additional assumptions helps to obtain an untainted result from this empirical dataset. However, although the pre-study location is not assumed, their (temporal) disease history can be used to inform susceptibility. Two cases that had onset but treated after month 90 were included as their inclusion as infectees was more important than excluding them because of their infector role at this late stage. The censored dataset covers 25,456 row entries for 24,759 people since some people can have 2 rows representing two study locations they migrated between during the study. Missing VL treatment times for those with VL onset pre-study were imputed from the empirical OT distribution [44]. For non-VL PKDL cases, their susceptibility ends at PKDL onset.

The drawbacks to time-censoring, apart from reducing the data duration by ∼17%, are i) the omission of PTAR spent in other VL-endemic areas before those in-migrators who have disease onset soon after arrival, these overlapping cases were omitted at the start of the study, and ii) the omission of disease events just after leaving the study. Nevertheless, the study period is still relatively long, covering ∼12 VL or ∼6 PKDL mean SIs.

### 4.2.3 Accommodating the secondary infectious stage (PKDL)

The rate estimator needs to account for PKDL for the distance-form (Eqn. 4.1) and provide the alterations to the odds & prevalence estimators (Eqns. 4.2 & 4.3), which are slightly different in their sum. The total rate $\lambda$ (for a new VL case, not PKDL case, caused by an existing VL or PKDL case) in an annulus is assumed to be the sum of the rate caused by VL ($\lambda_{\text{VL}}$) & PKDL cases ($\lambda_{\text{PKDL}}$) independently (Eqn. 4.1): this is reasonable as if an individual experiences both states, they will be separated by a time gap. So each VL-susceptible person VL$_j$ experiences an infection rate $\lambda$ as the addition of rates from the VL$_i$ infector $\frac{\text{№(VL}_i \texttt{ -> } \text{VL}_j \text{ transmission-related links)}}{\text{time at risk of all VL}_j \text{ under spatiotemporally-close VL}_i}$ and PKDL$_i$ infector $\frac{\text{№(PKDL}_i \texttt{ -> } \text{VL}_j \text{ transmission-related links)}}{\text{time at risk of all VL}_j \text{ under spatiotemporally-close PKDL}_i}$, and similarly for odds or prevalence estimators.

The lower infectivity of PKDL cases to sandflies is represented by weighting the count of a relevant pair as 0·64 rather than unity, calculated from xenodiagnosis studies [44, 96, 133]. Therefore, a PKDL$_i$ -› VL$_j$ link counts 0·64 as much as a VL$_i$ -› VL$_j$ one; regarding the sensitivity of this single value, based on later analyses where the infectivity of PKDL cases was parameterised as three different parameters according to PKDL lesion type, there was negligible change to results.

$$\hat{\lambda}(d_1, d_2) = \frac{\sum_{a=1}^{K} \sum_{b=1, k_l \neq k_m}^{K} \mathbb{1}(z_{ab}^{\mathrm{VL}_i \dashrightarrow \mathrm{VL}_j} = 1, d_1 \leq d_{ab} < d_2)}{\sum_{i=1}^{n_{\mathrm{VL}}} \sum_{j=1, j \neq i}^{N} \sum_{t=1}^{t_{\mathrm{end}}} \mathbb{1}(Z_{ij}^{\mathrm{VL}_i \dashrightarrow \mathrm{VL}_j}(t) = 1, d_1 \leq d_{ij}(t) < d_2)} + \frac{0{\cdot}64 * \sum_{a=1}^{K} \sum_{b=1, k_l \neq k_m}^{K} \mathbb{1}(z_{ab}^{\mathrm{PKDL}_i \dashrightarrow \mathrm{VL}_j} = 1, d_1 \leq d_{ab} < d_2)}{\sum_{i=1}^{n_{\mathrm{PKDL}}} \sum_{j=1, j \neq i}^{N} \sum_{t=1}^{t_{\mathrm{end}}} \mathbb{1}(Z_{ij}^{\mathrm{PKDL}_i \dashrightarrow \mathrm{VL}_j}(t) = 1, d_1 \leq d_{ij}(t) < d_2)} \tag{4.1}$$

where $z^{(\cdot)}$ or $Z^{(\cdot)}$ is as defined in Eqn. 2.6 with the additional condition that it only evaluates to 1 according to transmission from a '$\mathrm{VL}_i \dashrightarrow$ ' or '$\mathrm{PKDL}_i \dashrightarrow$ ' infector and has temporal relatedness $[0, \mathrm{VL\ mean\ SI}]$ or $[0, \mathrm{PKDL\ mean\ SI}]$ ranges, respectively.

$$\hat{\theta}(d_1, d_2) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 1, d_1 \leq d_{ij} < d_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 1, d_1 \leq d_{ij} < d_2)}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 0, d_1 \leq d_{ij} < d_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 0, d_1 \leq d_{ij} < d_2)} \tag{4.2}$$

$$\hat{\pi}(d_1, d_2) = \frac{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 1, d_1 \leq d_{ij} < d_2) + 0{\cdot}64 * \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 1, d_1 \leq d_{ij} < d_2)}{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbb{1}(d_1 \leq d_{ij} < d_2)} \tag{4.3}$$

These formulae similarly extend to the time-form (Eqns. 4.4 & 4.5):

$$\hat{\theta}(t_1, t_2) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 1, t_1 \leq t_{ij} < t_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 1, t_1 \leq t_{ij} < t_2)}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 0, t_1 \leq t_{ij} < t_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 0, t_1 \leq t_{ij} < t_2)}. \tag{4.4}$$

$$\hat{\pi}(t_1, t_2) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{VL}_i \to \mathrm{VL}_j} = 1, t_1 \leq t_{ij} < t_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(z_{ij}^{\mathrm{PKDL}_i \to \mathrm{VL}_j} = 1, t_1 \leq t_{ij} < t_2)}{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(\mathrm{VL}_i \to \mathrm{VL}_j, t_1 \leq t_{ij} < t_2) + 0{\cdot}64 * \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}(\mathrm{PKDL}_i \to \mathrm{VL}_j, t_1 \leq t_{ij} < t_2)}, \tag{4.5}$$

where the denominator of $\hat{\pi}(t_1, t_2)$ is more complex than its distance form since the proportion of VL vs. PKDL transmission may differentiate at different time bands for a specific distance band (implicitly set in $z_{ij}^{(t)}$), whereas the denominator of $\hat{\pi}(d_1, d_2)$ sums over all people, so counting the number of links is sufficient rather than their VL/PKDL transmission type.

### 4.2.4 Analysis

The tau-rate statistic is applied to the variable PTAR VL data and compared across estimator types. All analyses used the same distance band set $\mathcal{D}$ starting at 10m gaps from the case where $\tau$ can change rapidly, then 50m gaps from 100m–1km & 1km gaps from 1–6km; and same time band set $\mathcal{T} = \{0, 1, \ldots, 30\mathrm{mo}\}$. BCa CIs are constructed because the bootstrap distributions for $\tau$, $\hat{D}$ or $\hat{T}$ are non-Normal. The BCa algorithm uses adapted code from the `bcaboot` *R* package [65, 66]. Unlike standard percentile CIs, one also needs to perform diagnostic checks on the BCa output. Fortuitously, the BCa algorithm can also assess & provide bias correction to the empirical estimate.

Spatial bootstrapping using MMPSB (§3.2.4.1) is used to estimate the precision of a) tau point estimates (*e.g.* Fig. 4.5) across the distance band $\mathcal{D}$ (or time band $\mathcal{T}$) set, or b) its derived estimates—the clustering endpoint distance $\hat{D}$ (*e.g.* Fig. 4.8) or time $\hat{T}$ clustering, respectively. All bootstrapped

$\tau$ simulations intercepted with $\tau = 1$ within the distance $\mathcal{D}$ or time $\mathcal{T}$ band set, thus avoiding bias to the CIs constructed [165]. The jackknife applied in the BCa algorithm for choosing the 'acceleration factor' might be improved by applying the 'leave one out' algorithm not on data indices like the RISB but rather on local mark functions.

The analysis proceeds with i) descriptive analyses of incidence, population & time at risk (§4.3.1); then the inference order (Fig. 4.1) as set out in Pollington et al. [165]: ii) an indicative/diagnostic plot of $\tau^{(d)}_{\text{rate}}$ vs. distance (§4.3.2); iii) a graphical hypothesis test to assess the evidence against the null hypothesis of no spatiotemporal clustering nor inhibition (§4.3.3); iv) if the evidence is found then clustering endpoint distance $\hat{D}$ is estimated (§4.3.4). Additionally, $\hat{D}$ is compared to Chapman et al.'s model-based rate estimate [44]; v) contingent on $\hat{D}$ estimation, $\tau^{(t)}_{\text{prev}}$ estimates the duration $\hat{T}_{\text{prev}}$ (§4.3.5) for which the observed rate within a disc of radius $\hat{D}$ will stay higher than at any time afterwards. This information could inform control policy on when active surveillance could detect most cases after the index case [8]. It is accepted that there is inconsistency in switching estimator types from $\tau^{(d)}_{\text{rate}}$ (step iv above) to $\tau^{(t)}_{\text{prev}}$ (step v). This is because they are the best estimators in their distance/time classes, respectively and as there is no rate form of the tau-time estimator (§2.4.3).

#### 4.2.4.1 Temporal confounding of the tau-time analysis

By definition the high-season of VL incidence sees the highest incidence and the low-season the lowest. Therefore, cases during the high-season will have a heavier weighting on the pairwise tau-time statistic: disproportionately-more pairings occur between infectors in the high-seasons and infectees with onsets months afterwards during the low-seasons. In a similar way there are disproportionately-more pairings between (fewer) infectors in the low-seasons that pair with (far more) infectees that have onsets months afterwards during the high-seasons. This temporal heterogeneity by season confounds the tau-time analysis.

The following example uses the disease frequency measure of odds for the tau-time statistic—$\theta(t_1, t_2)$. For an epidemic with a 12-month period and serial interval of approximately 6 months, high-season infectors & low-season infectors would pair with infectees in low/high seasons, respectively, disproportionately more than the average, thus overestimating $\theta(t_1, t_2)$ at the peak times of $t_1, t_2 \approx [0, 1\text{mo}), [6, 7\text{mo}), [12, 13\text{mo}), +6\ldots$, compared to $\theta(0, \infty)$, and consequently overestimating $\tau(t_1, t_2)$. Whereas in-between these peak times the pairings would be in the mid seasons (between high-to-low & low-to-high and vice-versa), so would see less pairings than the average thus underestimating $\theta(t_1, t_2)$ and consequently underestimating $\tau(t_1, t_2)$ for $t_1, t_2 \approx [3, 4\text{mo}), [9, 10\text{mo}), +6\ldots$.

What should one do in light of this confounding? There is not a correction to the tau statistic that can be offered. Instead, my recommendation for tau-time graphs (i.e. Figs. 4.11 & 4.12) is to only give importance to the first seasonal period after the onset of the infector. For the example described above that would be 0–12 months and conditional on an average 6 month serial interval that values between 0–1·5mo, 4·5–7·5mo & 10·5–12mo are likely to be overestimates and 1·5–4·5mo & 7·5–10·5mo as underestimates.

In the case of this chapter's VL dataset the confounding problem worsens. Firstly, the periodogram of this chapter's VL dataset has a dominant non-trivial period at 3·4mo, however there are a number of credible second- & third-dominant, longer seasonal periods too. Secondly, it is less clear the direction of over or underestimation for high-season infectors, where their infectees would have onset on average 7 months later and thus not in synchrony with the low-season in onsets as described in the 12-month period epidemic above. All that can be done is to warn the reader of the possibility of bias in either direction in the first three months and to ignore $\tau^{(t)}$ values in later months (Figs. 4.11 & 4.12). Choosing a threshold limits the impact of repeat cycles of seasonal incidence that could be misinterpreted at later times in the tau-time graph, as representing other aspects of the disease instead e.g. generation time or changing immunity—thus limiting one effect of temporal confounding while others would remain.

## 4.3 Results

### 4.3.1 Incidence rate, population changes & person-time at risk

The incidence rate is a more valid measure of disease frequency in a variable PTAR setting and is the natural statistic for non-spatial purposes. Only ∼60% of the study participants spent the entire 108 months in the study with a mean time-in-study of 78 months; the times spent by the remainder were approximately uniformly distributed between 2–107 months (Fig. 4.2 inset). This PTAR variability motivates the exploration of the rate estimator herein.

The noisy incidence curve is correct and not a plotting error, as validated by Islam et al.'s [96] descriptive analysis of the same dataset (c.f. Islam et al. Fig. 1). Possible reasons for high lag-1 month volatility are i) annual household surveys during the visits of international co-PIs that may have increased surveillance efforts to improve detection, ii) temporal clustering around these times as reporting bias of self-reported onset dates meant people were more likely to remember recent symptoms as their first onset, iii) the incidence graph is an aggregation of 19 distinct paras (hamlets) and a temporal pattern of transmission across a serial interval of 7 months is unlikely to yield a strong lag-1 month autocorrelation in incidence and finally iv) the volatility in observed cases is within the expected margins of Poisson noise.

The study population only changed modestly: $n(t) = 18,571$–$21,661$ (Fig. 4.4). Those who became immune through observed disease represented up to 5%. There was an increase in the study population in the first 5 study years contributed by births mainly. In the final 3 years, births had dropped and internal movements were highest. VL deaths remained low throughout and contributed less than migrations out of the study area. The lines corresponding to the total population, births, deaths & migrations in Figure 4.4 were mostly constant after month 90 due to a winding down of demographic registration while disease surveillance continued. This wind-down could bias PTAR; therefore, only months 1–90 (inclusive) are considered—benefits & drawbacks of this censoring are discussed in §4.2.2. Note that only Figures 4.2 & 4.4 used the original dataset.

### 4.3.2 Indicative/diagnostic $\tau^{(d)}$ vs. distance plot

Figure 4.5 shows a decreasing trend in $\tau^{(d)}_{\text{rate}}$ with distance from an average case. For those living in the same case household, the rate is 7·20 times higher (95% BCa CI 5·87–8·38) than the rate at any distance. There is higher variability in $\tau^{(d)}_{\text{rate}}$ at distances shorter than 150m. This may be explained by the lower point pair density at these distances, whereas this density rises steadily from 150m up to ∼1km (Fig. 4.3), thus reducing the sampling error of the statistic. $\tau^{(d)}_{\text{rate}}$ within the household is singularly far higher than in its surroundings ($0 < d < 10$m), suggesting dominant within-household transmission. $\tau^{(d)}_{\text{rate}}$ values fall close to unity (no clustering/inhibition) at ∼400–500m. The transitions in $\tau^{(d)}_{\text{rate}}$ between the first (within-household) & second distance bands and at ∼400m are both 'statistically significant' since successive CIs do not overlap. The latter may represent the maximum dispersion of infectious sandflies, the average movement range of villagers while infectious or topographical barriers (*e.g.* rivers or uninhabited open land without tree cover).

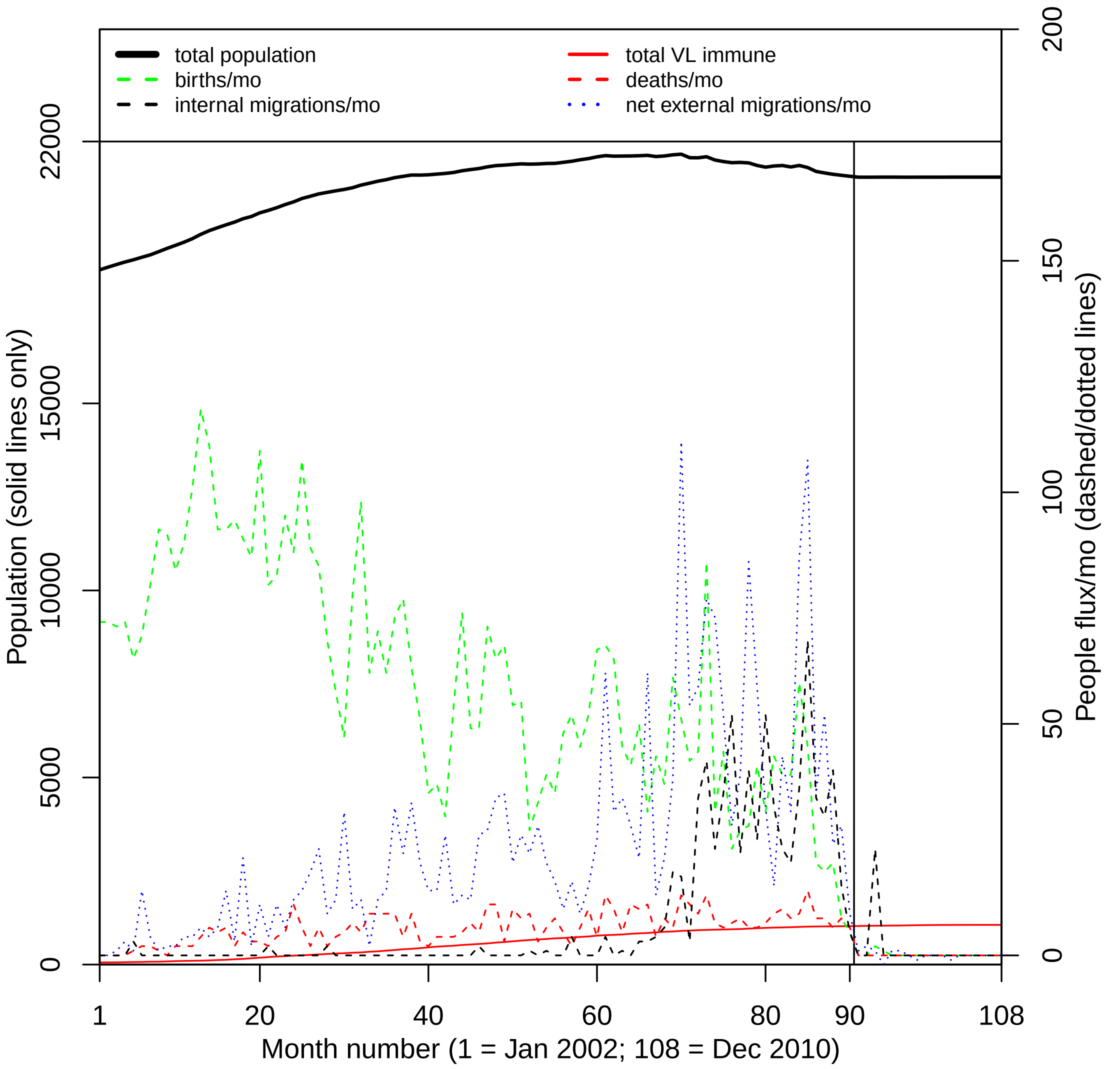


Figure 4.4: **Study population time series of the original dataset**. Analysed data is right-censored from month 91 (grey vertical line).

### 4.3.3 Graphical hypothesis test to assess spatiotemporal clustering using $\tau^{(d)}$

Commonly, simulations of the null hypothesis of no spatiotemporal clustering nor inhibition are produced by applying tau functions to time-scrambled data [165]. In the data's 0–6km range, clustering is suggested up to ~500m and inhibition between ~1,000–2,000m & between ~5,000–6,000m ($p$-value $\in$ [0, 0·02]) (Fig. 4.6). For $\tau^{(d)}_{\text{odds}}$ & $\tau^{(d)}_{\text{prev}}$ estimators, there is also evidence against $H_0$ ($p$-value $\in$ [0, 0·025] & [0, 0·02], respectively) (Fig. 4.7); thus leading to the same inference conclusion for analysis step iii in §4.2.4. Null simulations have the widest global envelope for $\tau^{(d)}_{\text{prev}}$, then $\tau^{(d)}_{\text{rate}}$ then $\tau^{(d)}_{\text{odds}}$ (Fig. 4.7). Having rejected $H_0$, one moves to the next stage to estimate $\hat{D}$.

We choose to evaluate tau over the data's spatial range of 0–6km to cover pairwise distances within the north-west or south-east paras' cluster (Fig. 4.6). Although unnecessary for indicative/diagnostic

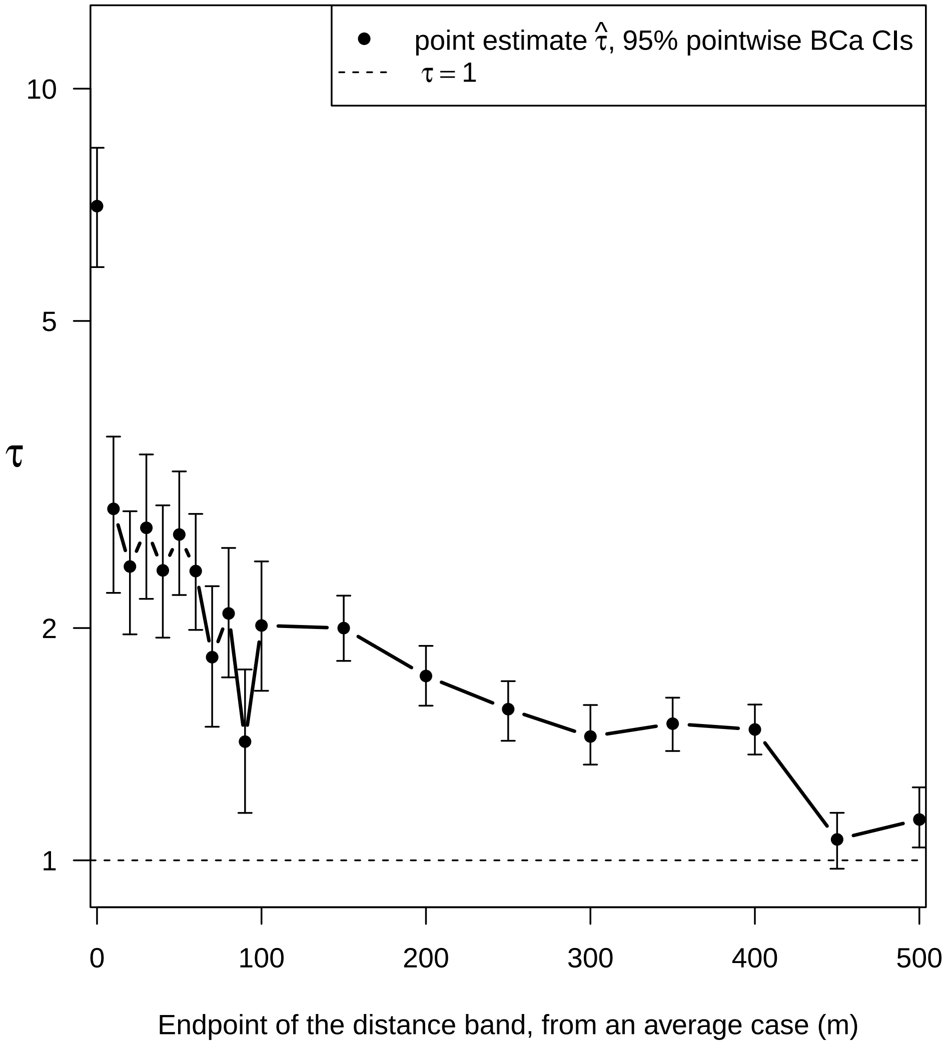


Figure 4.5: **Tau-distance statistic $\tau^{(d)}_{\text{rate}}$ (rate estimator)** for VL cases in Fulbaria village, Bangladesh, January 2002–June 2009. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo]. Distance axis values indicate the endpoint $d_m$ of each distance band $[d_l, d_m)$. $\tau\big([d_0, d_1)\big)$ is left unconnected as explained in §2.5.11. 2,000 bootstraps per pointwise CI. CIs' diagnostic checks passed.
Distance band set $:= \big\{[0, 10\text{m}), [10, 20\text{m}), \ldots, [100, 150\text{m}), [150, 200\text{m}), \ldots, [1, 2\text{km}), \ldots, [5, 6\text{km})\big\}$.

tau-distance plots where our main interest is the spatial range of spatiotemporal clustering (c.f. 500m for Fig. 4.5), it does matter for proper global hypothesis testing. The global envelope is calculated to account for the multiple hypothesis test over its entire spatial range so spatial truncation would invalidate the test.

Additionally, if there is (reasonable) concern about extreme data at far distances (on tau estimate bias/precision or the performance of the graphical hypothesis test), it is unclear at what distance to truncate. Without further knowledge of how to methodically choose distance bands (§2.5.8 &

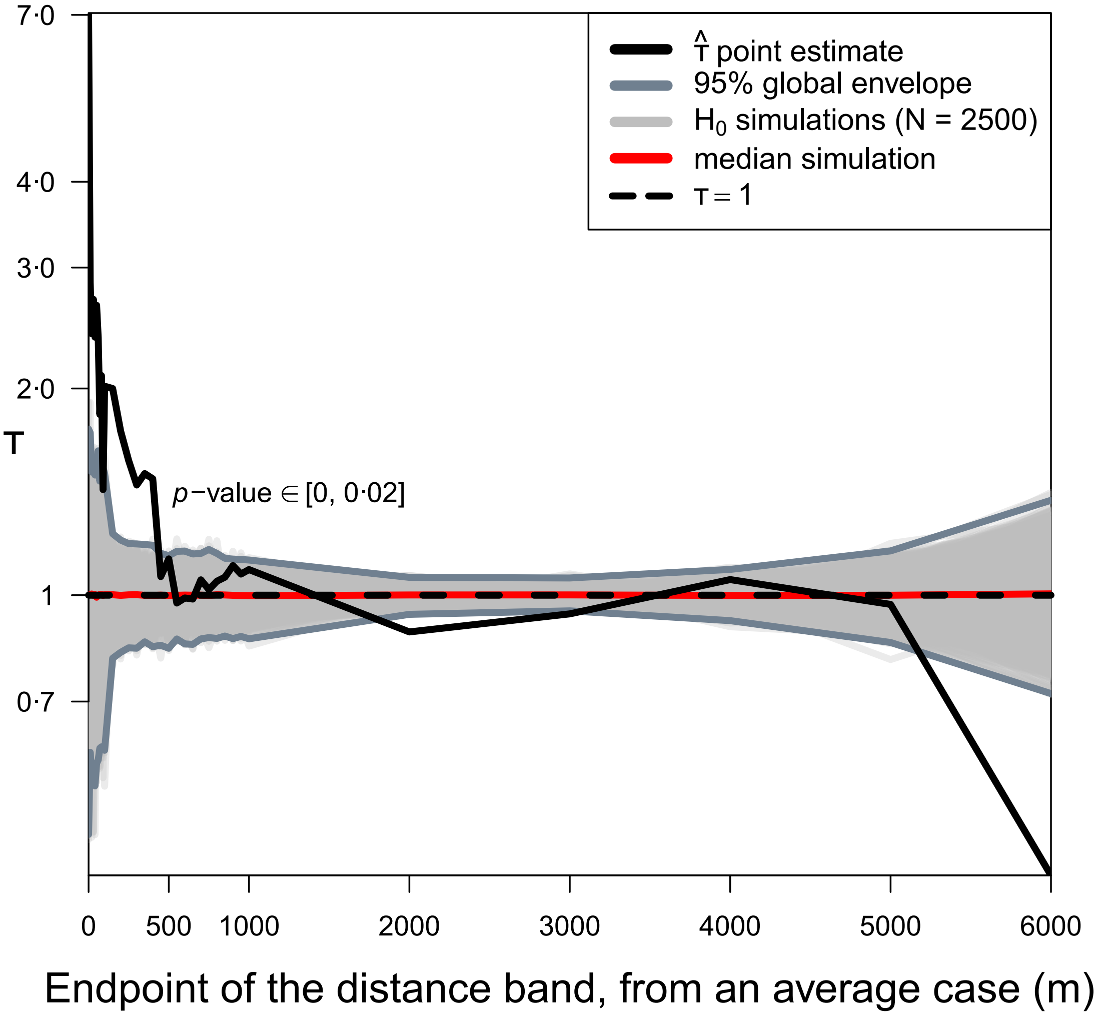


Figure 4.6: **Graphical hypothesis testing using $\tau^{(d)}_{\text{rate}}$.** Global envelope test, 'extreme rank' type, two-sided at 95% significance level [137] for $H_0$ (no spatiotemporal clustering nor inhibition, *i.e.* $\tau^{(d)} = 1$). Distance band set $:= \{[0, 10\text{m}), [10, 20\text{m}), \dots, [100, 150\text{m}), [150, 200\text{m}), \dots, [1, 2\text{km}), \dots, [5, 6\text{km})\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo]. The $H_0$ simulations are reliable as the median of simulations stays close to $\tau^{(d)} = 1$.

3.3.5) this would then be an arbitrary decision; surprisingly those at close distances could also be seen as 'extreme' ends of the data with respect to spatial separation, since the pairwise numbers at $0 < d_{ij} < 10$m are fewer than those separated by 4km (Fig. 4.3).

### 4.3.4 Estimating the clustering endpoint distance, $\hat{D}$

The $\hat{D}$ estimates by estimator were as follows: $\hat{D}_{\text{rate}} = 542\text{m} \sim 550\text{m}$ (95% BCa CI 448–1,414m) (Fig. 4.8); $\hat{D}_{\text{odds}} = 88\text{m}$ (95% BCa CI 72–555m) & $\hat{D}_{\text{prev}} = 1,779\text{m}$ (95% BCa CI 486–1,989m). The

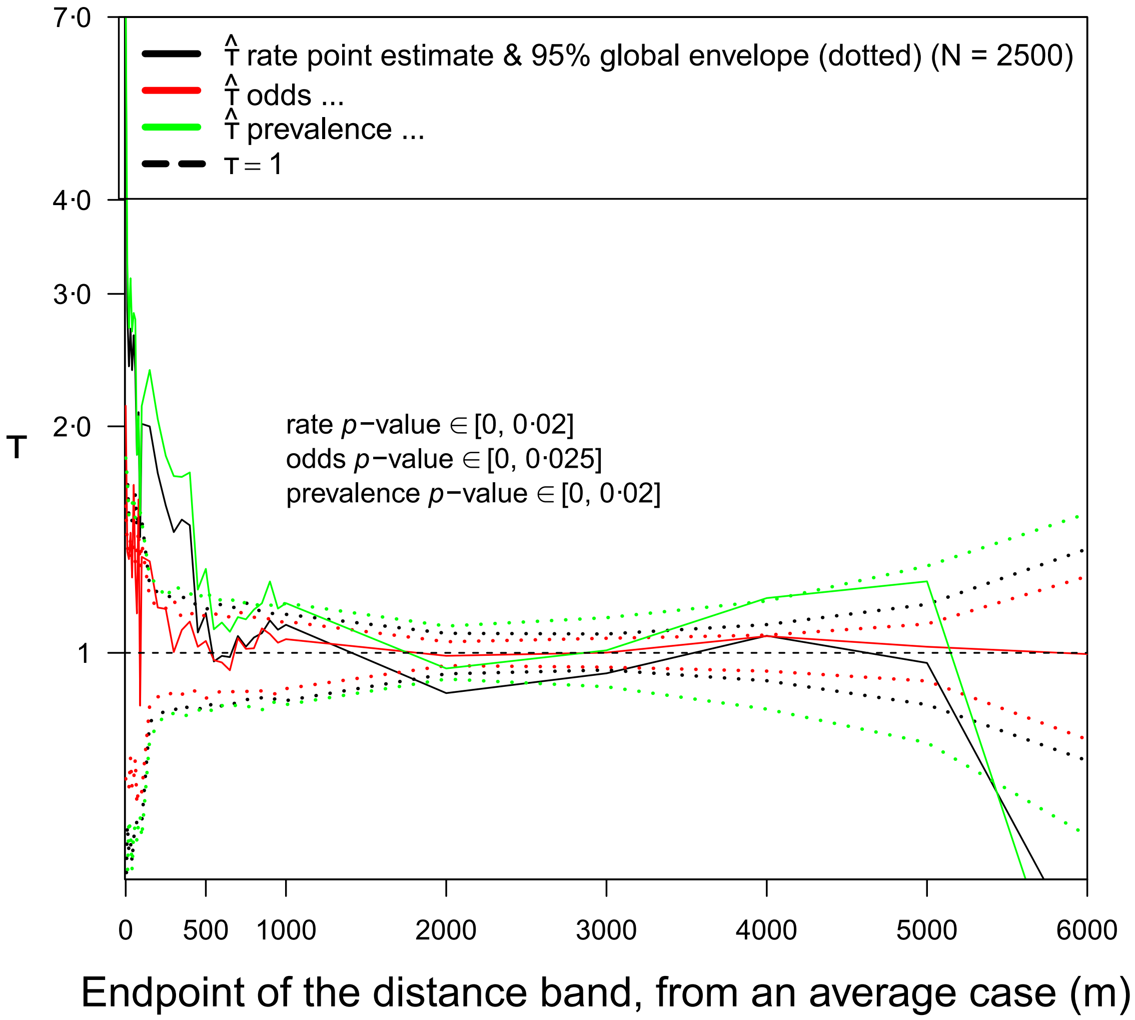


Figure 4.7: **Graphical hypothesis testing for all three distance-form estimators**. Global envelope test, 'extreme rank' type, two-sided at 95% significance level using 2,500 simulations [137] of the null hypothesis ($H_0$: no spatiotemporal clustering nor inhibition, *i.e.* $\tau^{(d)} = 1$). Distance band set := $\{[0, 10\text{m}), [10, 20\text{m}), \dots, [100, 150\text{m}), [150, 200\text{m}), \dots, [1, 2\text{km}), \dots, [5, 6\text{km})\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo].

standard deviation of the bootstrapped $\hat{D}_{\text{prev}}$ estimate was relatively large (88m) with instability in the upper bounds of the CIs for both $\hat{D}_{\text{odds}}$ & $\hat{D}_{\text{prev}}$. In the distribution of the derived estimates, $\underline{D}_{\text{rate}}$ in Figure 4.8 as well as $\underline{D}_{\text{odds}}$ & $\underline{D}_{\text{prev}}$ too, a small number of outliers at ∼1,500m can strongly influence the upper bound of $\hat{D}$, and sometimes its stability (Figs. 4.8 & 4.9). When all three estimators are plotted together, tight synchronicity is present in their local minima & maxima, respectively (Fig. 4.10). $\tau^{(d)}(d_1, d_2)$ variance remains higher at closer distances than far.

A pseudo-$\tau$ function represents the Chapman et al. [44] model-based estimate in Figure 4.10 by dividing their spatially-continuous force of infection $\big(syn.\ \text{rate}, \lambda(d)\big)$, by the traditional (non-

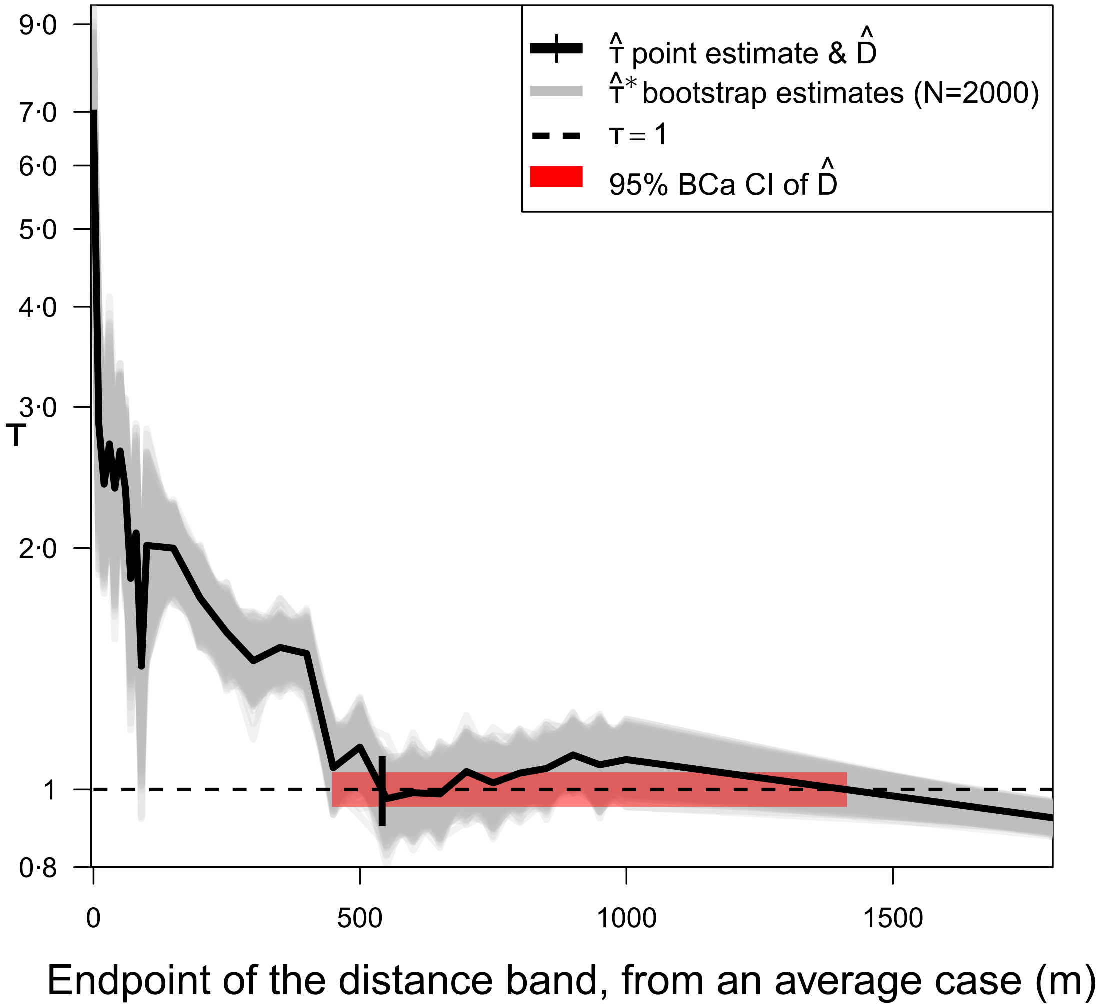


Figure 4.8: **Estimating the clustering endpoint distance** $\hat{D}_{\text{rate}}$. CI's diagnostic checks passed. Distance band set $:= \big\{[0,10\text{m}),[10,20\text{m}),\ldots,[100,150\text{m}),[150,200\text{m}),\ldots,[1,2\text{km}),\ldots,[5,6\text{km})\big\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo].

spatial) rate estimate evaluated over the entire study $\big(\lambda(0,\infty) := {}^{\text{total cases}}\!/\!_{\text{total person-months at risk}}\big)$. The model estimate $\tau^{(d)}_{\text{model}}$ is much higher than the other $\tau$ estimators at distances closer than 200m, while $\tau^{(d)}_{\text{model}} = 407\text{m}$ has the best agreement with $\tau^{(d)}_{\text{rate}}$, which is 135m higher.

### 4.3.5 Estimating the clustering endpoint time, $\hat{T}$ (in theory)

The proper inferential step is to estimate $\hat{T}$, contingent on having estimated $\hat{D}$, as more spatiotemporal data are available for a more precise $\tau^{(d)}$ to perform the graphical hypothesis test & estimate $\hat{D}$ than if $\tau^{(t)}$ was first used. No hypothesis testing for spatiotemporal clustering using $\tau^{(t)}$ is necessary using the time-form as established with $\tau^{(d)}$. While $\hat{D}$ is implicitly conditional on *temporal* relatedness

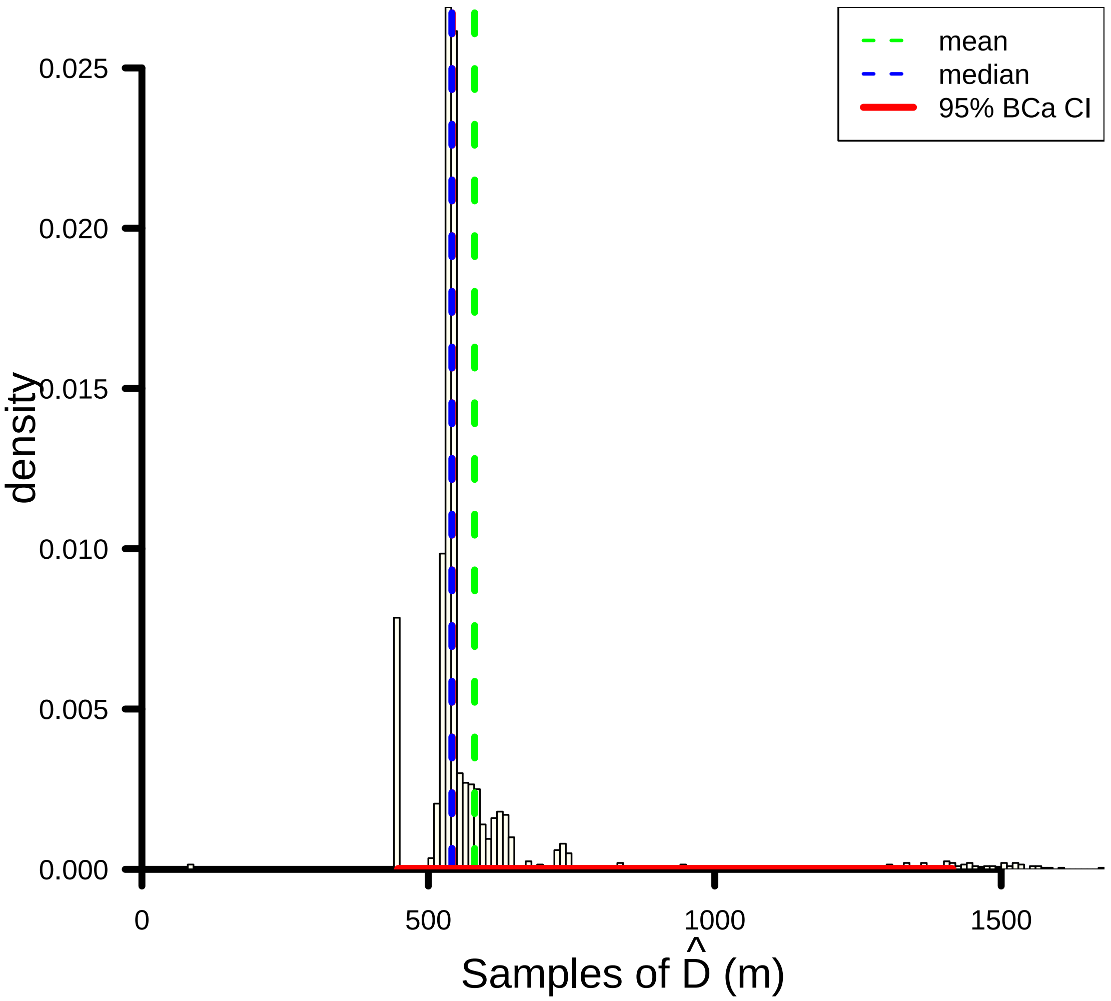


Figure 4.9: **Histogram of $\underline{D}$ estimates from Figure 4.8 using $\tau^{(d)}_{\text{rate}}$.** Point estimate $\hat{D}$ as a red dashed vertical line, with median (blue) & mean (green) of the bootstrapped distribution. The red horizontal line along the x-axis is the 95% BCa CI constructed from the 2,000 bootstrapped simulations which passed diagnostic checks. Distance band set $:= \big\{[0, 10\text{m}), [10, 20\text{m}), \ldots, [100, 150\text{m}), [150, 200\text{m}), \ldots, [1, 2\text{km}), \ldots, [5, 6\text{km})\big\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo].

parameters $T_1$ & $T_2$ sourced from epidemiological investigation, $\hat{T}$ is implicitly conditional on $d_1, d_2$ through *spatial* relatedness $z^{(t)}_{ij} = \big\{1 \text{ if } d_{ij} \in [d_1, d_2); \text{ else } 0\big\}$. The user could arbitrarily choose these between 0m & the maximum distance separation of the data. To guard against ‘data fishing’, it is sensible to start with tau-distance measures to obtain $\hat{D}$ as informed by prior biological/disease knowledge, and then estimate $\hat{T}$ conditional on this, with $z^{(t)}_{ij} = \big\{1 \text{ if } d_{ij} \in [0, \hat{D}); \text{ else } 0\big\}$.

A $\tau^{(t)}$ vs. time graph (*e.g.* Fig. 4.11) is different from $\tau^{(d)}$ vs. distance plots. The point estimate for a single distance band $\tau\big([d_l, d_m)\big)$ has little practical utility (as public health interventions are not delivered to selected (non-disk) annuli around an average index case). However, a time-version

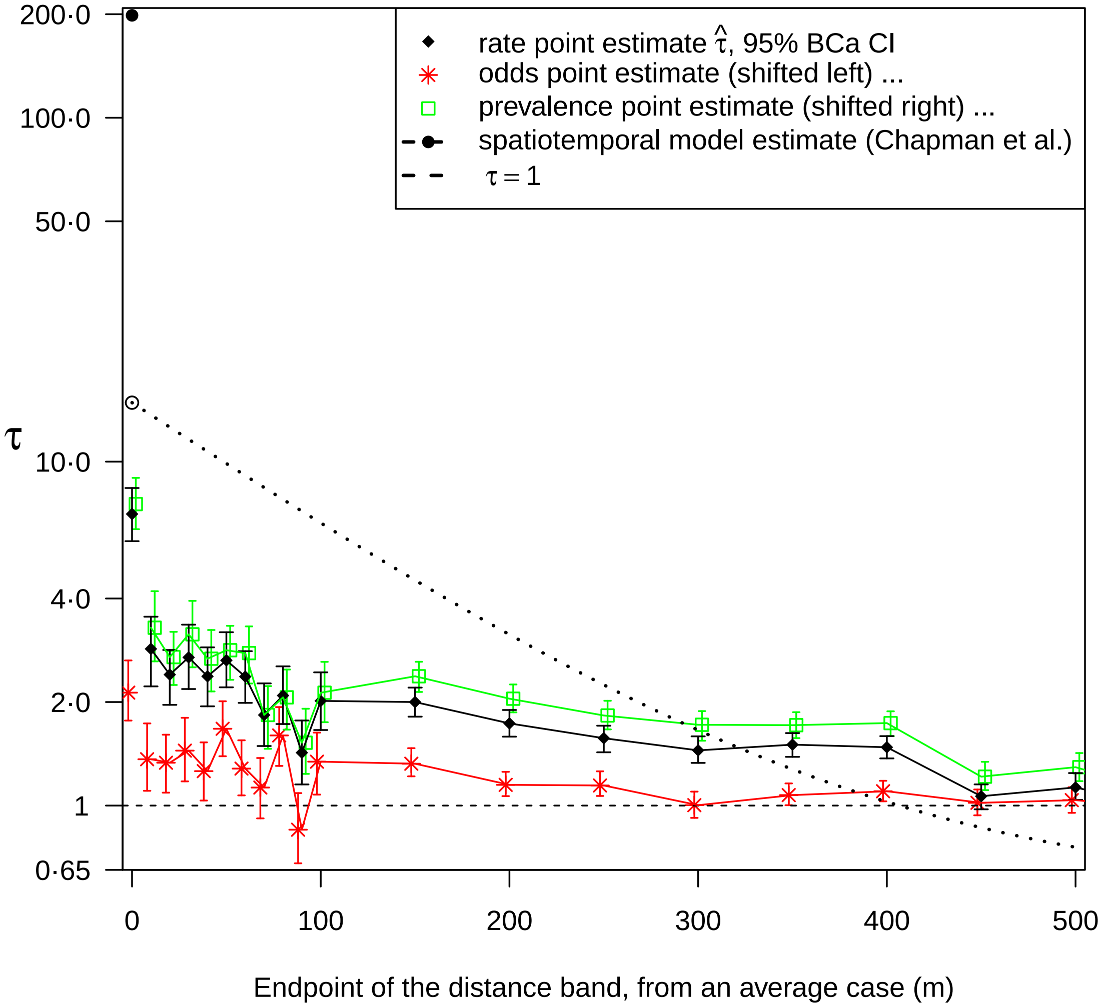


Figure 4.10: **Tau-distance statistic $\tau^{(d)}$ by estimator type**. The pseudo-$\tau$ function was included, based on the force of infection profile in Chapman et al. [44] for the same dataset; open circle plot '○' at $d = 0$ represents the rate *just-outside* household—solid circle '●' *within*. 2,000 bootstraps per pointwise CI. CIs passed diagnostic checks. Distance band set $:= \{[0, 10\text{m}), [10, 20\text{m}), \ldots, [100, 150\text{m}), [150, 200\text{m}), \ldots, [1, 2\text{km}), \ldots, [5, 6\text{km})\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo].

$\tau\big([t_l, t_m)\big)$ is more useful as one can envisage health workers returning to a disc region around an index case at a specific time after symptom onset. Note that, as $z_{ij}^{(t)}$ is defined $\big($using $\hat{D}$ obtained earlier from $\tau^{(d)}\big)$ without using SIs, it can no longer differentiate between temporal relatedness of VL or PKDL infectors.

In Figure 4.11 the $\tau_{\text{odds}}^{(t)}$ & $\tau_{\text{prev}}^{(t)}$ point estimates have slight separation, however $\tau_{\text{prev}}^{(t)}$ does consistently have lower variance and is recommended over $\tau_{\text{odds}}^{(t)}$ for $\tau^{(t)}$ analysis unless computational time is an issue.

During the first three months after onset, $\tau^{(t)}$ is significantly above 1 for $t \in \{0\text{–}1\}$, for all estimators (Fig. 4.11). Those times of operational significance indicated by their magnitude are $\tau^{(t)}_{\text{prev}}(0, 0\text{mo}) = 1{\cdot}27$ & $\tau^{(t)}_{\text{prev}}(1, 1) = 1{\cdot}11$. Although multiple ideal surveillance times are provided, a precision estimate would only be available for $\hat{T}$.

Unfortunately, the $\hat{T}$ result for VL & PKDL is invalid for use (Fig. 4.13). Although its first successful calculation here has demonstrated the calculation method, $\hat{T}$ & its CI is outside the threshold ($\leq$3mo) set to reduce the effects of temporal confounding (§4.2.4.1).

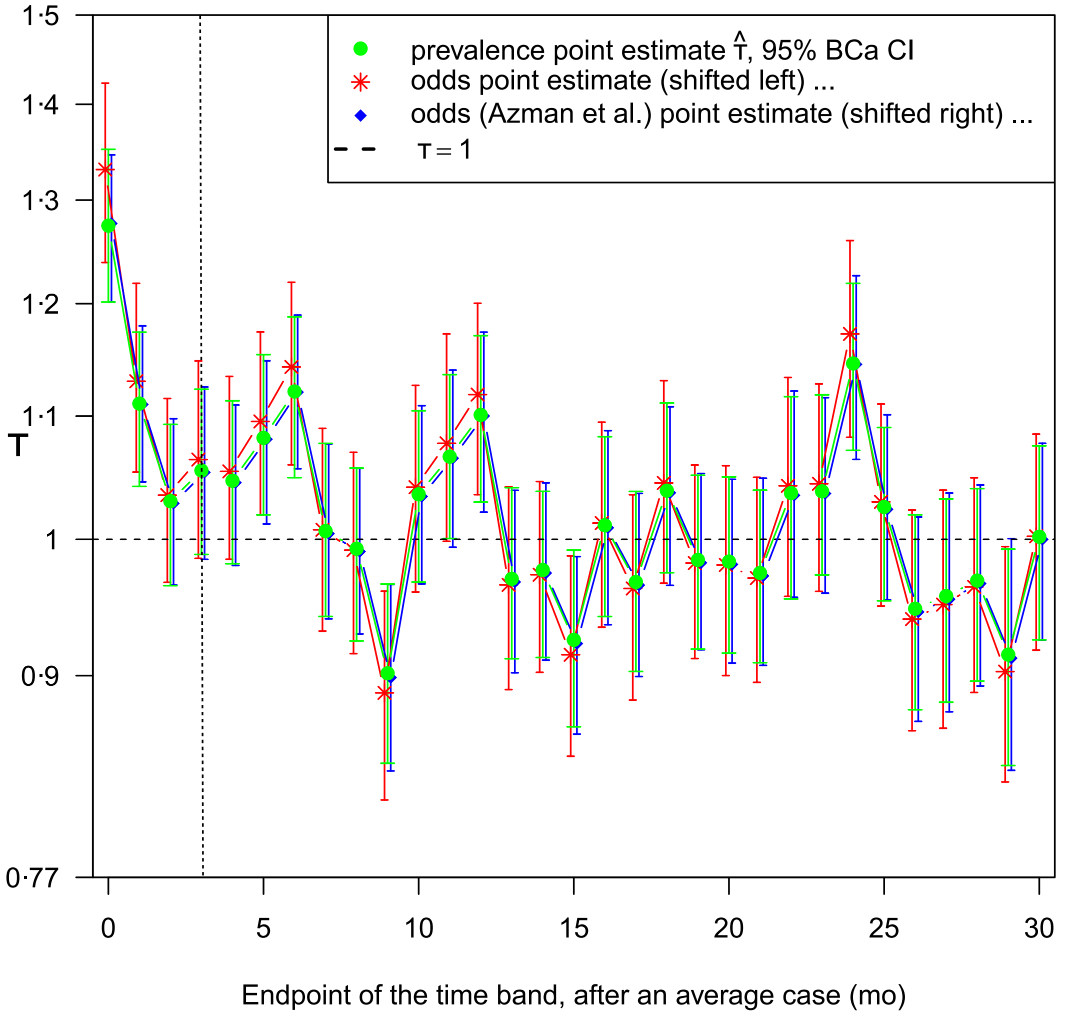


Figure 4.11: **Tau-time statistic $\tau^{(t)}$ by estimator type** for distances 0–550m around an average VL case in Fulbaria village, Bangladesh, January 2002–June 2009. CIs: using 2,000 bootstrap samples & diagnostic checks passed. Time band set := $\{[0, 1\text{mo}), [1, 2\text{mo}), \ldots, [30, 31\text{mo})\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo]. To reduce the impact of temporal confounding on results (§4.2.4.1), we warn the reader to only use $\tau^{(t)}$ values that are in the first three months of index case onset—to the left of & including the threshold indicated by the vertical dashed line.

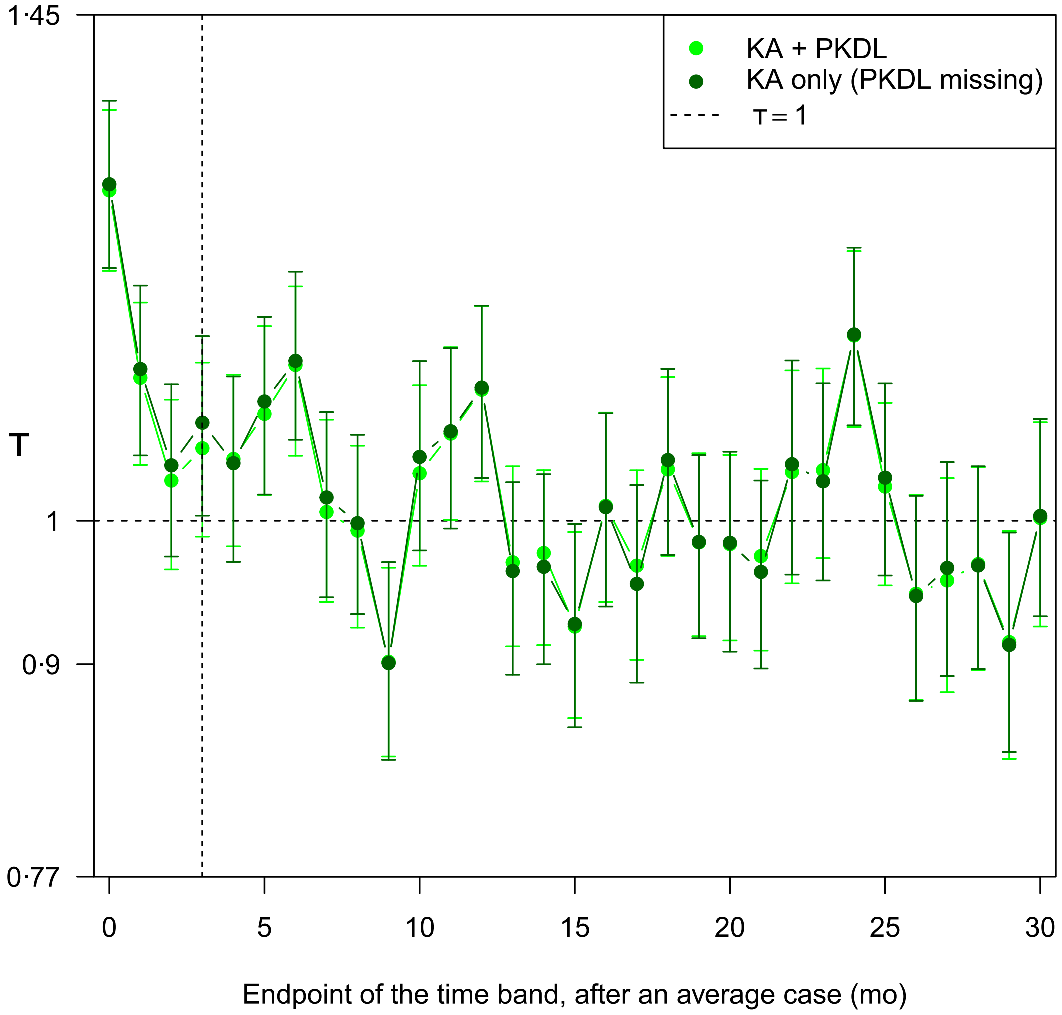


Figure 4.12: **Comparing the prevalence tau-time estimator $\tau^{(t)}_{\text{prev}}$ of all cases (VL + PKDL) vs. VL-only (PKDL-missing) data** for the distance disc 0–550m. 95% BCa CIs constructed from 2,000 bootstraps, which passed diagnostic checks. Time band set $:= \big\{[0, 1\text{mo}), [1, 2\text{mo}), \ldots, [30, 31\text{mo})\big\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo]. To reduce the impact of temporal confounding on results (§4.2.4.1), we warn the reader to only use $\tau^{(t)}$ values that are in the first three months of index case onset—to the left of & including the threshold indicated by the vertical dashed line.

### 4.3.6 Tau-time odds ratio estimator $\tau^{(t)}_{\text{odds}}$: corrected vs. Azman et al.'s

Although Chapter 2's $\tau^{(t)}_{\text{odds}}$ estimator (Eqn. 2.7) is preferred over Azman et al. [8] because of its logical construction (§2.4.4), there is little difference in their values, and Azman et al.'s has a lower variance. It is unclear if their previous cholera findings [8] would be different using our estimator (Eqn. 2.7).

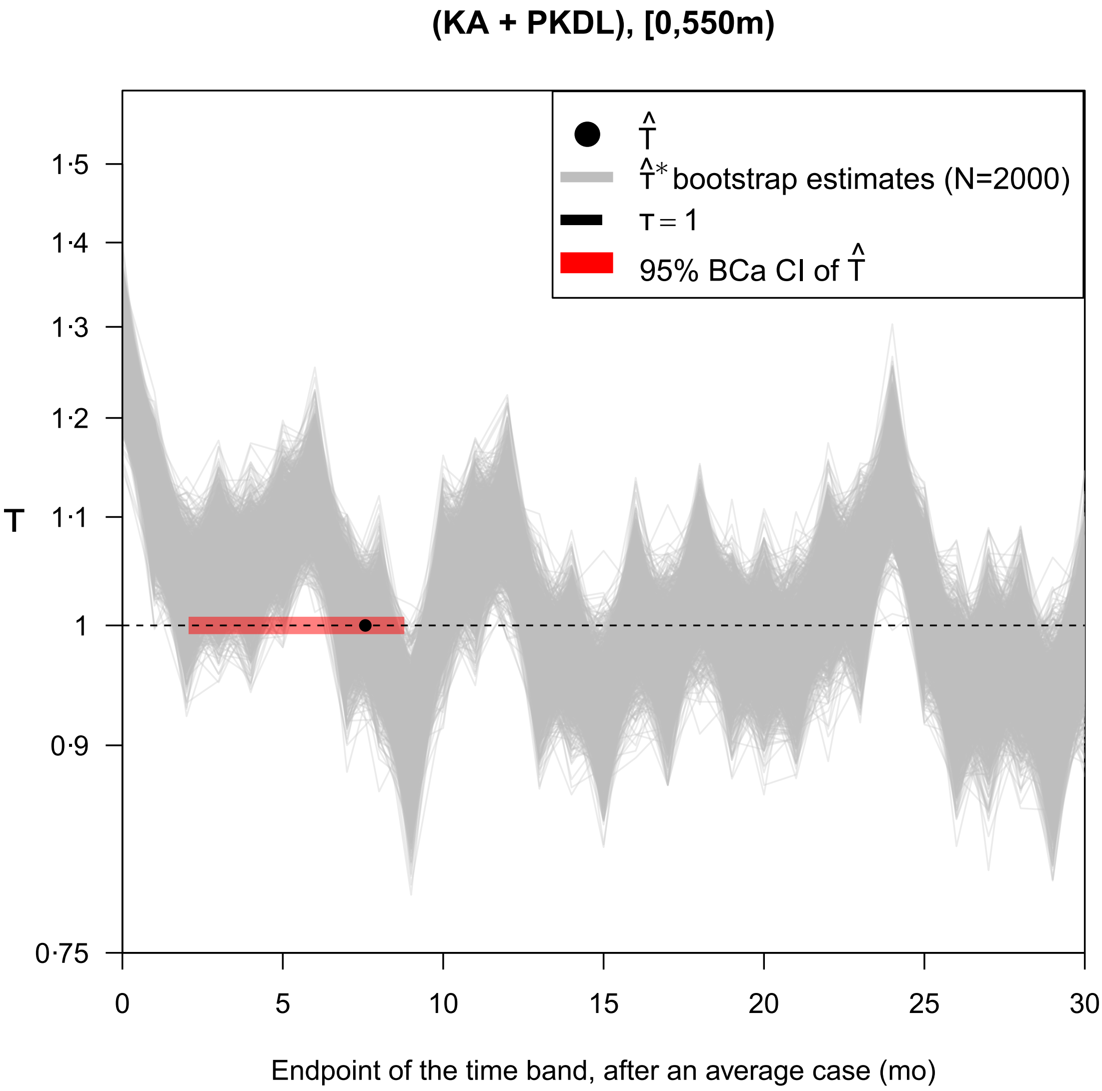


Figure 4.13: $\hat{T}_{\text{prev}}$ **estimate based on distances 0–550m around VL + PKDL cases**. CIs passed diagnostic tests. 2,000 bootstrapped simulations. Time band set $:= \big\{[0, 1\text{mo}), [1, 2\text{mo}), \ldots, [30, 31\text{mo})\big\}$. VL cases are considered temporally related within [0,7mo] & PKDL within [0,15mo]. $\hat{T}_{\text{prev}} = 7{\cdot}6$ months with a skewed CI (BCa 95% CI 2·1, 8·8mo); it is explained in §4.3.5 how although this figure demonstrates the first successful calculation of $\hat{T}$, it is far outside the threshold ($\leq$3mo) considered to avoid some of the effects of temporal confounding and so $\hat{T}$ should *not* be used to inform the results, only to demonstrate a successful method.

### 4.3.7 Computational speed

The odds estimator has the fastest computational speed due to the reduced case-only dataset. The additional loops to count PTAR makes computing the rate estimator longer than the odds estimator. For this dataset of around a thousand cases, on a high-performance cluster where 20 cores are used in parallel, the time to construct a BCa CI (using 2000 bootstraps) for a single point estimate of the distance-form is 44s for the odds estimator, 975s for the prevalence estimator & 548s for the rate estimator.

## 4.4 Discussion

### 4.4.1 Tau statistic estimate for $\hat{D}$

The clustering range estimated for VL using the non-parametric tau statistic $\tau^{(d)}_{\text{rate}}$ (rate estimator) is similar to a more complex, parametric model Chapman et al. [44] and a statistical hotspot analysis at the village-level finding a modal radius of ∼500m [33], which improves further the confidence of the evidence for a clustering range of VL of approximately 500m around a case. Note that the inferred infection times & serological states in the model are not validated and so the model estimate does not necessarily constitute a gold standard. For diseases with long & variable IPs like VL & HIV, the tau statistic satisfactorily captures a spatiotemporal signal with reasonable precision in its point estimates $\tau^{(d)}(d_1, d_2)$. However, the precision of its derived estimate $\hat{D}$ is much poorer, which may limit its usefulness for targeted disease control. Nevertheless, it is believed that the additional temporal information captured by $\tau^{(d)}_{\text{rate}}$ (time-varying location & entry/exit times) improves the accuracy of the clustering estimates it provides. Future epidemiological studies designed with the tau statistic in mind should collect all disease states relevant to infectiousness & susceptibility. Additionally, for the rate estimator, 'case & non-case' location & onset time, the entry/exit, treatment times & previous infectious episodes should be collected.

### 4.4.2 Comparison of the three estimators: $\tau^{(d)}_{\text{odds}}$, $\tau^{(d)}_{\text{prev}}$ & $\tau^{(d)}_{\text{rate}}$

As $\tau^{(d)}_{\text{odds}}$ only uses cases and thus less data, it therefore has a higher variance. Missing out non-cases appears to underestimate $\tau$, and thus $\hat{D}$, versus $\tau^{(d)}_{\text{prev}}$ or $\tau^{(d)}_{\text{rate}}$. $\tau^{(d)}_{\text{rate}}$ surpasses $\tau^{(d)}_{\text{odds}}$ in point estimation & precision: close to an infectious case, the PTAR (denominator of $\lambda$) would be lower than further away due to the depletion of local susceptibles, making $\lambda$ $\big($the numerator of $\tau^{(d)}_{\text{rate}}\big)$ at close distances larger and explaining why $\tau^{(d)}_{\text{rate}}$ is bigger than $\tau^{(d)}_{\text{odds}}$—yet this does not explain the $\tau^{(d)}_{\text{prev}}$ result. Whether $\tau^{(d)}_{\text{rate}}$ is better than $\tau^{(d)}_{\text{prev}}$ could be setting-dependent. If the average PTAR (as a function of time-in study & infection-acquired immunity) is a small proportion of the total study time, varies temporally (due to seasonal migration, say) or spatially (since the infection process causes susceptibility loss in neighbours and thus clusters of lower PTAR), then it is still hypothesised that $\tau^{(d)}_{\text{rate}}$ would provide more accurate estimates. This requires simulation models to understand if variable PTAR within the data is differentiated by tau estimator type and thus the results of the inference stages signposted in §4.3.2–4.3.5.

Model complexity may explain why the model-based pseudo-tau estimates up to 200m are higher than $\tau^{(d)}_{\text{odds}}$, $\tau^{(d)}_{\text{prev}}$ & $\tau^{(d)}_{\text{rate}}$. The model constructs infection chains that are self-consistent with a larger transmission chain, including unobserved asymptomatic cases, whereas the broader definition of relatedness in $z^{(d)}_{ij}$ or $Z^{(d)}_{ij}$ can only identify *single* pairs of *probable* transmission for observed cases. However, $\tau$ estimates can increase as more relatedness variables (*e.g.* serology or genetic relations) are added [118], so these tau values may be underestimated.

$\tau^{(d)}_{\text{rate}}$ does not perform optimally against $\tau^{(d)}_{\text{prev}}$, which may be explained because temporal relatedness $Z^{(d)}_{ij}$ is defined as the longer overlap of infectiousness *and* susceptibility rather than $z^{(d)}_{ij}$ defined on single pairs of onsets. Furthermore, basing $\tau^{(d)}_{\text{rate}}$ on observable symptom onset rather than infection time would misalign the start of the infector's infectious period and over-count the end of the infectee's susceptible period, both by a single latent period.

### 4.4.3 Relevance to VL surveillance & future elimination policy

Although the tau-time statistic can be used to assess additional temporal clustering within an infection-competent distance of an index case, a degree of care is needed when interpreting the results due to temporal confounding in our tau-time analysis (4.2.4.1). A straightforward approach to account for this confounding could not be identified. Therefore, it is sensible to restrict the time range to the periodicity of the seasonal pattern (*i.e.* 3mo) when reporting $\hat{T}_{\text{prev}}$.

Based on this analysis, performing ACD or vector control in a 550m radius around a case at 0–1 months after symptom onset could increase the number of secondary cases detected. These estimates provide information on household follow-up times for any future VL elimination survey policy planned for Bangladesh, extra to existing policy based on a crude non-spatial, empirical epidemiological estimate. Additionally, applying an estimate from a high-endemicity setting may not generalise to an elimination setting. Door-to-door surveillance within a 550m disc of an index case would cover on average 135 study households for this rural setting, which is feasible over a month for one community health worker. The invalid $\hat{T}$ estimates here are likely to be repeated for other VL analyses (due to the seasonal incidence pattern commonly observed) which occurs at a period similar to or shorter than a single infection generation. However, for infectious diseases during stable incidence, $\hat{T}$ analysis could be valuable for informing control policy in high-density or low-resource settings, where it is not possible to perform exhaustive house-to-house searches at all times or in all places.

This policy will prove impractical above a particular incidence, where separate index cases occur near to (*i.e.* within 550m of) each other and close together in time ($\leq$ 1mo apart), so a regular village-specific calendar time interval would be logistically easier than an index-case–led schedule.

Lessler et al. [118] warn that $\tau^{(d)}_{\text{prev}}$ indicates annuli of elevated *prevalence* only not *risk* of infection, and as a corollary, the rate estimator provides a measure of elevated rate, not risk. Disease frequency measures provide information of current disease burden, whereas to reduce new transmission, targeting reduction of infection risk is more relevant. Although $\hat{D}$ can still guide disease control, modelling that simulates risk and thus new cases (and asymptomatic transmission) like in [44] is required for a definitive intervention radius. $\hat{T}$, on the other hand, directly outputs the expected burden of disease in the first months following an index case, which is precisely what is needed for active surveillance. $\hat{T}$ should not be used as a direct measure to improve control *efficiency*, *c.f.* [8]. This would require outcome metrics like 'total cases averted within a period' or 'time to disease elimination threshold' and a way to account for varying endemicity—so again, a simulation study is essential. The *Policy-relevant items for reporting models in epidemiology of NTDs* (PRIME-NTD) summary (Appendix C.1) are provided to assist in communicating "the quality & relevance of modelling to stakeholders".

### 4.4.4 Sensitivity & robustness

Long-distance outliers (whether chance artefacts or real secondary cases of an index case) appear to strongly influence $\hat{D}$, pushing $\hat{D}_{\text{odds}}$'s upper CI bound to 1·4–2km. This distance seems improbable for direct transmission by a single sandfly vector since the furthest flight range observed for a single *P. argentipes* female is 309m (even though this was out of 223 sandflies released in a field experiment lasting only $2\frac{1}{2}$ days, so further spatial dispersal could have occurred [158]). There are three credible explanations: a) a moving (human) index case $i$ seeded cases in two separate areas which progressed in unison, producing apparent signals of spatiotemporal clustering when no real interaction occurred or b) a case $i$ infected an unobserved case $j$ over a very short SI, $j$ then infected observed case $k$ at a more considerable distance from $i$ across another small SI and the resulting observed onset difference $t_k - t_i$ was still within a single mean SI, or c) as viewed from a $\tau^{(d)}$ vs. distance graph, it is natural for the set of bootstrapped connected lines that varies with noise about the point estimate line, to graze $\tau^{(d)} = 1$ at a shallow gradient, which amplifies any 'vertical' tau variance into a much larger $\hat{D}$ variance 'horizontally'.

### 4.4.5 Recommendations for future study

To help in understanding outlier sensitivity of the tau statistic & what data points may contribute to this, one could analyse the inferred transmission chains in the Chapman et al. model to see the proportion that transcended 1·4–2km. Inspired by leverage & influence concepts from regression analysis, to understand the sensitivity of $\tau$ to outlier data points, the top 1% of 'influential' data points could be identified using a 'leave-one-out' algorithm. What unique characteristics do these outliers have compared to the rest of the data? Does dropping them from the dataset en-masse significantly affect the point estimate or range of the $\underline{D}$ distribution? The $\hat{D}$ estimate could be made more robust by trimming the $\underline{D}$ distribution by a 99% highest density interval [128] to attenuate the distribution's tail.

The inconsistency in the range of clustering computed from the $\tau_{\text{prev}}$ or $\tau_{\text{odds}}$ estimators with the model-based estimate (if the latter is true) is contrary to simulation-based findings in Lessler et al. [118]. However, new cases are allowed to arise *anywhere* in space in their simulation and not restricted to household locations—*i.e.* a discrete spatial process. This may weaken their results that demonstrates $\tau$'s robustness to underreporting or heterogeneous sampling, which were also only available for $\tau^{(d)}_{\text{odds}}$ & $\tau^{(d)}_{\text{prev}}$ estimators at the time [118].

The rate estimator uses the infectious period of onset time to treatment time; however, this underrepresents the actual infectious period & its infectiousness profile. Future tests with a disease with this known period & profile could indicate how this could be improved.

It is unknown how the increasing fraction of asymptomatics (as inferred by Chapman et al.'s spatiotemporal model [44]) that deplete the locally-available susceptibles for new infection, would bias $\hat{D}$. The role of asymptomatics could significantly affect $\hat{D}$. If found to be significant then new protocols could be developed where rK39 community surveys perform a serological survey on a subsample of a population to estimate the proportion asymptomatic and thus tailor subsequent control or new vaccination strategies accordingly.

## 4.5 Conclusions

A new rate form of the tau statistic for measuring global spatiotemporal clustering is developed & applied, $\tau^{(d)}_{\text{rate}}$, which accounts for variable PTAR. There are encouraging signs from this analysis that it could be an appropriate estimator for other PTAR-variable datasets with migration or changing immunity. For the field of VL, a previous model-based estimate of clustering is partially-validated—confirming that it was household dominant & ∼400–500m—but the tau estimators showed poor precision. Following principled inferential stages, an additional step is formalised to estimate $\hat{T}$. Important warnings on extrinsic seasonal incidence patterns advise not to analyse beyond a single seasonal period of incidence.

In addition to the open issues surrounding the tau statistic already identified in §2.7 & §3.4, it is unclear the threshold at which the rate estimator becomes beneficial for variable PTAR; and the sensitivity of the clustering range endpoint estimate to long-distance outliers. Also, it is unknown how informing control using the spatial or temporal clustering range estimates could impact public health outcomes. This completes the study of the tau statistic and its application to a household-level VL dataset. In the next chapter, how the disease clusters at the district level informs a (spatiotemporal) statistical model to answer some pressing VL research questions.

CHAPTER 5

# Impact of intensified control on VL in a highly-endemic district of Bihar, India: an interrupted time series analysis (ITSA)

*Continuing with VL control but at the district spatial scale, a recent VL intervention in Bihar, India is analysed. Using a spatiotemporal statistical model of intensified control I ask what the additional impact of this intervention was in a highly-endemic district. VL incidence decreased faster in the pilot district compared to other districts, with several hundred VL cases estimated to have been averted during 2015–2017. These recent findings are relevant following the new WHO NTD 2021–2030 roadmap. So this result is framed in how it may contribute to policy evidence for future VL intervention planning.*

## Abstract

VL is declining in India as the WHO's EPHP target, set for 2020, has nearly been achieved. Intensified combined interventions might help reach elimination, but their impact has not been assessed. WHO's NTD 2021–2030 roadmap provides an opportunity to revisit VL control strategies. The combined effect of a district-wide pilot of intensified interventions in the highly-endemic Vaishali district is estimated, where cases fell from 3,598 in 2012–2014 to 762 in 2015–2017. The intensified control approach comprised IRS with improved supervision; VL-specific training for accredited social health activists (ASHAs) to reduce onset-to-diagnosis (OD) time; and increased Information, Education & Communication (IEC) activities in the community. The rate of incidence decrease in Vaishali is compared to other districts in Bihar state via an ITSA with a spatiotemporal model informed by previous VL epidemiological estimates.

Changes in Vaishali's rank among Bihar's endemic districts in terms of monthly incidence showed a change pre-pilot (4$^{\text{th}}$ highest out of 33 reporting districts) vs. during the pilot (11$^{\text{th}}$) ($p < 1 \times 10^{-10}$). Counterfactual model simulations suggest an estimated median of 352 cases (IQR 234–477) were averted by the Vaishali pilot between January 2015 & December 2017, which was robust to modest changes in the onset-to-diagnosis distribution. Strengthening control strategies may have precipitated a substantial change in VL incidence in Vaishali and suggests this approach should be piloted in other highly-endemic districts.

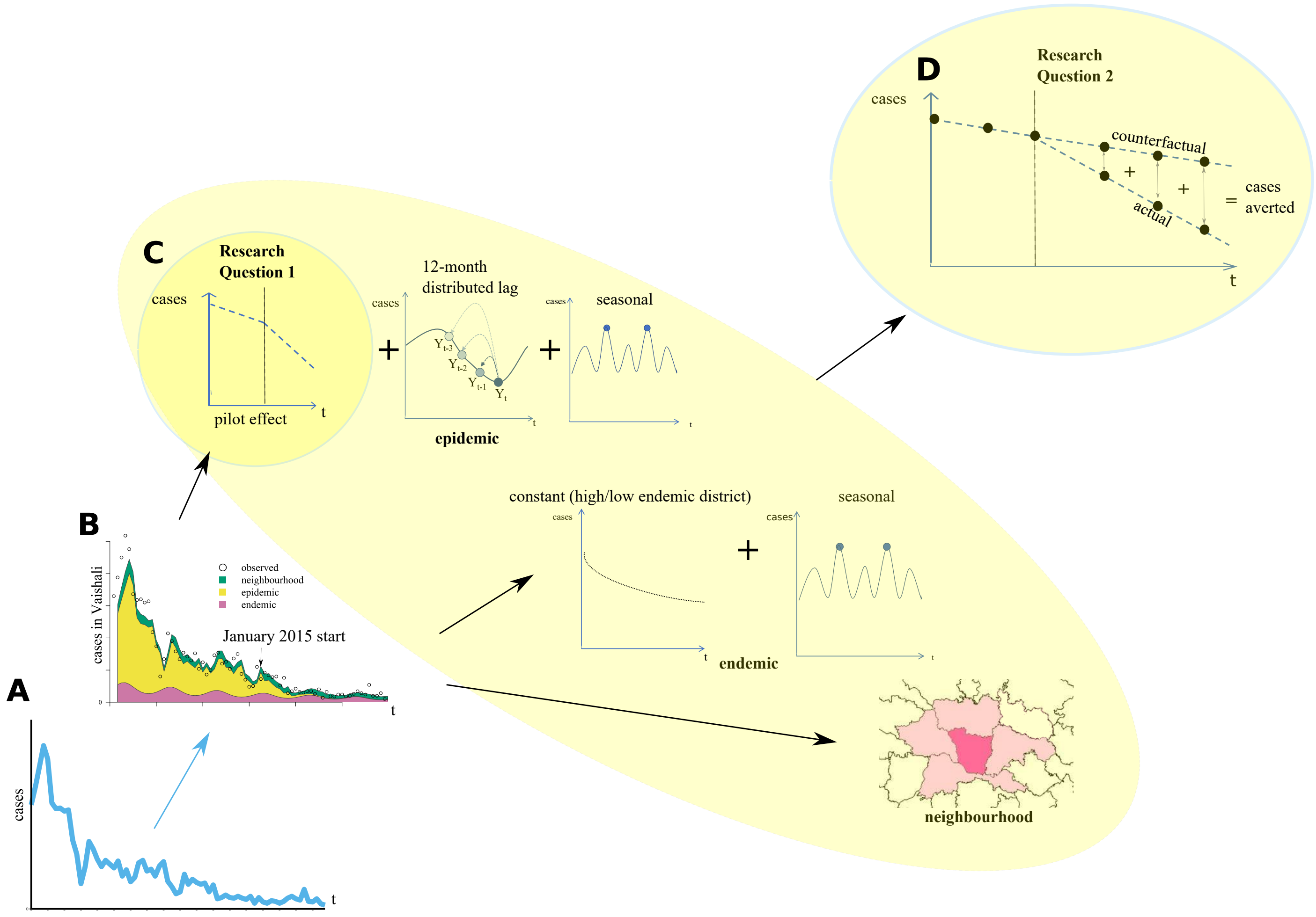


Figure 5.1: **Summary of the modelling framework**. Vaishali district incidence (blue line in **A**); model composed of epidemic, endemic & neighbourhood infection processes (**B** & **C**); pilot effect (**C**) represents the additional contribution that intensified control made amidst the declining state-wide incidence; the counterfactual model is used to predict the cases that would have happened had no pilot occurred—to estimate the cases averted (**D**). Research questions are fully described in §5.1.4.

## 5.1 Motivation

India had an estimated 146,700–282,800 VL reported cases annually between 2004–2008, most of which were from Bihar state [6]. VL cases have declined since 2011 but have plateaued slightly in recent years (2014–2017) [142, 205]. The WHO 2020 target for elimination of VL as a public health problem ($<$ 1 case/10,000 people/year at block (subdistrict) level) [92, 201] has now passed, and only ∼2% of blocks are still above the target (February 2021), but resurgence may still occur—as seen in previous decadal cycles [61]. Therefore, this recent intensified pilot analysis is still relevant for future control policy. Current interventions implemented by the National Vector Borne Disease Control Programme (NVBDCP) for routine VL control in Bihar state involve biannual IRS of insecticide at state-level, passive case detection (PCD) at (block-level) by primary health centres & ACD by ASHAs; or via annual mobile camps (§5.1.2).

### 5.1.1 Existing research base for interventions

An IRS review in Bangladesh, Nepal & India showed that the IRS had an impact on sandfly densities when properly conducted but did not significantly impact VL case incidence [154]. The only large-scale randomised control trial of a vector control intervention on infection incidence (the KALANET project) found no evidence that large-scale distribution of long-lasting insecticidal nets provided additional protection over existing control practices [155, 156]. A multi-site ACD screening intervention by ASHAs in highly-endemic Muzaffarpur & Saran districts discovered 6·7–17·1% more cases than PCD alone [93]. Overall, robust evidence is lacking on intervention effectiveness from field trials. Nevertheless, it is hypothesised that a *combination* of strengthening the ACD referral system through VL-specific training for ASHAs, higher quality IRS by well-trained and supervised spray teams & IEC community activities (§5.1.3) could produce measurable incidence reductions. It is expensive to run a control programme of this scale: requiring coordination between RMRIMS and the Ministry of Health & Family Welfare for administrative & logistical support and needing 166 spray squads for several weeks, twice a year (§5.1.3). Therefore, any policy decision to apply this costly intervention to highly-endemic districts in future requires an evidence base.

### 5.1.2 VL control under the national programme

#### 5.1.2.1 Routine control activities

- Early case detection & management, primarily as passive surveillance followed by annual ACD with a 'camp approach', which is less sensitive & uses a weak referral system. Liposomal amphotericin B in a single dose of 10mg/kg was the first-line VL treatment, and combination therapy (paromomycin-miltefosine injection for 10 days) as the second-line treatment followed by other regimens, *e.g.* amphotericin B emulsion, miltefosine (28 days) & amphotericin B deoxycholate in multiple doses as per availability; this was also the case for Vaishali district[1] under the pilot study.

- IRS using DDT (dichlorodiphenyltrichloroethane) in earlier rounds (50% wettable powder applied at 1g/m$^2$) and then alpha-cypermethrin (a synthetic pyrethroid: 5% wettable powder at 25mg/m$^2$) was introduced at different times across Bihar state in 2015 once DDT resistance in

[1] Note that throughout the chapter, appendices & code, all mentions of 'Vaishali' refer to the district within Bihar state, rather than the smaller Vidhan Sabha constituency with the same name, which is within Vaishali district.

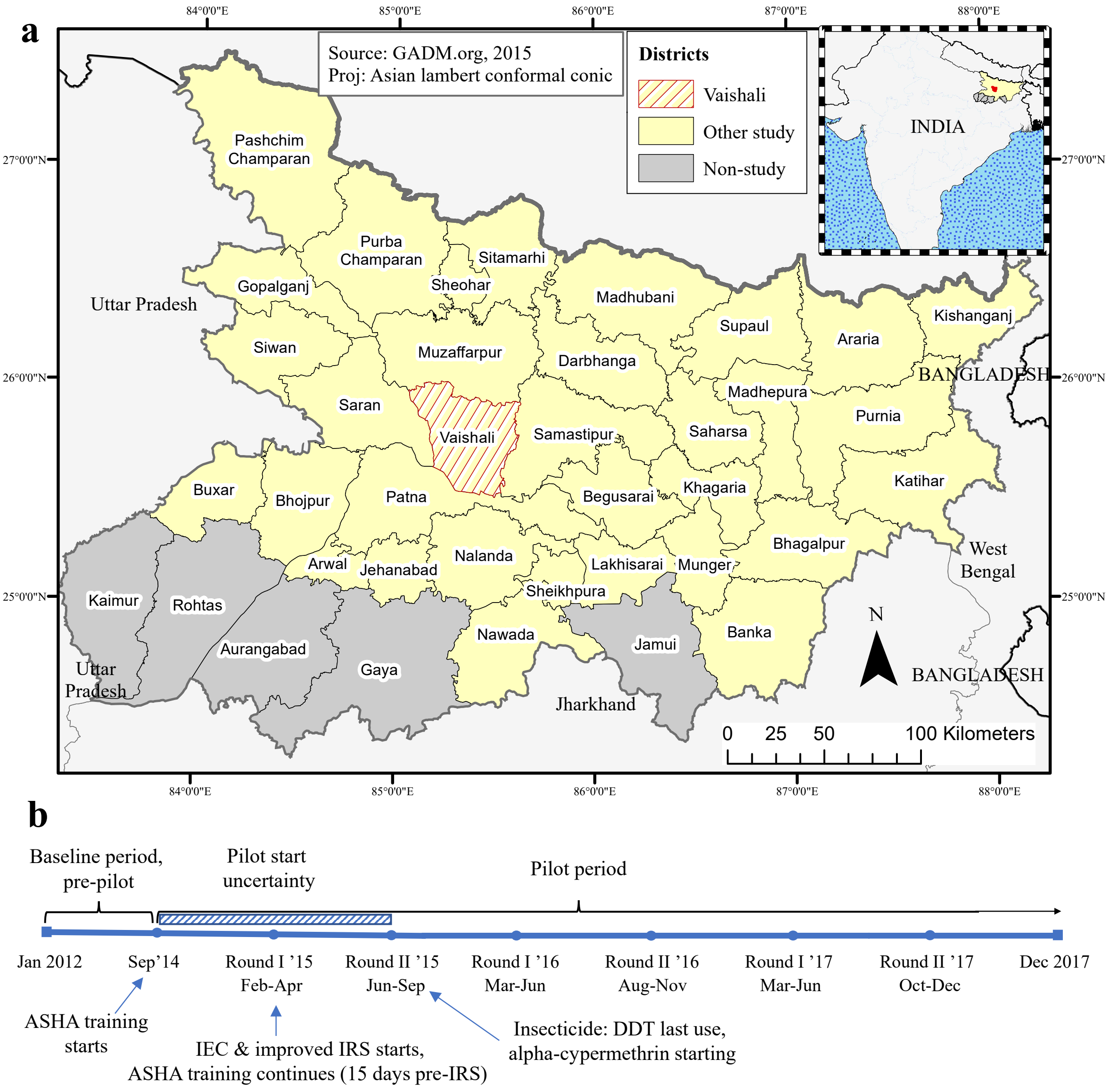


Figure 5.2: **Study map & timeline**. **a)** The pilot district of Vaishali is the hashed region. GADM shapefile [75]. **b)** Annotations indicate the start months of the intensified control elements, and circular dots mark the biannual ASHA training, IEC & IRS training rounds. The hatched bar marks the period of pilot scale-up when the combined methods would unlikely have reached full impact. Made in ArcMap™.

sandflies was detected. During the two IRS rounds, insecticide was sprayed in human dwellings & cattle sheds up to 1·8m in height. Usually, the first IRS round started February–March and then May–June for the second. Village selection was based on passive case reports, *i.e.* any village or hamlet reporting VL in the past 3 years qualified for 100% IRS coverage in that round. Districts of Bihar typically received varying levels of supervised IRS since there was no squad-level supervision, only at a block level.

- Unstandardised IEC activities with low coverage.

Neither the pilot nor comparison districts were known to have been supplied insecticide-treated nets by RMRIMS, NVBDCP or others. However, the WHO TDR programme did provide logistical support across Bihar.

#### 5.1.2.2 Management hierarchy

At the district level, IRS was monitored by one District Vector Borne Disease Control Officer & one District Vector Borne Disease Consultant. However, at the block (Public Health Centre or PHC, subdistrict)-level, IRS activity was managed by one Kala-azar Technical Supervisor, with at most one roving camp for ACD at any time.

#### 5.1.2.3 Early case detection & management

There had previously been accredited social health activist (ASHA) training in Bihar since 2012. A Grand Challenges Canada®-funded project in March–April 2012 & October–December 2013, conducted ASHA training in Paroo (Muzaffarpur district) & Marhoura (Saran district) blocks; whereas Sahebganj (Muzaffarpur) & Baniyapur (Saran) blocks received single training in October–December 2013 [55]; but as the training was not implemented comprehensively in these districts, it has not been included in the statistical model. From these four blocks, approximately 1,000 ASHAs were trained in groups of 100–150 by RMRIMS in VL/PKDL identification, transmission, treatment & IRS [55]. In 2014 the following districts' blocks also received two rounds of ASHA training ending in September 2014: Muzaffarpur (1/16), Saran (1/20), Siwan (1/19), Khagaria (1/7), Saharsa (7/10) & Vaishali (1/16) (blocks trained/total in parentheses); the single Vaishali block of Raghopur received two more rounds of ASHA training in September 2014 as part of this pilot study's intensified intervention for all 16 Vaishali blocks as described in §5.1.3 [54, 81].

### 5.1.3 Intensified control for Vaishali district under this pilot

RMRIMS conducted an observational study on the impact of intensified VL control covering 1,569 villages in all 16 blocks in Vaishali district in late 2014–early 2015 (when 15 blocks were above the elimination threshold), while standard control by the NVBDCP continued in other districts (Fig. 5.2 & §5.1.2 & §5.1.3) [108]. The triad of ongoing interventions, which began asynchronously (Fig. 5.2b & §5.1.3), are:

- specialised ASHA training (21–29 September 2014)
- improved IRS (from 15 February 2015) &
- IEC (19–21 February 2015).

Each block had their own VL control programme supervisor. Additionally, under intensified control, block-level supervisors were selected by and originated from RMRIMS based in Patna or from their respective block. However, all spraying squads were recruited from each block. Similarly, insecticide & pump equipment were delivered through the District Vector Borne Disease Control Office, Vaishali to blocks and then villages—in 2015, RMRIMS provided equipment in 7 Vaishali blocks out of 16, but from 2016, it was given to all blocks by the District Vector Borne Disease Control Office. Vaishali district had a total population of 3·50 million in 2011 [81]. RMRIMS staff supervised the pilot, which was composed of three elements:

#### 5.1.3.1 Early case detection & management

The case referral system was strengthened through ASHA training in various ACD approaches. A total of 2,431 new & existing ASHAs across all 16 blocks received two training rounds between 21–29 September 2014. Each trained ASHA worked exclusively on VL & PKDL case detection covering 200 households and was linked by name to the village's microplan, so they were separate from other ASHAs in Vaishali who were outside of this study and monitored other diseases. This training programme was repeated 15 days before each IRS round during the study. Complicated cases of VL were referred to the Samrat Ashoka Tropical Disease Research Centre Hospital (RMRIMS), Patna, Bihar.

The particular ACD approach used was context-dependent: house-to-house screening (blanket approach) in villages with five or more VL cases; or for newly-detected VL villages, the index case approach—where 50 m surrounding a newly-detected VL case is actively surveilled throughout the year. Furthermore, when high incidence was recorded in a focal area, then temporary mobile roving teams (camp approach) would intervene using four camps over 3–4 months, starting on the same months as IRS (Table 5.1). In the absence of cases, the standard passive surveillance was followed.

#### 5.1.3.2 Improved IRS

In Vaishali district, the IRS insecticides, their concentration & mode of application and village selection were identical to the national programme (§5.1.2) apart from the additions detailed here.

To conduct IRS activities in Vaishali, 24 block supervisors were selected & trained (as per WHO IRS monitoring & supervision criteria [200]) by RMRIMS and assigned to blocks. For IRS monitoring, monitors were also selected & trained by RMRIMS and assigned to each squad for each block. All block supervisors had a motorbike with daily fuel provision from RMRIMS. Monitors were selected from the locality for spraying, whereas squads were still recruited from the national IRS programme at the district level by the Vector Borne Disease Control Office (VBDCO), Hajipur. Insecticide, spray pumps & other equipment were delivered through the VBDCO to blocks and then to villages. The total IRS coverage during the study was 1,145 villages, however, this varied between rounds as villages' endemic status changed. IRS coverages by round were as follows: 1,144 villages (Round I 2015); 1,078 (Round II 2015); 995 (Round I 2016); 1,001 (Round II 2016); 1,046 (Round I 2017); 1,067 (Round II 2017).

Initially, supervised IRS with DDT was conducted at 90% household coverage within a block. Later, alpha-cypermethrin was introduced with quality checks, as earlier insecticide sensitivity tests had found DDT resistance. During the DDT era, the usual reason for not reaching full coverage was refusal or locked households, as the residents were away working in their fields [107]. Urban/peri-urban properties had higher refusal rates as people with lower socioeconomic status were worried about the effect on their retail goods. In contrast, people with a higher socioeconomic status did not think that it affected them or they had protection from living in concrete-walled properties. This indicates how supervision is key to IRS coverage.

| **Date** | **IRS round** | **Notes** |
|---|---|---|
| February–April 2015 | I | DDT.<br>IQK for 8 blocks.<br>All districts: use stirrup pump except Vaishali where 7 blocks use compression pumps (RMRIMS-obtained) & 9 use stirrup. 1 squad/monitor |
| June–September | II | Saran/Muzaffarpur/Vaishali: DDT for 1st 15 days then SP.<br>Other districts: DDT |
| March–June 2016 | I | SP.<br>Compression pump (LSTM-obtained) used in all districts.<br>2 squads/monitor |
| August–November | II | |
| March–June 2017 | I | |
| October–December | II | |

Table 5.1: **IRS schedule for Vaishali district**. DDT = dichlorodiphenyltrichloroethane; IQK = insecticide quantification kit; SP = synthetic pyrethroid.

A summary of the IRS schedule is described in Table 5.1. Two IRS rounds were conducted annually with the help of 166 spray squads. DDT was used in the first round of 2015 for all districts and in the second only for the first 15 days in Saran, Vaishali & Muzaffarpur districts, then alpha-cypermethrin after that. All districts received alpha-cypermethrin from 2016 onwards. In Vaishali district, IRS was performed using stirrup & compression pumps. In the first & second rounds of 2015, seven blocks used the compression pump while nine continued with the stirrup pump. From the first round of 2016 onwards, all 16 blocks used 'Hudson® X-Pert® Sprayer' compression pumps. To monitor the quality & coverage of the spraying activity, there were 166 monitors; in 2015, one monitor was assigned to each squad, whereas in 2016, a monitor covered two squads. A programme supervisor at the block level oversaw the work of the monitors.

Each squad comprised six members: five Field Workers & Senior Field Worker. In the teams with stirrup pumps, two pairs operated a pump each, one person mixed the insecticide, and the Senior Field Worker maintained the register and marked stencils onto the entrance of the sprayed house. Spraying with the compression pump enabled three people to spray with three pumps, while a pair made the solution and another acted as the Senior Field Worker.

Village selection in Vaishali district used GIS-based mapping of endemic & non-endemic villages for VL trends and hotspot analysis to prepare the microplans. In addition to endemic villages of the last three years, periphery villages of hotspot villages within 500 m of the endemic village boundary that had had a case in the previous year were also included [124]. This algorithm provided a list of villages for the spray team. NVBDCP decided the start of the first round according to the first seasonal surge in *P. argentipes* sandfly densities. The start of the second round was often dictated by the end of the first round plus a gap of twelve weeks, and by access, *e.g.* if flooding risk from the monsoon rains had diminished; conversely, it could not be too late due to people's reluctance to allow spraying before they re-decorated interior walls for October/November festivals.

During the study period, the existing stirrup pump was trialled against a compression pump, and the latter was found to deliver more uniform results [46]. To assess the quality of spraying, four methods were employed:

i) Visual checks by the senior field worker on the same or next day following spraying and re-spraying as necessary.

ii) Random checks of spraying quality by the squad monitor of 1 in every 10 of the 60–80 households in each village sprayed; they were re-sprayed if the quality was deemed poor on a household basis; if more than 50% of households were poor, then the whole village was re-sprayed.

iii) Sandfly density tests (as an indirect measure) using CDC light traps were performed in six houses in one village per block (*i.e.* 16 villages × 6 households = 96 households). One of each of three dwelling types was recorded: 'human only', 'human + cattle in same dwelling' or 'human only + adjacent cowshed'. Samples were taken during a single night at 15 days before IRS and 2, 4 & 12 weeks afterwards. Later, sandfly densities by *P. papatasi*, *Sergentomyia spp.* or *P. argentipes* were determined.

iv) IQKs (insecticide quantification kits) were used to take 3,000 household samples, each from four interior walls of one sleeping room per sampled household at 1·8m, 1·1m & 0·3m height, from eight districts during the first round of 2015. Previous research has shown the IQK's performance as comparable to high-performance liquid chromatography—the gold standard [97].

#### 5.1.3.3 IEC activities

Advice on preparing for the forthcoming IRS was first conducted during 19–21 February 2015 using audio broadcasts from auto-rickshaws. It was conducted at 1,196 locations, including marketplaces, private hospitals, block- & panchayat-level health centres, rural childcare centres (anganwadis), schools, state & central governmental offices, and households covered down to the ward level. Supporting literature over the course of the study covered banners (block-level $n = 44$; village-level $n = 495$), hoardings (marketplaces & government offices $n = 52$), posters ($n =$ 47,840), leaflets ($n =$ 95,680) & stickers ($n =$ 47,840). These activities continued 2–3 days before every spray round.

### 5.1.4 Research questions (RQs) & rationale for the spatiotemporal model

This study estimates:

RQ1: *whether intensified control additionally contributed to the decline in VL cases in Vaishali versus other districts*, and

RQ2: *how many VL cases are expected to have been averted by the pilot?*

Answering these questions is complicated since the incidence was already falling in Vaishali before the pilot started (Fig. 5.6). Crude calculations indicate decreasing case counts year-on-year: 664 in 2014, falling by 38·1% to 411 in 2015, and by 56·4% to 179 in 2016 [109]. To estimate the impact of the pilot while accounting for the decreasing background secular trend, Vaishali is compared with other districts rather than analysing it in isolation. The dynamics of case counts before & during the pilot was evaluated using a spatiotemporal framework [32, 86, 87, 130, 168]. The *pilot model* (§5.2.3.2) (*i.e.* the final model of the pilot study) is informed by prior VL epidemiology & spatiotemporal features of the setting (§5.2.3.1.1) [17]. To estimate the number of cases averted, the same model is fitted to a subset of pre-intervention months and counterfactual predictions of case counts made, with which observed case counts can be compared (Fig. 5.1 & §5.2.4.1).

## 5.2 Methods

### 5.2.1 Longitudinal dataset

Monthly VL case counts (by diagnosis date) for 33 out of the 38 districts of Bihar from January 2012–December 2017 were provided by the State Vector Borne Disease Office [82]. The analysis included HIV-VL cases from January 2015–December 2016 and HIV/TB-VL cases (coinfection of all three) from January 2017–December 2017 but excluded PKDL cases (Appendix D.2). The 33 study districts formed a contiguous island of transmission without the five remaining districts (Aurangabad, Gaya, Jamui, Kaimur & Rohtas are considered non-endemic) (Fig. 5.2a). Monthly district populations were estimated from 2001 & 2011 censuses [81]: The population is estimated by a monthly geometric progression using the decadal change between the 2001 & 2011 censuses; we assume a constant population increase during each decadal period. District shapefiles provided adjacency information [75]. The Institutional Ethical Committee of RMRIMS approved the intensified control programme (03/RMRI/EC/2018). University of Warwick's Biomedical & Scientific Research Ethics Committee (REGO-2018-2231) approved this analysis. Analysis code is available at github.com/t-pollington/ITSA. The study design, analysis & modelling are detailed in §5.1.2 & Appendix D.2 (DOI: 10.5281/zenodo.5701378) and a PRIME-NTD summary is provided (Table D.1).

### 5.2.2 Descriptive analyses

These crude analyses provided essential information prior to spatiotemporal model development:

- districts were compared by their ranked incidence levels & year-on-year changes in monthly incidence (§5.2.2.1). Changes in rank position enabled a crude comparison of the *relative* changes of Vaishali to other districts in the context of a state-wide medium-term decline in incidence. Using the two-sample two-tailed Wilcoxon test with continuity correction, it was assessed if the ranks before & during the pilot were different.

- evidence for global spatial correlation in incidence was assessed before & during the pilot with a Global Moran's I statistic hypothesis test (§5.2.2.2).

- the effective reproduction number $\hat{R}_e(t)$ for Vaishali & non-pilot districts was estimated to explore temporal patterns in transmission that may have been affected by interventions or seasonality (§5.2.2.3) [47, 48].

#### 5.2.2.1 Year-on-year comparisons of monthly incidence before/during the pilot

Ranked analyses were used as some differences are not visually discernible in time-series graphs featuring multiple districts. Incidences were computed by dividing monthly case numbers by the interpolated district population for that month. For each month, the year-on-year comparison of incidence was computed, then a rank was given to the districts according to which had: a) the highest incidence (highest = $1^{\text{st}}$ rank) & b) the largest negative percentage change in its incidence (largest negative change = $1^{\text{st}}$ rank). Comparisons before & during the pilot were then made by taking the rounded mean rank of each district for each period.

#### 5.2.2.2 Global Moran's $I$ for between-district spatial correlation

Alongside choropleth maps, Global Moran's I statistic was used to assess global spatial correlation in case numbers, for which a positive value ($I > 0$) indicates clustering of similar case numbers; a negative value indicates clustering dissimilarity in case numbers; and a zero value indicates no correlation and serves as the null for the subsequent hypothesis test.

#### 5.2.2.3 Effective reproduction number $\hat{R}_e$ time profiles

The effective reproduction number $\hat{R}_e(t)$ is the average number of people that someone infected at time $t$ could expect to infect if conditions remain unchanged during an epidemic [48]. It is the ratio of new infections $I_t$, to the total infectiousness of infecteds at time $t$ given by $\sum_{s=1}^{T} I_{t-s}w_s$, where the 'infectivity function' $w_s$ is represented here by the diagnosis-to-diagnosis (DD) distribution with $T = 12$ months. We used the `epiEstim` *R* package was used to estimate $\hat{R}_e(t)$ [47]. Rather than use the generation time distribution for $w_t$ which would have described the infector-to-infectee time interval for a single transmission chain, it is a reasonable corollary for a time series of VL *diagnoses* $I_t$, that the appropriate match be the DD distribution (Eqn. 5.6 & §5.2.3.1.1)—it better portrays the extra time variation caused by convolving the infection process with the time-to-diagnosis process. A 7-month sliding window was used to smooth the estimate that matched the mean DD. As the time of the estimate was still in the diagnosis time domain, it was shifted by a fixed single 7-month lag (mean of IP + OD distribution) to show the results in the infection time domain.

### 5.2.3 Spatiotemporal model

#### 5.2.3.1 Base model

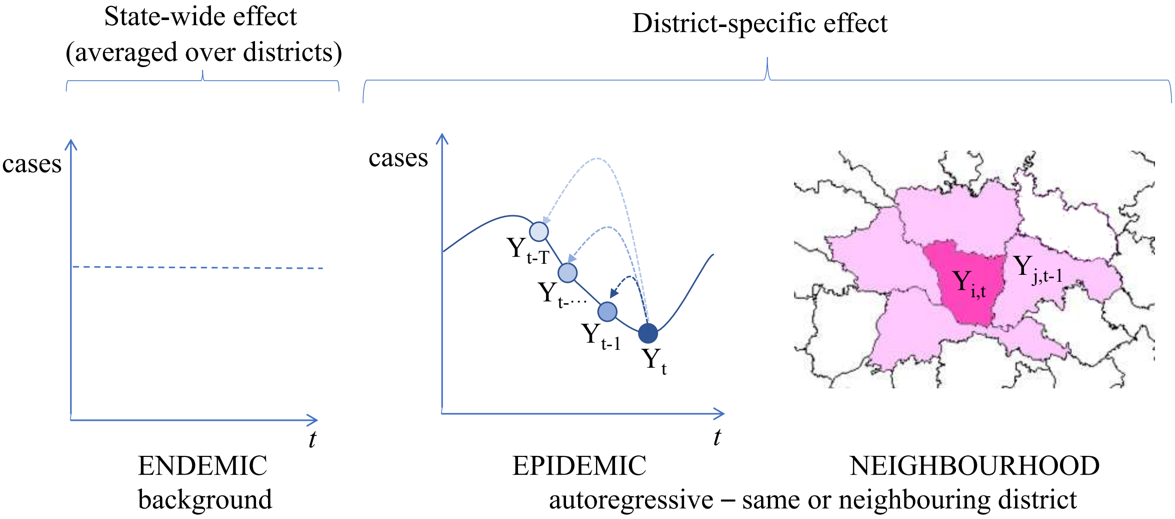


Figure 5.3: **Spatiotemporal model composition**. The respective three components of the Held et al. spatiotemporal model framework [86] that inspired the base model.

The *base model* of monthly district case counts observed in district $i$ in diagnosis month $t$ (Eqns. 5.1–5.5) represents ongoing direct transmission between cases ('epidemic' component, $\lambda_i$) while accounting for the typical VL (diagnosis-to-diagnosis) serial interval, hidden transmission from unobserved or asymptomatic cases ('endemic' component, $\nu_{i,t}$) with high/low-incidence stratification

$\alpha^{(\nu)}_{i,t} \in \left\{\alpha^{(\nu)}_{\text{low incid.}}, \alpha^{(\nu)}_{\text{high incid.}}\right\}$, effects of directly-adjacent districts ('neighbourhood' component, $\phi$), and changing district-specific population effects $e_{i,t}$ (Fig. 5.3).

The process producing observed cases $Y_{i,t}$ is assumed to follow a Negative Binomial distribution with mean $\mu_{i,t}$ & variance $\sigma^2_{i,t}$ conditional on a weighted sum of cases from the previous 12 months $\sum_{T=1}^{12} D_T Y_{i,t-T}$, where $D_T$ is the weight for the cases $T$ months ago (*i.e.* *distributed-lag* autoregression) (Eqns. 5.1–5.2). We were not extend the lag beyond 12 months because of model instability. This distributed-lag distribution represents the DD distribution, *i.e.* the distribution of times between VL *diagnoses* of infector & infectee (§5.2.3.1.1), akin to a 'diagnosis' SI distribution. It better represents the temporal correlation of diagnosis times than a naïve lag-1 autoregression [32]. The normalised DD distribution $D_T$ is informed by an estimated IP (mean = 6 mo) [42, 44], which broadly agrees with literature estimates [28], and an onset-to-diagnosis (OD) distribution (mean = 1·47 mo) from a Bihar study in the third quarter of 2012 [99]. With the DD distribution being a central assumption of all our models, we focussed a sensitivity analysis on the OD distribution to assess impact on RQ1 & RQ2 (§5.2.5). Focussing on the OD distribution made more sense than the IP distribution, both of which contribute to the DD distribution. This is because the OD distribution represents a quantity more liable to change during the epidemic as treatment behaviour & public health resources change, whereas the incubation period was seen as a more implicit property of the disease history, of which we have no information about its heterogeneity during an epidemic.

The cases observed in the previous 12 months cannot fully account for those observed in the current month because of noise in the temporal correlation of a district's cases with itself or its immediate (first-order) neighbours. In using this framework [86], it is assumed that there is a directly-observed process of autoregressive effects from the same district or its neighbours (epidemic & neighbourhood components, respectively) and an indirectly-observed process of background transmission (endemic component) from unobserved symptomatic or asymptomatic individuals. These three components (Fig. 5.3) sum to give the conditional mean $\mu_{i,t}$ (Eqn. 5.2) and the full process observed. The directly-observed process can be inferred from spatiotemporally-*local* information of the case numbers of the last 12 months in the district & its immediate neighbours; whereas the indirectly-observed process is inferred by fitting to the pre-specified time-varying membership of the high/low-incidence stratum across Bihar state, to estimate a district-averaged monthly contribution.

$$Y_{i,t}\big|\{Y_{i,t-12},\ldots,Y_{i,t-1}\} \sim \text{NegBin}(\mu_{i,t},\sigma^2_{i,t}), \tag{5.1}$$

$$\mu_{i,t} = e_{i,t}\nu_{i,t} + \lambda_i \sum_{T=1}^{12} D_T Y_{i,t-T} + \phi_i \sum_{j\neq i}\left(\omega_{ji}\sum_{T=1}^{12} D_T Y_{i,t-T}\right) \tag{5.2}$$

$$\text{where: } \ln(\nu_{i,t}) = \alpha^{(\nu)}_{i,t} \quad \text{(endemic component)}, \tag{5.3}$$

$$\ln(\lambda_i) = \alpha^{(\lambda)}_{\text{other}} + \mathbb{1}_{\{i=\text{Vaishali}\}}\alpha^{(\lambda)}_{\text{Vaishali}} \quad \text{(epidemic component)}, \tag{5.4}$$

$$\ln(\phi_i) = \alpha^{(\phi)}_i \quad \text{(neighbourhood component)}, \tag{5.5}$$

$$D_T = \text{DD interval distribution weightings (normalised)}.$$

with population offset $e_{i,t}$; $\omega_{ji} = 1$ if $j$ neighbours $i$, else 0; and two overdispersion terms $\psi_{\text{high}}$, $\psi_{\text{low}} > 0$, s.t. $\sigma^2_{i,t} = \mu_{i,t}(1+\psi_k\mu_{i,t})$ for $k \in \{\text{high}, \text{low}\}$ endemicity districts.

**Unpacking the base model's** three components (Eqns. 5.1–5.5):

**'endemic' ($e_{i,t}\nu_{i,t}$):** cases in the same district caused by a time-specific background unobserved transmission term $\alpha^{(\nu)}_{i,t} \in \left\{\alpha^{(\nu)}_{\text{low incid.}}, \alpha^{(\nu)}_{\text{high incid.}}\right\}$ that takes a lower or higher value according to whether district $i$ in month $t$ has fewer than 11 cases, or 11 or more, respectively. A population offset

$e_{i,t}$ accounts for the higher case numbers expected in districts with larger populations.

**'epidemic'** $\left(\lambda_i \sum_{T=1}^{12} D_T Y_{i,t-T}\right)$**:** cases correlated with a weighted sum of the last 12 months' cases in the same district $i$. This component was represented by two fixed intercepts: one for other districts $\alpha^{(\lambda)}_{\text{other}}$ & a combined one for Vaishali, $\left(\alpha^{(\lambda)}_{\text{other}} + \alpha^{(\lambda)}_{\text{Vaishali}}\right)$. The offset term was not included here since the study's cases occur in a minority subgroup of the district's population, so are assumed to be dependent on the cases arising in the epidemic process rather than the wider population.

**'neighbourhood'** $\left(\phi_i \sum_{j \neq i} \left(\omega_{ji} \sum_{T=1}^{12} D_T Y_{i,t-T}\right)\right)$**:** recent case importation from adjacent districts $j$ ($\omega_{ji} = 1$ if $j$ neighbours $i$, else 0).

**5.2.3.1.1 Obtaining the DD distribution** In detail, the sum of normalised weights of 12-month lagged data accounts for the DD distribution for VL, $D_T$, which was estimated as {0·08 [1st lag at $t-1$], 0·11, 0·12, 0·13, 0·12, 0·10, 0·09, 0·08, 0·06, 0·05, 0·04, 0·03 [12th lag at $t-12$]}. This is simulated from:

i) previous IP model estimate from a Bangladesh study using a zero-truncated Negative Binomial distribution with shape parameter = 3 & probability of success = 0·35 [42, 44] and maximum time of 24 months, which produces a sample distribution with a mean of 6 months & standard deviation of 4 months.

ii) OD time estimate [99] from Bihar in Q3 2012 using a Lognormal $(\mu, \sigma)$ distribution with log mean $\mu = 3{\cdot}5$ & log standard deviation $\sigma = 0{\cdot}8$ and maximum time 492 days, produces a sample distribution with mean 1·5 months & standard deviation 1·4 months. We assume that the period of infectiousness starts at onset and finishes at treatment—additionally we assume that the time of diagnosis & treatment occurred at the same time as field studies suggest they are only separated by 1–2 days [99].

From these simulated distributions, the DD distribution, $D_T$, can be generated (Eqn. 5.6 & Fig. 5.4). Note that the OD distribution was randomly drawn twice for the independent time intervals of case 2 (infectee) & case 1 (infector). By truncating the DD below one month and above 12 months, only ∼10·8% of DD intervals are missed from this range. See Table 5.3 for the results of a sensitivity analysis on the main results for research questions 1 & 2, based on changes in the OD distribution.

$$D_T = \text{DD interval}_{1\rightarrow 2} \sim \text{IP}_{\text{case 2}} + \text{OD}_{\text{case 2}} - {}^1\!/_2\text{OD}_{\text{case 1}}. \tag{5.6}$$

**5.2.3.1.2 Overdispersion cut-off choice** The Negative Binomial distribution produces non-negative predictions and accounts for overdispersion arising from increased variability due to unobserved covariates or time-aggregated incidence [85]. As the endemicity distribution of districts (histogram of districts' total cases during the study) resembles a two-component mixture model (Fig. 5.5), overdispersion was dichotomised into $\psi_{\text{high end.}}, \psi_{\text{low end.}}$, for high & low endemicity districts, respectively, at a cut-off of 1,000 total cases during the study. This parsimonious approach was preferred rather than introducing a separate overdispersion parameter ($\psi_i$) for every district.

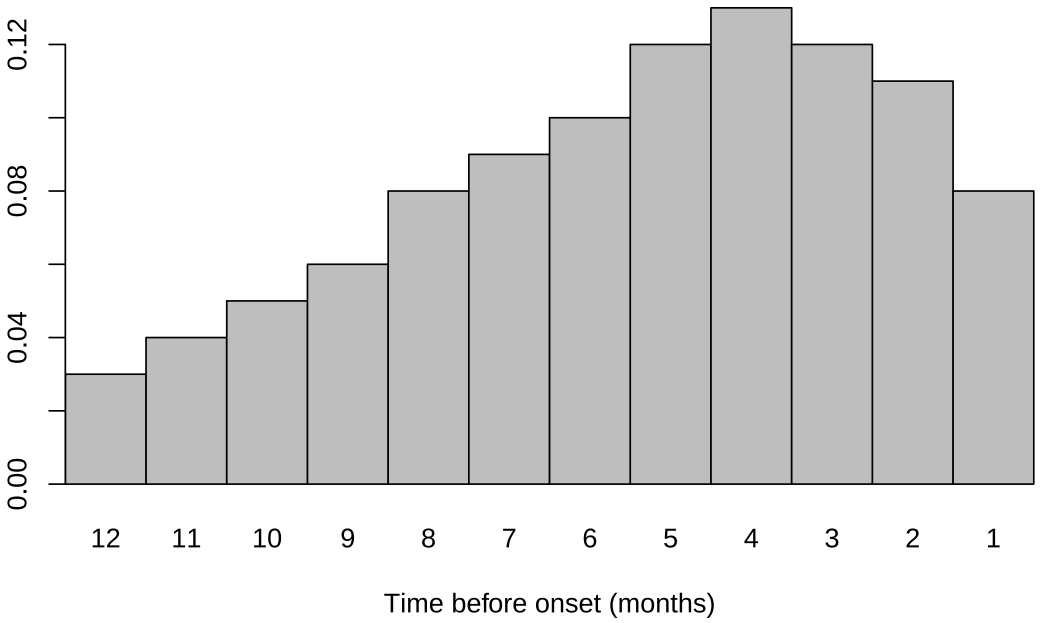


Figure 5.4: **Normalised weights of the 12-month diagnosis-to-diagnosis distribution**.

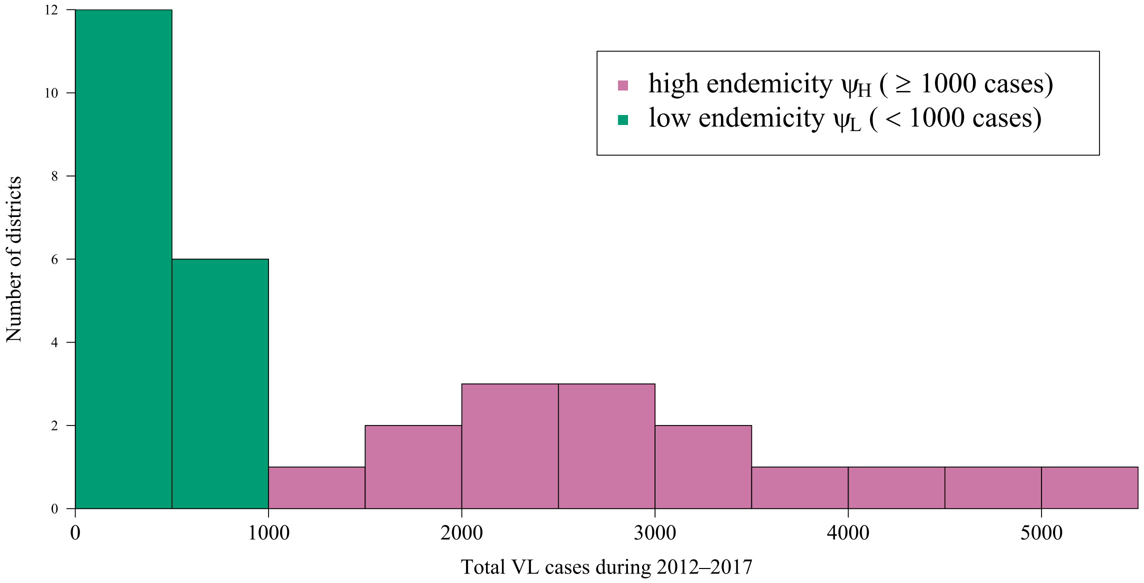


Figure 5.5: **Overdispersion classification**. District frequency by total VL cases to decide cut-off for high/low overdispersion assignment.

#### 5.2.3.2 Interrupted time series analysis (ITSA): the pilot model

An *ITSA* is a subset of regression discontinuity analysis that was applied to the districts' longitudinal case counts to assess the impact of this non-randomised pilot while adjusting for existing trends [13, 19, 105]. The *assignment variable* in this ITSA is the calendar time $\tau$[2] (Eqn. 5.10) of the start of the pilot implementation in Vaishali. The base model (§5.2.3.1) was expanded with the pilot effect $\left(a\ priori \text{ primary variable, } \alpha^{(\lambda)}_{\text{pilot}}\right)$ and annual seasonality in the epidemic & endemic components (Eqns. 5.10 & 5.9) to form the pilot model (§5.2.3.2.1):

$$Y_{i,t}\big|\{Y_{i,t-12},\ldots,Y_{i,t-1}\} \sim \text{NegBin}(\mu_{i,t},\sigma^2_{i,t}),\ \text{unchanged from Eqn. 5.1} \tag{5.7}$$

$$\mu_{i,t} = e_{i,t}\nu_{i,t} + \lambda_{i,t}\sum_{T=1}^{12} D_T Y_{i,t-T} + \phi_i \sum_{j\neq i}\left(\omega_{ji}\sum_{T=1}^{12} D_T Y_{i,t-T}\right) \tag{5.8}$$

where the epidemic component can now be time-changing *i.e.* $\lambda_{i,t}$

The endemic, epidemic & neighbourhood components were formulated as:

$$\ln(\nu_{i,t}) = \alpha^{(\nu)}_{i,t} + A_{\text{END}}\sin\left(\frac{2\pi}{12}t + \Phi_{\text{END}}\right) \qquad \text{(endemic)}, \tag{5.9}$$

$$\ln(\lambda_{i,t}) = \alpha^{(\lambda)}_{\text{other}} + \mathbb{1}_{\{i=\text{Vaishali}\}}\left(\alpha^{(\lambda)}_{\text{Vaishali}} + \mathbb{1}_{\{t\geq\tau\}}c_t\cdot\alpha^{(\lambda)}_{\text{pilot}}\right) + A_{\text{AR}}\sin\left(\frac{2\pi}{12}t + \Phi_{\text{AR}}\right) \qquad \text{(epidemic)}, \tag{5.10}$$

$$\ln(\phi_i) = \alpha^{(\phi)}_i,\ \text{unchanged from Eqn. 5.5} \qquad \text{(neighbourhood)}; \tag{5.11}$$

with $\tau$ starting pilot month, $A_{(\text{END/AR})}$ annual sinusoid amplitude & phase $\Phi_{(\text{END/AR})}$,

$$\text{and } c_t = \begin{cases} 0 & \text{if} \quad t < \tau, \\ \frac{\sum_{p=1}^{p=t-\tau+1} Y_{i,t-p}}{\sum_{p=1}^{12} Y_{i,t-p}} & \quad\ \ \tau \leq t < \tau + 11, \\ 1 & \quad\ \ \tau + 12 \leq t. \end{cases} \tag{5.12}$$

which corrects for the first 12 months of the pilot due to delayed-lag intervention effects (§5.2.3.2.1).

District differences are encapsulated as control effectiveness, which includes the pilot effect in Vaishali in Eqn. 5.10 (epidemic component), two high/low endemicity levels in Eqn. 5.9 (endemic component), and strength of transportation links & human flow between districts in Eqn. 5.11 (neighbourhood component).

**5.2.3.2.1 Model building—how the pilot model was obtained** Each fit used all the districts' data for the whole time series (unless indicated) to ensure a good fit to Vaishali's neighbours. Sometimes Vaishali-specific information was extracted from the output for further calculation or graphs. Parameters were inferred using an iterative scheme built into the `surveillance` & `hhh4addon` packages[3] that minimised the model's likelihood using the `nlminb` optimisation method in

[2] not to be confused with the tau statistic with the same symbol in earlier chapters

[3] `hhh4addon` is a branched development, additional to `surveillance`

the hhh4_lag() function [149]. The fitted values of the pilot model's case counts were plotted against their residuals to assess heteroskedasticity. Starting from the base model (Eqns. 5.1–5.5), the model fit was sequential to reach the final pilot model (Eqns. 5.7–5.12).

Model development can be summarised as follows:

**Pilot model = Base model + ...**

**pilot effect:** the impact was modelled via a single step change in the intercept in the epidemic component for Vaishali only, to capture changes in the time series of cases when the pilot began (Eqn. 5.10: epidemic component). This answered RQ1 (§5.1.4)).

**pilot start month:** due to the uncertainty of when each of the three control elements may have started to impact diagnosed cases, they were treated as having a single combined effect, assumed in January 2015 (Fig. 5.2b). Thirteen possible pilot start months $\tau$ (September 2014–September 2015 inclusive) were tested, and the best model with $\tau^*$ chosen that minimised the AIC. To assess the sensitivity of the pilot model's parameters to the start time, their range was also reported when varying the start month (§5.3.4). September 2014 was the lower bound as ASHA training had started by then, while September 2015 was the upper bound: if IRS in the first half of 2015 had failed due to DDT resistance, the second round might still have become effective after insecticide change (Fig. 5.2b).

**seasonality:** annual sinusoid in epidemic and endemic components combinations was trialled.

Model selection was based on the lowest Akaike information criterion (AIC) of each candidate model versus the best-performing model from the previous selection step while monitoring changes in parameter uncertainty, particularly for the primary variable $\alpha^{(\lambda)}_{\text{pilot}}$ [149].

The selection of the predictive model for RQ2 was based on *predictive performance*. Firstly the predictive performance of the pilot model was compared against the base model to establish that the pilot model was a relative advancement. PIT histograms were also used to check if the pilot model had a constant predictive performance throughout the predicted range [52, 130].

The predictive performance of simulations of the (pilot) null model & alternative models were then compared with alternative parametric functional forms (linear, Geometric & Poisson) for the DD distribution by summing the '*one-step-ahead*' sequential model scores [149]. However, the DD distribution was shown to have the lowest ranked probability score (RPS) value. The 'one-step-ahead' approach fits up to and including month $t$ and predicts the cases in the next month $t + 1$; it uses the difference between observed & predicted cases to form a model score; the model is then refitted with the extra observed data at $t + 1$ and the next month's cases predicted, and so on. This approach thus sequentially trials the two models repeatedly, takes the mean of these scores, then compares the pair's scores. The RPS value was used for comparison since it gives less weight to extreme departures from the trend of the observed [1]. As the histograms of the distribution of score differences were non-Normal, the non-parametric Permutation Test (using 10,000 simulations) was used to assess the significance of the difference. A $p$-value less than 0·05 was considered to indicate the alternative model had a reasonably better predictive performance. It turned out that the model attained through AIC selection (pilot model) was also optimal for predictive performance. Therefore the counterfactual model stayed the same as the pilot model, apart from omitting the pilot effect.

### 5.2.4 Counterfactual model

A *counterfactual model*, formed by omitting from Eqn. 5.10 the $\left(\mathbb{1}_{\{t\geq\tau\}}c_t \cdot \alpha_{\text{pilot}}^{(\lambda)}\right)$ term, was used to predict the number of cases that would have occurred had there been no intensified control in Vaishali. Informed by the most likely start month $\tau^*$, the cases averted in Vaishali during January 2015–December 2017 was estimated by summing the monthly differences between simulated case counts from this counterfactual (fitted on January 2013–December 2014), versus the pilot model (§5.2.4.1). The cases averted was also presented as a percentage of those that would have occurred under the counterfactual model (Fig. 5.1:D).

When developing a single model to both infer the pilot effect (RQ1) & generate counterfactual predictions (RQ2), it is unclear how to weight AIC & predictive performance (§5.3.6) for these respective purposes [139]. So pragmatically, the fit was optimised for first; and prediction second, working from the pilot model. For model validation, the base, pilot & counterfactual models' goodness of fit were compared by AIC for January 2013–December 2017 (§5.3.6). The predictive performance of the base & pilot models for January 2015–November 2017 was also compared (§5.2.3.2.1).

#### 5.2.4.1 Estimating cases averted

To answer RQ2 (§5.1.4), we took the difference in monthly 'one-step-ahead' January 2015–December 2017 forecasts for pairs of i) simulated cases from the counterfactual model fitted to all data from January 2013–December 2014, initialised with the Mersenne-Twister random number generator seed; & ii) simulated cases from the pilot model. The paired difference (ii−i) was summed for all forecast months to obtain the estimated case numbers averted. 100,000 pairs were simulated to produce an estimated median & IQR for the cases averted.

### 5.2.5 Sensitivity analysis on the key results

Only the mean parameter of the OD distribution ($\mu$) was varied by ±5% & ±10% while leaving the standard deviation ($\sigma$) constant so that the the coefficient of variation too stayed constant (it is only $\sigma$-dependent for the Lognormal distribution). The impact of this sensitivity analysis on the key outputs that informed the first & second research questions (5.1.4) are in Table 5.3.

## 5.3 Results

### 5.3.1 Trends in diagnoses

In most districts, including Vaishali, case counts fell from 2012–2017 (Fig. 5.6). During the pilot years 2015–2017, monthly cases in Vaishali declined substantially in absolute terms compared with cases in its highly-endemic neighbours and the district mean of the rest of the state (Figs. 5.2a & 5.6).

Before the pilot and then during the pilot, Vaishali had the 13th and then 12th largest year-on-year percentage reduction in monthly VL incidence, respectively, out of 33 reporting districts (averaged over 36 monthly incidence ranks from 2015–2017 and over 24 months from 2013–2014, respectively). However, it also had the 11th highest VL incidence during the pilot period (averaged over 36 monthly incidence ranks from 2015–2017) versus pre-pilot when it was the 4th highest (averaged over 36 months from 2012–2014, $p < 1 \times 10^{-10}$ Wilcoxon test) (Fig. 5.7).

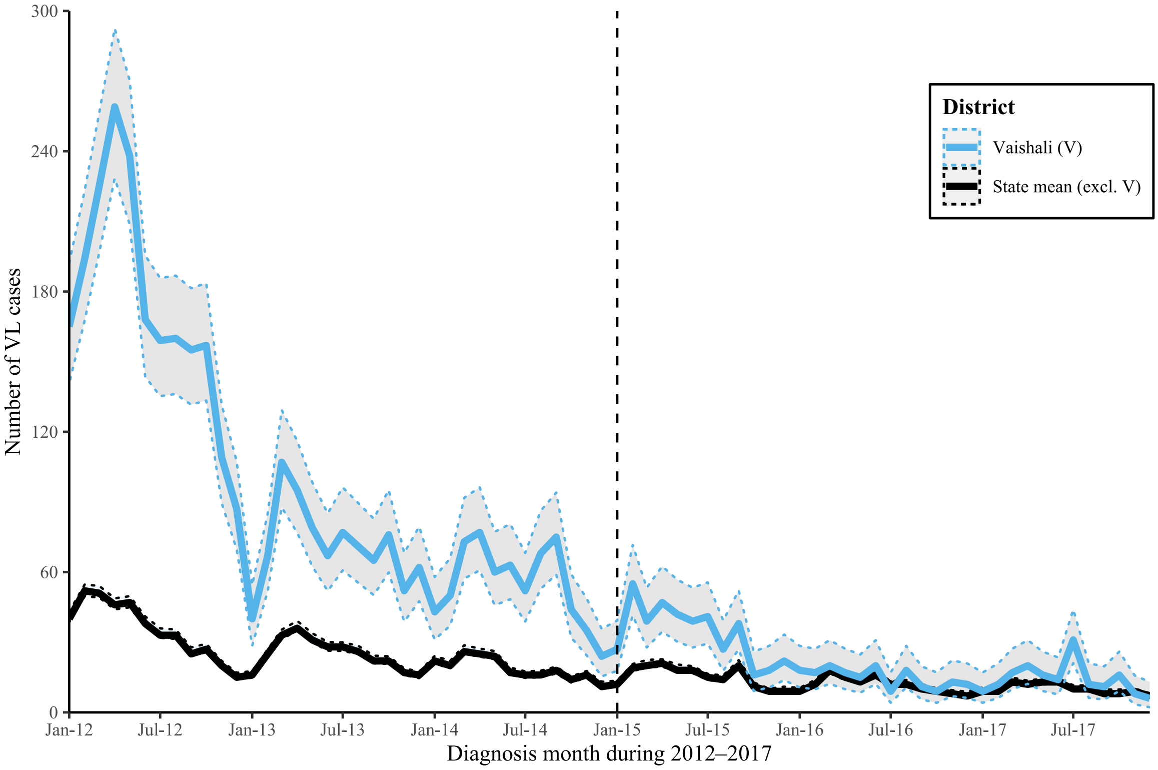


Figure 5.6: **VL time series** for Vaishali district & the rest of Bihar state. Monthly case counts [82]. Note that HIV-VL cases are included from 2015–2016 and HIV/TB-VL from 2017; monthly HIV/TB-VL case proportions out of all VL cases are shown in Fig. D.1. The state mean excludes Aurangabad, Gaya, Jamui, Kaimur, Rohtas & Vaishali districts. The dashed vertical line indicates the start of the modelled intervention.

### 5.3.2 Seasonality & spatial correlation

Across Bihar, an annual seasonality in case counts was apparent, whose signal weakened as endemicity fell (Fig. 5.6). However, at the district level the strength of the seasonal signal varied and was only recognisable for some high-endemicity districts (*e.g.* Saran had a solid seasonal signal while others did not—note not presented in figures).

Spatial correlation in incidence between neighbouring districts was apparent both before & during the pilot; Global Moran's $I = 0{\cdot}36$, $p = 0{\cdot}002$ (10,000 simulations) & $I = 0{\cdot}40$, $p = 8 \times 10^{-4}$, respectively. This supports the use of the between-district neighbourhood component in Eqn. 5.5. Vaishali was surrounded by neighbours with a range of endemicities, which either remained constant (e.*g.* Saran) or declined. Although incidence was declining in Vaishali, it was also declining in many other districts, yet clustering remained among the other districts, while Vaishali was dissimilar to its neighbours.

### 5.3.3 Effective reproduction number

The estimated district-specific effective reproduction numbers $\hat{R}_e(t)$ generally follow an annual seasonality (Fig. 5.8b) which supports using seasonality in the model (Eqns. 5.9 & 5.10). Compared to the average trend of the other 32 districts, Vaishali saw sustained $\hat{R}_e < 1$ during 2012/3, with the next noticeable sustained reduction around the pilot start (Fig. 5.8a): after summer 2015, Vaishali's

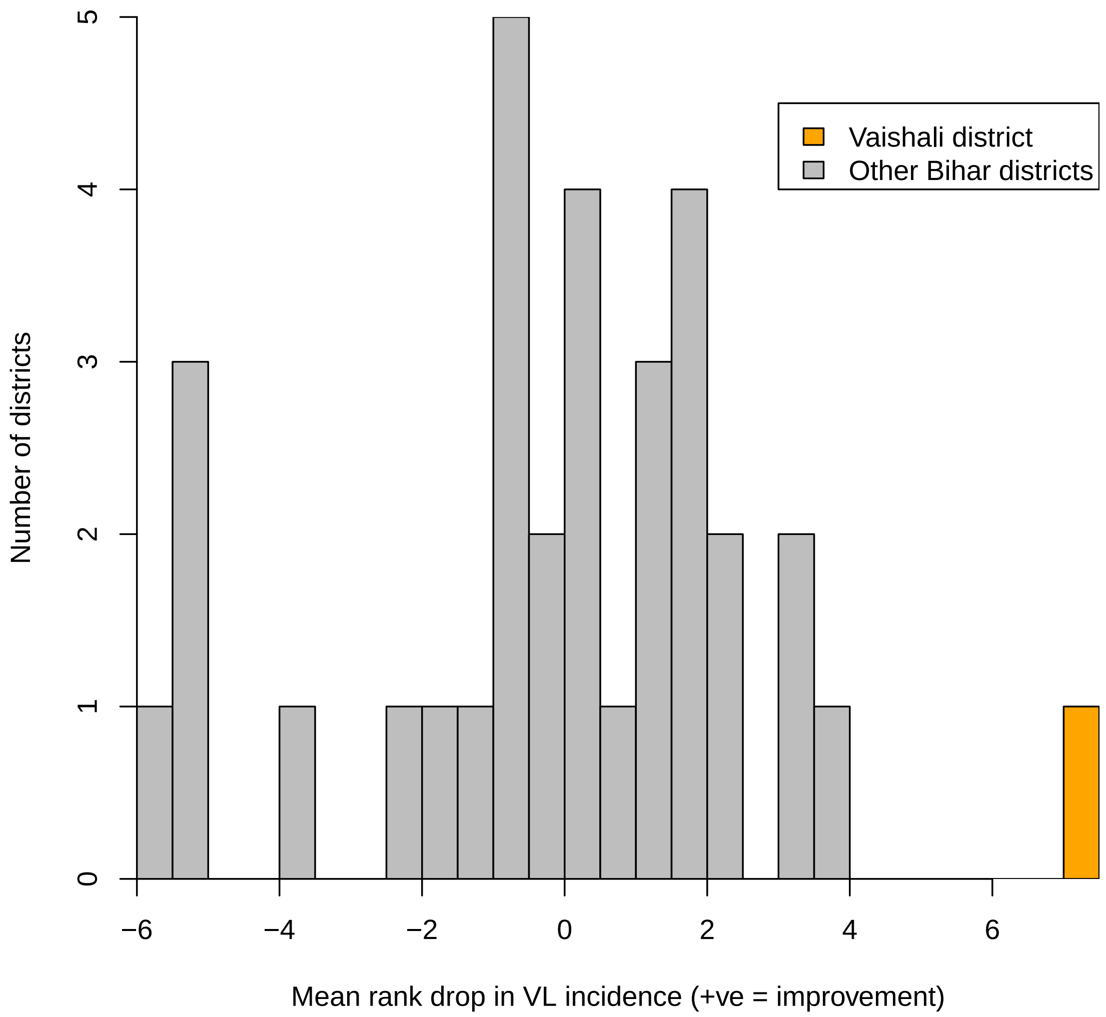


Figure 5.7: **Distribution of Bihar districts' rank change in incidence** from before the pilot in Vaishali district (averaged over 36 monthly incidence ranks during 2012–2014) to during the pilot (averaged over 36 months during 2015–2017). Vaishali district clearly shows a large improvement versus the distribution of the remaining 32 Bihar districts.

$\hat{R}_e$ did not return to a seasonal peak around January 2016 unlike the mean of the other 32 districts. However, this effect only lasted the 2015/6 season and Vaishali's $\hat{R}_e$ resurged at the end of 2016. Given 2017's lower incidence, the impact of this above-one $\hat{R}_e$ in terms of new cases would have been less than if it had occurred at 2015's case levels.

### 5.3.4 Pilot model estimation

The pilot model selected consisted of a Negative Binomial distribution with population offset, annual sinusoid in epidemic & endemic components to account for seasonality, time-specific endemic intercept for high/low incidence, a distributed-lag epidemic component with a single change-of-intercept in

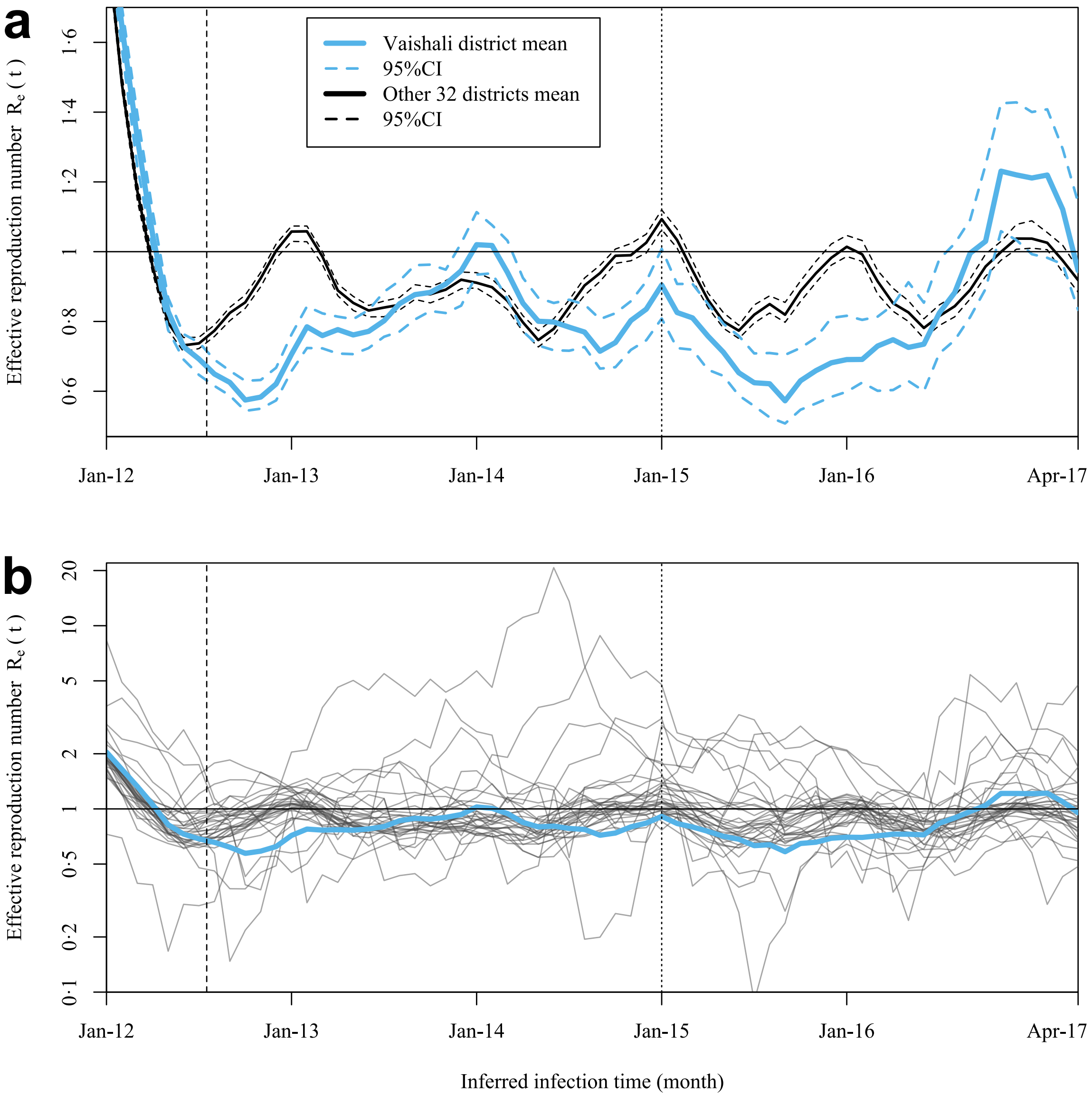


Figure 5.8: **Effective reproduction number** $\hat{R}_e(t)$ for Vaishali district & the other 32 districts as means (a) and as 33 separate districts (b). In (a), the mean $R_e(t)$ of the 'other 32 districts' was computed by applying `epiEstim` to the simple mean of the 32 time series. The x-axes' inferred infection times were calculated by subtracting the mean IP & mean onset-to-diagnosis time (7-month total shift) from the diagnosis month. The dashed vertical line indicates the length of the 7-month sliding window used to smooth the $\hat{R}_e$ estimates: $\hat{R}_e$ estimates before this date are unreliable as they only include partial data within this sliding interval. The dotted vertical line indicates the start of the modelled intervention.

Vaishali in January 2015, a constant distributed-lag contribution from directly-adjacent districts in the neighbourhood component, fixed intercept means in the epidemic component (one for Vaishali & one for the other 32 districts), and overdispersion by high/low-endemicity districts (Eqns. 5.7–5.12). The relative contributions of the time-varying components were evaluated through a plot of the fitted components alongside the observed cases (Fig. 5.9). The pilot model fitted better than the base model ($\Delta$AIC $= -308{\cdot}3$) and showed a reasonable fit to the observed case counts for Vaishali (Fig. 5.9) but,

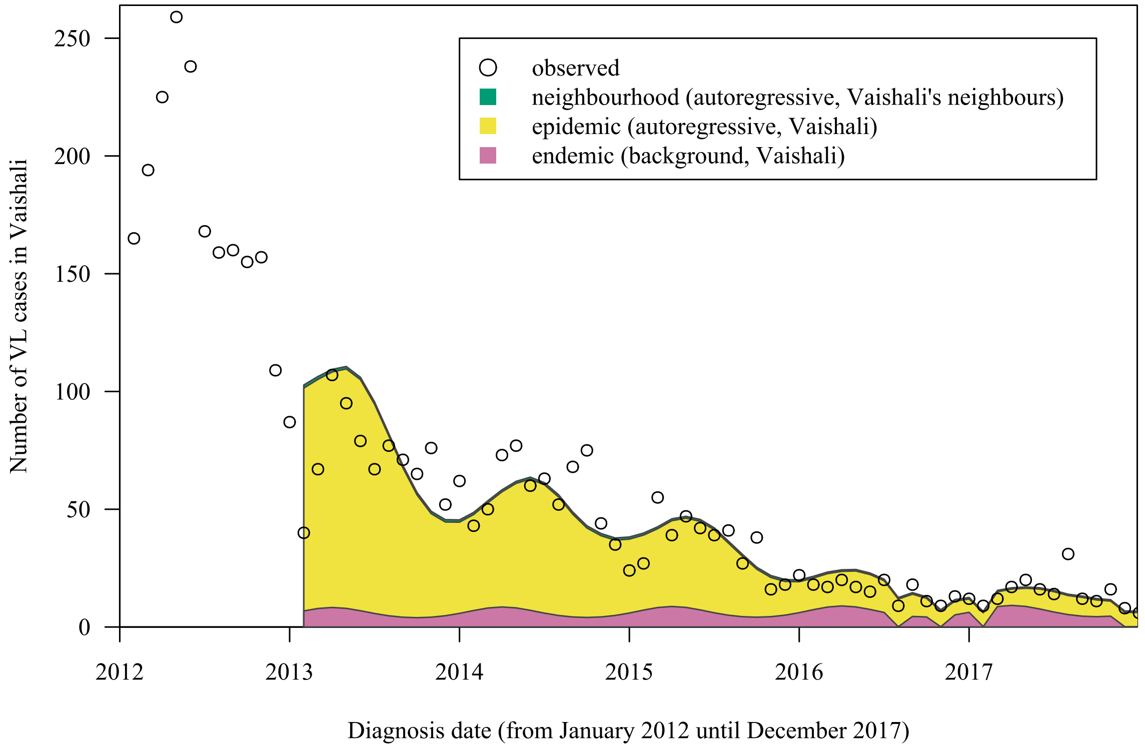


Figure 5.9: **Model decomposition**. The epidemic component of the model in Vaishali (yellow area) is sandwiched between the neighbourhood (green slither) & endemic (pink area) components. After 12 months of lagged data, the model component fit starts from January 2013. January 2015 was the chosen pilot start month $\tau^*$ for the pilot model, however, due to the distributed-lag intervention structure, the reduction in the epidemic component for the first 12 months is gradual using corrections via $c_t$ in Eqns. 5.12. This figure was produced using `surveillance` package in *R* [130, 168].

across all districts, was prone to overestimating low counts (Fig. 5.12). January 2015 was chosen as the pilot start month $\tau^*$ as it had the lowest AIC.

Table 5.2 shows the parameter estimates for the intervention effect, which can be interpreted as follows. For Vaishali pre-pilot, an estimated average of 67·8% of the weighted sum of the previous 12 months' case counts contributed towards the current month's case count, versus 70·8% (95%CI 67·7–73·8%) for other districts. This means a hypothetical same-sized epidemic in any of the other districts would take slightly longer to die out on average than if it was to occur in Vaishali. During the pilot, this 67·8% contribution was estimated to fall (by 27·3%) to 49·3% for January 2015 onwards, where the pilot effect CI represents a significant drop (*i.e.* 8·8–45·8%). The estimated endemic contribution per district since January 2012 (based on a mean district population of 2·8 million) was practically nil ($\sim 0$ cases/mo) for low-incidence settings ($< 11$ observed cases/mo), whereas high-incidence settings ($\geq 11$ observed cases/mo) are estimated to get 3/6 cases/mo for low/high seasons, respectively. In absolute terms the seasonality term contributed more to the epidemic component than endemic (Fig. 5.9)—0·77× at the November minimum & 1·31× at the May maximum of the epidemic component. Each district received an estimated 0·5% average contribution from each of the adjacent districts' weighted sum of their previous 12 months' cases. The standard errors of all parameters were within reasonable bounds. Parameters were mostly insensitive to pilot start month $\tau$; however, the pilot effect on the epidemic component $\alpha^{(\lambda)}_{\text{pilot}}$ & Vaishali-specific intercept $\alpha^{(\lambda)}_{\text{Vaishali}}$ did differ by up to 13% & 7%,

| **Parameter** | **Adjusted estimate** | **SE** |
|---|---|---|
| **Epidemic component** | | |
| Pilot effect, $\exp\left(\alpha^{(\lambda)}_{\text{pilot}}\right)$ (change-of-intercept) | 0·727 | 0·094 |
| Fixed intercept mean: Vaishali, $\exp\left(\alpha^{(\lambda)}_{\text{other}} + \alpha^{(\lambda)}_{\text{Vaishali}}\right)$ | 0·678 | —$^{\dagger}$ |
| Fixed intercept mean: Other 32 districts, $\exp\left(\alpha^{(\lambda)}_{\text{other}}\right)$ | 0·708 | 0·016 |
| Seasonality: | | |
| Amplitude, $A_{\text{AR}}$ | 0·269 | 0·022 |
| Phase, $\Phi_{\text{AR}}$ | -0·678 | 0·046 |
| | | |
| **Endemic component** | | |
| Intercept mean, $\exp\left(\alpha^{(\nu)}\right)$: | | |
| high-incidence, $\exp\left(\alpha^{(\nu)}_{\text{high incid.}}\right)$ | $1{\cdot}55 \times 10^{-6}$ | —$^{\dagger}$ |
| low-incidence, $\exp\left(\alpha^{(\nu)}_{\text{low incid.}}\right)$ | $3{\cdot}72 \times 10^{-8}$ | $1{\cdot}48 \times 10^{-8}$ |
| Seasonality: | | |
| Amplitude, $A_{\text{END}}$ | 0·377 | 0·132 |
| Phase, $\Phi_{\text{END}}$ | 0·517 | 0·171 |
| | | |
| **Neighbourhood component** | | |
| Fixed intercept mean, $\exp\left(\alpha^{(\phi)}\right)$ | $4{\cdot}75 \times 10^{-3}$ | $1{\cdot}37 \times 10^{-3}$ |
| | | |
| **Overdispersion** | | |
| High-endemic district, $\psi_{\text{high end.}}$ | 0·060 | 0·005 |
| Low-endemic district, $\psi_{\text{low end.}}$ | 0·115 | 0·018 |

Table 5.2: **Pilot model parameter estimates**. Those referenced in §5.3.4 are highlighted. Mathematical notation explained in Eqns. S1 & S2. Some parameters ($^{\dagger}$) are combined from individual ones for interpretability, but their standard errors are not provided, as the `surveillance` package only provides single parameter estimates.

respectively, versus their value for a January 2015 start. The changes in the mean parameter of OD distribution had negligible effect on both RQ1 & RQ2 (Table 5.3).

The fit of the pilot model was superior (AIC = 10281·7) to the base model (AIC = 10590·0) and similar to the counterfactual model (AIC = 10286·3). The final & counterfactual models were also better in RPS (RPS = 2·50 & 2·46, respectively) than the base model (RPS = 2·81) in prediction for 2015–2017 ($p = 1 \times 10^{-5}$, permutation test).

### 5.3.5 Estimating cases averted

The counterfactual model showed reasonable predictive performance (§5.2.3.2.1 & 5.3.6) before the pilot, notwithstanding the last four months of 2014, where the limited duration of this test period caused convergence issues (Fig. 5.13). The pilot model (Fig. 5.14b) produced forward predictions for

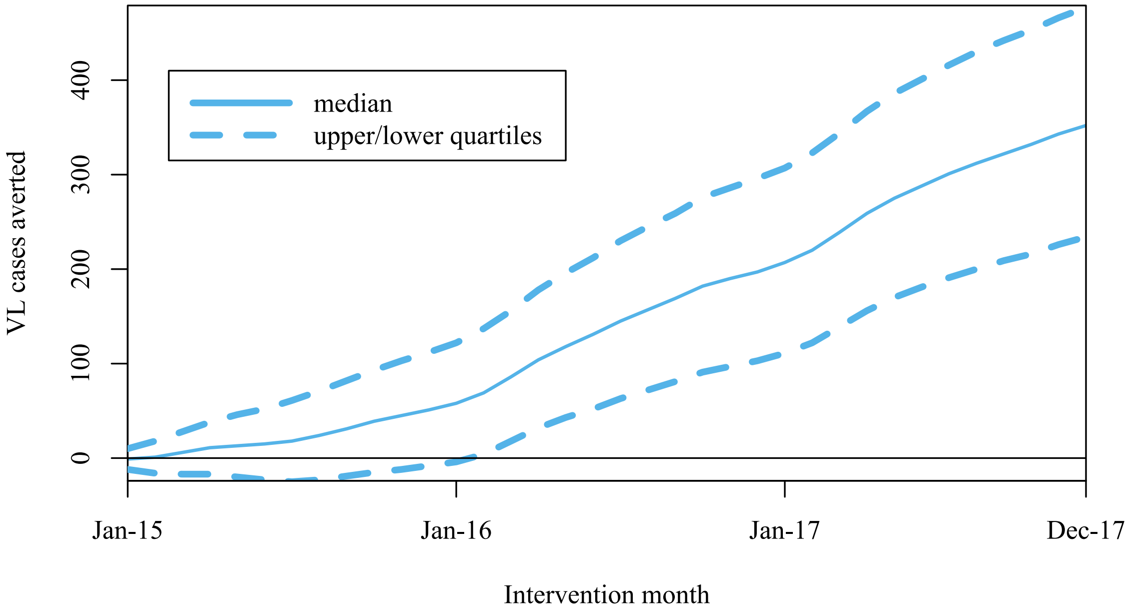


Figure 5.10: **Estimated cumulative cases averted since the pilot started**.

2016–2017 generally sharper & closer to the observed time series than those of the counterfactual model (Fig. 5.14a). Predictions of both models into 2017 were less robust to large departures from the mean trend, *e.g.* the unexpected July 2017 peak in observed cases (Fig. 5.14b), as the epidemic component was diminished by this point. The predictive performance of the pilot model was poorer at extrema, especially high counts (Fig. 5.11).

Simulations comparing pilot & counterfactual models suggest a median of 352 (IQR 234–477) cases were averted in Vaishali during the pilot from January 2015 (2% of 100,000 simulations had estimated negative cases averted), which would have accounted for an estimated 31% of cases if there had been no intensified control (Fig. 5.10).

Analysing the year-on-year incidence decreases that could have occurred under the counterfactual model, the pilot was estimated to have averted additional cases, as a median percentage of the total cases estimated under the counterfactual model, of 93·9% (IQR 37·5–203·3%) from 2015–2016 & 29·0% (IQR -42·9–137·5%) from 2016–2017 (Fig. 5.10).

### 5.3.6 Model validation

Heteroskedasticity was present in the pilot model since a higher variance of model residuals was present when the model was fitted to low numbers of cases (Fig. 5.12).

The final counterfactual model made reasonable 'one-step-ahead' sequential forecasts of the monthly case numbers in Vaishali district in 2014 based on a fit to the 2013 data, as assessed visually (Fig. 5.13), suggesting that the model captured the essential features of the process giving rise to the case counts and could be relied upon to make counterfactual predictions from 2015, based on the 2013–2014 status quo.

Time series plots were used to visually assess the difference between the pilot & counterfactual

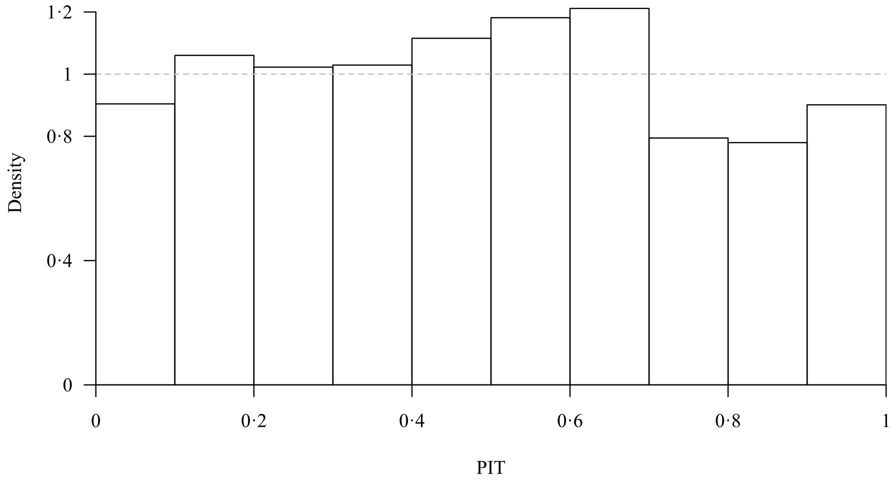


Figure 5.11: **PIT histogram** for the pilot model. The upside-down ‘U’ shape indicates overdispersion; therefore model predictions at extreme counts will be less reliable.

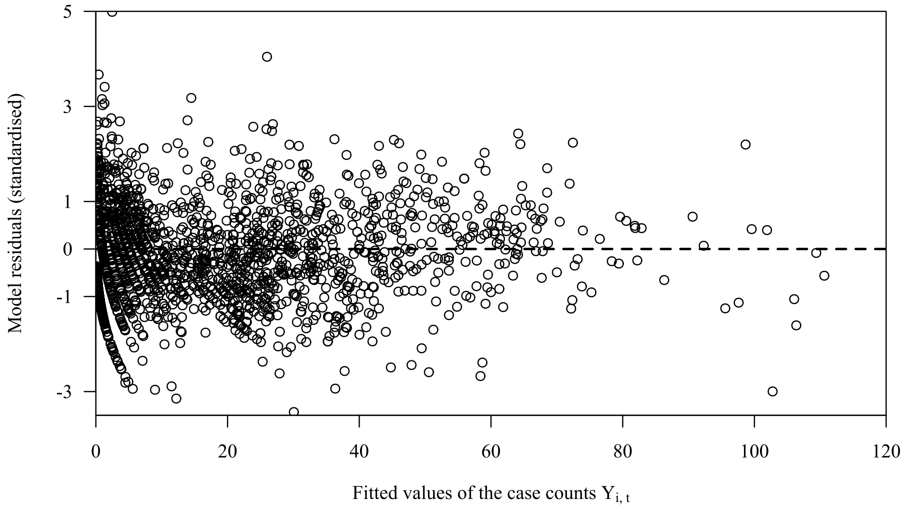


Figure 5.12: **Pilot model heteroskedasticity** when fitted to all 33 districts during January 2013–December 2017.

models versus the observed case time series with fanplots of the distribution of simulations from sequential ‘one-step-ahead’ forecasts of January 2016–December 2017 [1] (Fig. 5.14).

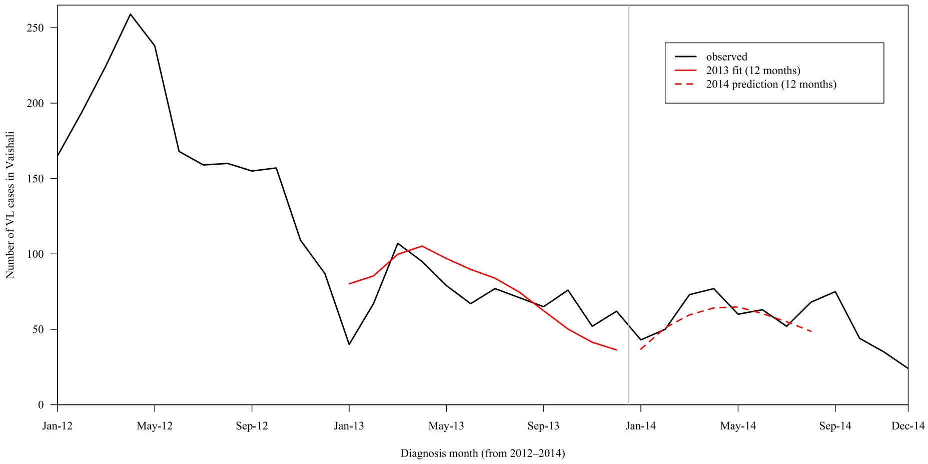


Figure 5.13: **Counterfactual pre-intervention goodness of fit & predictive performance**. The counterfactual model was fitted to the observed data (black line) during 2013 (on the left-hand side of the vertical grey line) to produce the initial fit (solid red line) for that time series. 'One-step-ahead' forecasts were then sequentially made for progressing months (dashed red line) according to the complete observed time series for the 12 previous months (§5.2.3.2.1) [149]. Model convergence problems meant that predictions could not be made for the latter 4 months of 2014—this arose from the limited test dataset and did not appear when using the entire dataset.

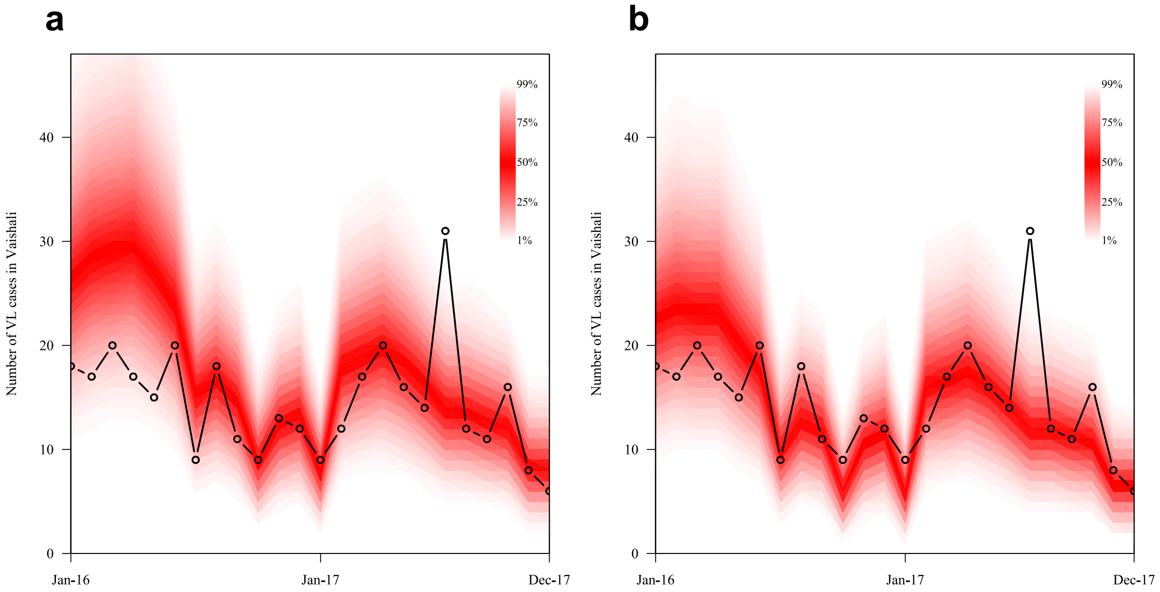


Figure 5.14: **Fanplots**. Traditional forecast plots show a distribution of forecast time series emanating (and diverging with time from the mean forecast) from a single observed time. The fanplot, however is a companion to the 'one-step-ahead' model scoring and only plots the sequential one months' worth of forecast counts ahead. Sequential probability distributions for **a)** counterfactual model & **b)** pilot model. The connected black line represents observed cases, and the red gradient band indicates sample quantiles about each month's predicted values. This figure was produced using the `fanplot` package in $R$[1].

| × **OD log mean** ($\mu$) | **mean OD (mo)** | **mean DD (mo)** | **Vaishali pre-pilot** $\exp\left(\alpha^{(\lambda)}_{\text{other}} + \alpha^{(\lambda)}_{\text{Vaishali}}\right)$ | **% reduction** $1 - \exp\left(\alpha^{(\lambda)}_{\text{pilot}}\right)$ | **Pilot effect** $\exp\left(\alpha^{(\lambda)}_{\text{pilot}}\right)$ | **Cases averted median & IQR** |
|---|---|---|---|---|---|---|
| 0·9 | 1·04 | 6·3 | 0·684 | 27·2 | (8·8, 45·6) | 351 [231, 479] |
| 0·95 | 1·24 | 6·4 | 0·681 | 27·2 | (8·8, 45·7) | 351 [232, 478] |
| 1 | 1·47 | 6·5 | 0·678 | 27·3 | (8·8, 45·8) | 352 [234, 477] |
| 1·05 | 1·75 | 6·7 | 0·674 | 27·4 | (8·9, 45·9) | 353 [235, 477] |
| 1·10 | 2·07 | 6·8 | 0·670 | 27·5 | (8·9, 46·0) | 354 [238, 476] |

Table 5.3: **Sensitivity analysis on outputs for research questions 1 & 2**. The first 3 columns describe the changes made to the OD distribution that fed through to DD distribution changes. The next 3 columns describe outputs pertaining to the first research question, and the last column the second (§ 5.1.4).

## 5.4 Discussion

This study comes at a critical point in VL elimination, where high-endemicity districts are predicted to be the hardest in which to reach the elimination target [114]. This analysis of the Vaishali pilot study suggests that combining existing interventions with special attention to quality might contribute to additional reductions in VL incidence.

Descriptive analyses suggested a significant change in the case counts in Vaishali for the first two pilot years 2015–2016 relative to other districts, which is supported by this detailed spatiotemporal analysis that accounts for decreasing trends in cases pre-pilot and neighbouring district effects. When the study started, 15 out of 16 blocks in Vaishali were above the elimination target of 1 case/10,000 people/year, but all blocks apart from Raghopur (where flooding interrupted the pilot in August 2017) were below the target at the end of 2017. Model simulations characterising the pilot period suggest that several hundred cases have been averted since 2015, which was robust to changes in the OD distribution.

One cannot conclusively attribute the additional decline in case counts in Vaishali from 2015 to the intensified control programme because this is an observational study. For internal validity of an ITSA, the continuity assumption must be met so that one is reasonably confident that “no other interventions or confounding covariates than the treatment of interest in analyses changed” at the intervention start month [13]. As the pilot & initial decline were concurrent and because no other widespread interventions were in place (§5.1.2), one concludes that the additional decline was most likely due to the intensified interventions.

### 5.4.1 Limitations

This study does not apportion how much each of the pilot’s triad of interventions contributed to the decline nor does it include covariates that describe the time-varying susceptibility of sandflies to the deployed insecticides. Modelling suggests this pilot’s high 90% household coverage per block (§5.1.3:2) would have been insufficient alone to reach disease elimination [71]. In addition, a recent study in two highly-endemic districts of Bihar suggests IRS, as implemented under the national control programme, has a negligible impact on sandfly abundance [158]. The examiners inquired how the efficacy of this intervention may realistically vary for the same pathogen. It is expected that the efficacy of VL interventions targetting case detection or reducing sandfly densities (as covered by the triad here) could depend (possibly non-linearly) on the endemicity of VL cases in the community or seasonal environmental conditions. Vector abundance, insecticide susceptibility & IRS coverage data from Bihar’s districts and other surrounding Indian districts with lower endemicities would allow further investigation.

‘Single-world’ matching of counterfactual simulations to their corresponding pilot simulations could produce a similar point estimate for cases averted but with lower stochastic variation [102], producing the averted estimate with a narrower uncertainty band (c.f. Fig. 5.10) but is beyond this study’s scope. A control group is also lacking as the 32 comparison districts could have unobserved confounders distributed heterogeneously across them, which limits the external validity of the analysis. Furthermore, inferences are made from a pilot in a single district.

The treatment information of some Vaishali cases that chose nearby district hospitals or other districts’ cases migrating into Vaishali is unknown, affecting the estimated contributions of the epidemic & neighbourhood terms in the model to Vaishali’s case counts. It is also unclear how drug supply may have impacted incidence since the national programme introduced single-dose liposomal amphotericin B in 2015–2016. Some of the largest differences among the 32 non-Vaishali districts are the VL endemicity & mean OD [99]; however, this model does not account for these heterogeneities. If ASHA training reduced OD times and thus infectious durations and subsequent incidence, this would have also shortened the DD distribution, meaning that the inferences & predictions here are biased. However, any large reductions expected in the infectious duration would only marginally affect the OD distribution as the mean infectious period was only 11% of the mean DD.

Case underreporting is estimated at 15–18% in Vaishali in 2012–2013, with a non-uniform age distribution that may have affected these results [53, 99]. However, NVBDCP introduced mandatory VL reporting state-wide for the public sector on 7 January 2016. Although HIV/TB-VL coinfection data is included in the monthly cases, stratification of their status in the model was not performed due to this data only being available since 2015. In a Vaishali district hospital in 2011–2013, VL admissions who were unknowingly HIV+ had OD times on average 3 weeks longer [35]; their underdiagnosed HIV-VL status accounted for 2·4% of admissions, rising to 5% in middle-aged men. This may also be important if HIV-VL-coinfected individuals contribute disproportionately to transmission [36]. If they do, then the pilot effect in 2017 for Vaishali, a district with a rising proportion of HIV-VL coinfections, may be underestimated. Furthermore, PKDL cases are not incorporated into the analysis as case counts were unavailable from 2012 but recent studies suggest they contribute substantially to transmission as VL incidence declines [44, 133].

Despite these limitations, further pilots are recommended in highly-endemic settings with additional collection of time-varying district covariates and assessment of cost-effectiveness. Widening research questions & study design to the recent 2030 goals (§1.1) would support India's elimination efforts.

## 5.5 Further model developments

Introducing a time-varying overdispersion term $\psi_{i,t}$ to account for districts like Vaishali whose endemicities change throughout the study would likely improve the model fit (similar to the time-specific endemic component intercept $\alpha_{i,t}^{(\nu)}$ in Eqn. 5.9; as would weighting neighbourhood adjacency ($\omega_{ji}$ in Eqn. 5.5) by the proportion of the shared edge to the perimeter of district $i$. Unfortunately, the former is not possible under the current version of the `surveillance` package.

The selected island of 33 districts could underestimate the full neighbourhood effects for two reasons. Firstly, the five unsurveilled districts in south-west Bihar may have had unreported cases. Secondly, the effect of neighbouring states like Uttar Pradesh, Jharkhand & West Bengal or the Nepalese border, which, albeit are relatively low-endemicity zones, is not accounted for [141, 205]. The latter could be addressed within the `surveillance` framework by modelling entire neighbouring states as additional 'district' units.

Given that the disease has a relatively long & varied IP, it is reasonable to expect that cases are temporally-related through the months. As case diagnosis dates have been used, it is unclear how this correlation may be obscured by unobserved changes in the OD time distribution.

## 5.6 Conclusion

Can intensified control reduce VL incidence more quickly in a highly-endemic district? This robust analysis shows that observed VL case counts did fall more quickly in Vaishali district than other districts, in line with previous crude analyses [108, 109] and estimates an additional outcome indicator as ‘cases averted’. Since the design of this study (2014), VL policy now covers PKDL burden & VL mortality (§1.1). To meet these policy updates, there is justification for piloting this approach in other highly-endemic settings, contingent on improvements in study design & analysis (§5.4.1 & §5.5). This is the final research chapter, and so in the next chapter the findings of this thesis are summarised & concluded.

# CHAPTER 6

# Conclusions

*I have performed VL analyses at two spatial scales that have included disease clustering. I have also made contributions to the tau statistic and provided a model example of an interrupted time series analysis (ITSA) using the popular Held et al. spatiotemporal modelling framework &* `surveillance` *R package. Finally, I outline future avenues of exploration, covering VL, the tau statistic & Held et al.'s spatiotemporal modelling framework.*

## 6.1 An overview of the research progress made

The review in **Chapter 2** covered the first eight years' use & development of the tau statistic. Some open issues remain regarding the implementation & use of the tau statistic. These include its non-uniqueness to the choice of distance band set, time window selection for temporal relatedness of case pairs, and comparison with other modern spatiotemporal statistics. Epidemiologists using the statistic need to be careful not to introduce faulty assumptions through pair-relatedness parameters or that which would cause misclassification of pairs.

Statistical inference covers point estimation, interval estimation & hypothesis testing. Chapters 2–3 have advanced statistical inference methods for the tau statistic. Chapter 2 introduced the tau rate estimator, specifically for studies with participants exposed to varying times at risk, either due to migration or immunological status. In **Chapter 3**, I obtained a narrower (confidence) interval estimate for the clustering endpoint distance using an updated spatial bootstrapping schema and separating point estimation from graphical hypothesis testing. For active case detection strategies whose intervention radius & length could be informed by this statistic, improvements in bias & precision of estimates could help save valuable public health resources. This will benefit epidemiologists or infectious disease modellers who wish to characterise spatiotemporal clustering for their descriptive analyses.

This work has given a spatiotemporal treatment of two VL datasets to answer two research aims—understanding how VL disease risk varies with distance, and developing spatiotemporal models for VL. In quantifying the VL risk profile as a function of distance (the first research aim), **Chapter 4** provides important information for guiding spatially-targeted VL interventions, in terms of active case detection & indoor residual spraying (IRS) around cases. It also partially validates an earlier spatial kernel transmission model estimate [44], by providing a similar estimate of the clustering endpoint distance with fewer assumptions. In the process of analysis, I made contributions to theory and use of the tau statistic: introducing a rigorous inferential framework that can be used to assess if clustering is present, estimate the precision of the clustering estimate and suggest times when the onsets of cases after the index case are likely to be highest. I have also contributed open source code for these advances, which has been submitted as a provisional pull request to the `IDSpatialStats` *R* package [166]. Pending publication of Chapter 4, this pull request will be amended and go live. Person-time at risk is a common feature of longitudinal epidemiological datasets with continual study recruitment or migratory populations that could bias the tau statistic—this has not until now been addressed in its formulation. This is addressed by the development of the new tau rate estimator in Chapter 4 which has made the statistic more versatile to different data types—three main estimators for the distance-form are now available. I hope these advances can serve other infectious diseases too. A concluding theme of Chapters 3–4 is the need to validate the tau statistic against other spatiotemporal statistics. This could be through applying a range of statistics including the tau statistic & Ripley's spatiotemporal K function to simulated data from well-known point process models [57]. Another issue is how to choose distance band sets. Disease control programmes have already benefitted from the tau statistic and these improvements and further research will safeguard future health decisions informed by it. Answering just some of the remaining open questions will foster confidence and, if warranted, will boost adoption of what appears to be a very useful statistic for health research.

In **Chapter 5** I analysed a pilot study of intensified VL control interventions in Vaishali district, India conducted by RMRIMS during 2015–2017.

There are a number of novel aspects of this analysis. It used a customisation of the Held et al. `hhh4` spatiotemporal modelling framework for the ITSA research question. Firstly, it used the distributed-lag extension of the model from the recent `hhh4addon` package to account for the influence of previous months' cases on the current month's cases, using the diagnosis-to-diagnosis distribution for the distributed-lags as the data was observed at the diagnosis month. Also, due to the distributed delay between infection & diagnosis, the full effect of the intervention is not fully seen in the observed data for up to a year and instead builds up gradually. So this subtle effect had to be captured in a $c_t$ correction term for the weighting of previous months' cases in the autoregressive component of the model for the first 12 months of the intervention in Eqn. 5.12 (§5.2.3.2). Descriptive analyses were presented justifying the application of a spatiotemporal model to the dataset that includes transmission from immediate neighbours & seasonality.

There was no available information on the deployment of IRS on a district basis, changes in sandfly insecticide susceptibility with time, nor a detailed timeline & expenditure on information, education & communication activities. This left no choice but to group all three control activities together and assume they all started synchronously (January 2015).

Two ways in which this research could be extended and improved are to: i) apply the single-world matching approach of Kaminsky et al. [102] to produce a more precise estimate of the cases averted; & ii) compare the impact HIV-VL cases have on incident VL cases versus the force of infection from PKDL cases or prevalent VL cases. HIV-VL coinfected cases have been available in India's routine VL data since 2015. The time series of VL & HIV-VL incidence could be modelled as a bivariate time series using the `surveillance` package [150].

The results of Chapter 5's analysis are encouraging for the significant benefits intensified interventions could bring to these affected populations. They are also highly relevant to policymakers during the current policy window in which control targets for 2030 have recently been formulated following the latest WHO roadmap for 2021–2030. Furthermore, it could contribute to control strategy across the Indian subcontinent (ISC). These modern modelling techniques applied to a district-wide pilot that accounted for secular trends & spatial heterogeneity would be of interest to epidemiologists & medical statisticians. Until now the `surveillance` & `hhh4addon` packages have only been used to fit static-intervention models. Chapter 5 in parts provides a pedagogical explanation more lengthy than a traditional paper which is complementary to the existing package vignettes [100, 131]—benefitting those who code in the *R* language and wish to employ this modelling framework to answer ITSA research questions for other interventions.

## 6.2 Wider limitations & new research avenues

I expand on the chapter-specific limitations already discussed by considering the wider limitations of this thesis and gaps in the current research base. Both of the datasets I have used focus on human VL cases only, while missing the connection with the sandfly vector & environmental factors. Despite the rapid review not covering Brazilian studies, when briefly read, many had integrated epidemiological, entomological & environmental data, unlike their Indian study counterparts. This would require a coordinated approach for study design within the ISC: maintaining sandfly traps & testing facilities on a regular & spatially-representative sample, in conjunction with a study's human VL cases & environmental variables [37]. Through a similar approach (regular sandfly trapping and testing of spray quality & insecticide susceptibility), the Vaishali intervention could have disentangled how IRS contributed to the impact of the intensified control intervention.

Throughout this thesis the tau statistic has only used onset date to infer probable transmission pairs. As Salje warned during this thesis' viva (personal comm.), this can make the tau statistic vulnerable to spatiotemporal signals from non-primary spatially-close transmission chains from the same (point source epidemic) or unrelated transmission chains (propagated epidemic). I accept this, however the choice of the tau statistic as a non-parametric estimator of spatiotemporal dependence appeared to be the best choice out of the range of statistics available as detailed in §2.2. Unfortunately for VL, no relatedness variables like serotype & genotype are available unlike for dengue and other diseases reviewed in Chapter 2. For other diseases with the same difficulty as VL, an improvement to the statistic based on simulation, not study design will solve the problem. Despite the expected underestimate in the magnitude of tau at close distances, the statistic has still been able to capture $\hat{D}$ remarkably well when compared to a model-based estimate [44].

Furthermore, Salje highlighted the reliance of this thesis on the same serial interval (either as the 7 month mean in Chapter 4 or as a distribution in Chapter 5). I decided to ignore lower-quality estimates in the literature that lacked a source themselves or which had reportedly (Bern, personal comm.) been decided by consensus among a panel of experts. The examiners suggested a sensitivity analysis in Chapter 5 which was accepted and made to the infection period that made up the (diagnosis-to-diagnosis) serial interval distribution which was a key input to the spatiotemporal model and further strengthened that analysis.

Salje (personal comm.) makes a useful point of whether a tau statistic based on solely temporal-relatedness conditions would work in "sparsely sampled settings". Although it has been shown in Lessler et al. [118] to be robust at just 1% of cases observed with respect to both magnitude of tau & clustering range, their simulations presumably used serotype or genotype to identify probable transmission pairs. Again, simulation could verify if robustness also holds for temporal-relatedness conditions only.

Until now the tau statistic (including in this thesis) has focussed on the extent of spatial dependence (Salje, personal comm.) and/or the magnitude of the spatial effect. Salje has critiqued this thesis' "focus on the extent of dependence" while ignoring its magnitude, however I believe it is subject to the disease in question. For exploratory research into a disease of an unknown origin, I agree that the magnitude of the spatial effect would be more important to answer the infectious disease hypothesis and describe likely modes of transmission. However, for policymakers of an established disease like VL, they know that any untreated cases are likely to lead to death and so knowing the spatial extent to optimise surveillance is more relevant. Also as a temporally-related variable (date of onset) was only available for VL, I did not wish to focus on the magnitude of clustering, knowing that it would be an underestimate due to Lessler et al.'s investigations when additional relatedness variables are available [118]. A third aspect which has not yet been explored is to estimate the number of "additional cases" (Salje, personal comm.) that arise due to spatial clustering, to quantify the maximum cases that could be averted for a spatially-targetted intervention; this would apply more to diseases that have a range of control options that need comparison e.g. active case detection, school/work-based interventions etc.

Better data & greater model sophistication could better inform VL infection dynamics—not just for current control, but also new prophylaxis & vaccine delivery protocol. The best marker of protection (the leishmanin skin test [16]) is not currently available and lacks laboratory standardisation thus preventing inter-study comparison [43]. While this epidemiological tool is unavailable and asymptomatic infection is lurking, one cannot know for sure when a village may next be susceptible to a VL epidemic in years to come. Here the strength of modelling is its ability to provide quantitative answers based on the uncertainties in data around such hidden immune states.

The Bangladeshi study could have better accounted for changes in sandfly populations through a parallel sandfly capture study. Both dataset analyses assume fixed geolocations of cases & non-cases, yet humans may be bitten at night away from their residence or when sleeping outdoors overnight. Cheap GPS devices now exist that can be distributed to study participants [184] to collect contact patterns by age & sex. By neglecting (hourly) human movement, the spatial clustering found here & in Chapman et al. [44] can only be explained by sandfly movement between fixed household, thus the current clustering range estimates are likely to have been overestimated.

It is unclear how the spatial extent of clustering $\hat{D}$, would change over time. Provisional testing of a month-based infector onset as a remedy to temporal confounding in the estimation of $\hat{T}$ drew numerical instabilities as there were insufficient data points within the distance × time bands—a similar outcome is expected for yearly estimates of $\hat{D}$. Hypothetically, $\hat{D}$ would expect to widen compared to the early years of the Fulbaria epidemic in Chapter 4 as spatially-local depletion of susceptibles around infectors [44]. Yet in the transmission trees inferred by the Chapman et al. spatiotemporal model, rather than transmission spreading radially from infectors with time, "short and long jumps in space" per infection generation were inferred.

Earlier in §3.2.4, spatiotemporal inhibition was discounted as it did not directly relate to control activities around a household. However, the process of inhibition that arises as a consequence of infection has had little attention and could reveal important features of an infectious disease—such as the proportion of those with immunity & its rate of change, the level of protection afforded and topographical barriers for humans & vectors. Understanding the sensitivity of the startpoint of spatiotemporal inhibition to these attributes could indicate how much a descriptive analysis of a spatially-heterogeneous infectious disease using this parameter can tell us.

HIV status, although at low prevalence in the ISC compared to other regions globally, could disproportionately contribute to transmission through HIV-VL coinfected superspreaders [36]. They would likely have extended (VL) infectious periods & VL treatment failure. However, the parasite load & existing disease histories (including HIV status) of presenting VL patients are not routinely collected by epidemiological studies.

Finally, I made the grounding assumption that cases only arise from a single infectious contact [127]. Examining alternative processes is possible yet unexplored under the Chapman et al.

model and could also link with within-host VL parasite modelling. High VL endemicity is another common theme linking both datasets and it is unclear whether the results of the analyses would translate to areas with lower endemicities. The spatial infection process requires a vector and it is likely that clustering estimates seen at high infection densities are non-linear with incidence and transmission may even be halted at certain mean density thresholds.

Chapter 4's estimates should be treated with caution for control policy purposes: our values during a time of high VL endemicity in Bangladesh in the mid-2000s, will unlikely generalise to the far lower VL endemicities (thankfully) seen there today and the rest of the Indian subcontinent; this thesis clearly shows that the tau clustering statistic is still in need of further development to address its calculation and sensitivity to bias under different implementations; finally targeting interventions around and within a period of time after index case detection where incidence is 'spatiotemporally' above average may not necessarily be the optimal to avert new cases or for programme efficiency. However, Chapter 4 does validate a recent & more complex model-based clustering estimate. These values can also provide prior information for new simulation models or epidemiological trials planning index-case targeted approaches.

WHO has set an elimination as a public health problem (EPHP) goal for VL. Their reticence for more ambitious targets is explained in their 2021–2030 roadmap [206]. They identify gaps in the scientific "understanding of disease epidemiology and pathology" that "would hinder progress towards achieving [even these EPHP] targets", let alone considering true 'elimination' (interruption of transmission & zero cases within a country) or 'eradication' (around the globe of *L. donovani* parasite). The new roadmap goal to cut the VL case fatality rate below 1% is vulnerable to "perverse incentives": by reporting the cause of death under other associated causes, the goal can artificially be met [140]. Additionally, there has not been an attempt in models to simulate VL deaths, owing to lack of recent data, vulnerable to the underreporting just mentioned [207]. Further work on simulating from the Chapman et al. model [44] is underway, which can help answer questions on elimination given different proportions of susceptibles and treatment delays.

* * *

The task to deliver VL modelling insights is not an easy one: it requires strong collaborations between modellers & policymakers to understand the most relevant research questions of the day, and with close links between modellers and those designing subsequent epidemiological studies to ensure the correct variables are collected & bias avoided. This modelling-derived knowledge could have a considerable impact on development & welfare, relieving the economic burden of disrupted lives & costly disease control while significantly reducing the death toll, potentially leading to true elimination. Modelling & quantitative analyses can contribute important insights to support the development, deployment & evaluation of more effective control interventions, and support the sustainable elimination of VL as a public health problem in India. The results of this thesis will contribute to this process through important insights into the spatiotemporal transmission of the disease and impact of control interventions that can be used to design more effective & efficient control policies.

# CHAPTER 7

# Afterword

This piece of research has endeavoured to be of high quality by following best practice. I co-conceptualised 'Unit testing for infectious disease epidemiology' [122], that highlighted a lack of unit testing concepts in the field using toy examples. I have made analysis code available to the public on GitHub repositories, added unit tests to my code to detect errors and use Mersenne Twister random number generators for confidence in the randomisation procedure. Code was also reviewed by Lloyd Chapman for Chapter 3. Modern modelling methods are employed with limitations described in the chapter-specific conclusions. Uncertainty in parameters is expressed. I include open source packages used in the cited references to ensure they get the academic credit they deserve. The CRediT framework is used to ensure that significant non-author contributions are properly credited and to detail the exact contributions of authors that can be ambiguous from author list order.

At the culmination of this PhD, what I have learnt in this training programme & other achievements are as important to me as the research outputs. In 2017 I was successful in the Newton Bhabha PhD programme and awarded British Council funding to spend four months in Bihar, India, hosted within the national VL control programme. This was a competitive award and enabled collaboration with epidemiologists & vector biologists leading to the manuscript associated with Chapter 5. I participated in the American Mathematics Society 'Mathematics Research Communities' virtual research group in May 2021 into the parameter identifiability of compartmental models. I reached the *3 Minute Thesis* competition final at the University of Warwick where I summarised my research to a lay audience.

# References

[1] Abel, G. J. (2015). `fanplot` *R* Package v4.0.0. *R J.*, 7(2):15–23.

[2] Alam, M., Hasan, N. A., Sadique, A., Bhuiyan, N. A., Ahmed, K. U., Nusrin, S., Nair, G. B., Siddique, A. K., Sack, R. B., Sack, D. A., Huq, A., and Colwell, R. R. (2006). Seasonal Cholera Caused by Vibrio cholerae Serogroups O1 and O139 in the Coastal Aquatic Environment of Bangladesh. *Appl. Environ. Microbiol.*, 72(6):4096–4104.

[3] Aldstadt, J. (2007). An incremental Knox test for the determination of the serial interval between successive cases of an infectious disease. *Stoch. Environ. Res. Risk. Assess.*, 21(5):487–500.

[4] Aldstadt, J., Yoon, I.-k., Tannitisupawong, D., Jarman, R. G., Thomas, S. J., Gibbons, R. V., Uppapong, A., Iamsirithaworn, S., Rothman, A. L., Scott, T. W., and Endy, T. (2012). Space-time analysis of hospitalised dengue patients in rural Thailand reveals important temporal intervals in the pattern of dengue virus transmission. *Trop. Med. Int. Health*, 17(9):1076–1085.

[5] Allen, L., O'Connell, A., and Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learn. Publ.*, 32:71–74.

[6] Alvar, J., Vélez, I. D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., den Boer, M., and the WHO Leishmaniasis Control Team (2012). Leishmaniasis Worldwide and Global Estimates of Its Incidence. *PLoS ONE*, 7(5):1–12.

[7] Anderson, N. H. and Titterington, D. M. (1997). Some Methods for Investigating Spatial Clustering, with Epidemiological Applications. *J. Royal Stat. Soc. Ser. A*, 160(1):87–105.

[8] Azman, A. S., Luquero, F. J., Salje, H., Mbaïbardoum, N. N., Adalbert, N., Ali, M., Bertuzzo, E., Finger, F., Toure, B., Massing, L. A., Ramazani, R., Saga, B., Allan, M., Olson, D., Leglise, J., Porten, K., and Lessler, J. (2018). Micro-Hotspots of Risk in Urban Cholera Epidemics. *J. Inf. Dis.*, 218(7):1164–1168.

[9] Azman, A. S., Rumunu, J., Abubakar, A., West, H., Ciglenecki, I., Helderman, T., Wamala, J. F., de la Rosa Vázquez, R., Perea, W., Sack, D. A., Legros, D., Martin, S., Lessler, J., and Luquero, F. J. (2016). Population-Level Effect of Cholera Vaccine on Displaced Populations, South Sudan, 2014. *Emerg. Infect. Dis.*, 22(6):1067–1070.

[10] Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* CRC Press/Taylor & Francis, Boca Raton, first edition.

[11] Baddeley, A. and Turner, R. (2005). `spatstat` v1.61-0: An R package for analyzing spatial point patterns. *J. Stat. Softw.*, 12(6):1–42.

[12] Barnett, P. G., Singh, S. P., Bern, C., Hightower, A. W., and Sundar, S. (2005). VIRGIN SOIL : THE SPREAD OF VISCERAL LEISHMANIASIS INTO UTTAR PRADESH , INDIA. *Am. J. Trop. Med. Hyg.*, 73(4):720–725.

[13] Bärnighausen, T., Oldenburg, C., Tugwell, P., Bommer, C., Ebert, C., Barreto, M., Djimeu, E., Haber, N., Waddington, H., Rockers, P., Sianesi, B., Bor, J., Fink, G., Valentine, J., Tanner, J., Stanley, T., Sierra, E., Tchetgen, E. T., Atun, R., and Vollmer, S. (2017). Quasi-experimental study designs series—paper 7: assessing the assumptions. *J. Clin. Epidemiol.*, 89:53–66.

[14] Bates, P. A. (2007). Transmission of leishmania metacyclic promastigotes by phlebotomine sand flies. *Int. J. Parasitol.*, 37(10):1097–1106.

[15] Berman, J. (2006). Visceral leishmaniasis in the New World & Africa. *Indian J Med Res*, (123):289–294.

[16] Bern, C., Amann, J., Haque, R., Chowdhury, R., Ali, M., Kurkjian, K. M., Vaz, L., Wagatsuma, Y., Breiman, R. F., Secor, W. E., and Maguire, J. H. (2006). Loss of leishmanin skin test antigen sensitivity and potency in a longitudinal study of visceral Leishmaniasis in Bangladesh. *Am. J. Trop. Med. Hyg.*, 75(4):744–748.

[17] Bern, C., Courtenay, O., and Alvar, J. (2010). Of Cattle, Sand Flies and Men: A Systematic Review of Risk Factor Analyses for South Asian Visceral Leishmaniasis and Implications for Elimination. *PLoS Negl. Trop. Dis.*, 4(2):1–9.

[18] Bern, C., Hightower, A. W., Chowdhury, R., Ali, M., Amann, J., Wagatsuma, Y., Haque, R., Kurkjian, K., Vaz, L. E., Begum, M., Akter, T., Cetre-Sossah, C. B., Ahluwalia, I. B., Dotson, E., Secor, W. E., Breiman, R. F., and Maguire, J. H. (2005). Risk Factors for Kala-Azar in Bangladesh. *Emerg. Inf. Dis.*, 11(5):655–662.

[19] Bernal, J. L., Cummins, S., and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.*, 46(1):348–355.

[20] Bhoomiboonchoo, P., Gibbons, R. V., Huang, A., Yoon, I.-K., Buddhari, D., Nisalak, A., Chansatiporn, N., Thipayamongkolgul, M., Kalanarooj, S., Endy, T., Rothman, A. L., Srikiatkhachorn, A., Green, S., Mammen, M. P., Cummings, D. A., and Salje, H. (2014). The Spatial Dynamics of Dengue Virus in Kamphaeng Phet, Thailand. *PLoS Negl. Trop. Dis.*, 8(9):6–11.

[21] Bhunia, G. S., Chatterjee, N., Kumar, V., Siddiqui, N. A., Mandal, R., Das, P., and Kesari, S. (2012). Delimitation of kala-azar risk areas in the district of Vaishali in Bihar (India) using a geo-environmental approach. *Mem. Inst. Oswaldo Cruz*, 107(5):609–620.

[22] Bhunia, G. S., Kesari, S., Chatterjee, N., Kumar, V., and Das, P. (2013). Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. *BMC Inf. Dis.*, 13(1).

[23] Bivand, R., Keitt, T., and Rowlingson, B. (2021). `rgdal` *R* package v1.5-27. `CRAN.R-project.org/package=rgdal`.

[24] Bivand, R. and Rundel, C. (2021). `rgeos` *R* package v0.5-8. `CRAN.R-project.org/package=rgeos`.

[25] Bivand, R. and Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association, `spdep` *R* package v1.1-11. *TEST*, 27(3):716–748.

[26] Bland, M. (2015). *An Introduction to Medical Statistics*. OUP Oxford, Oxford, UK, fourth edition.

[27] Bode, B. (2002). Analyzing power structures in rural Bangladesh. `pqdl.care.org/CuttingEdge/Analyzing%20Power%20Structures%20in%20Rural%20Bangladesh.pdf`. Accessed: 2019-01-09.

[28] Boelaert, M. and Sundar, S. (2014). Leishmaniasis. In Farrar, J., editor, *Manson's Tropical Diseases*, chapter 47, pages 631–651. Elsevier/Saunders, Philadelphia, Pennsylvania, 23$^{\text{rd}}$ edition.

[29] Boettcher, J. P., Siwakoti, Y., Milojkovic, A., Siddiqui, N. A., Gurung, C. K., Rijal, S., Das, P., Kroeger, A., and Banjara, M. R. (2015). Visceral leishmaniasis diagnosis and reporting delays as an obstacle to timely response actions in Nepal and India. *BMC Inf. Dis.*, 15(1):1–14.

[30] Boily, M. C., Baggaley, R. F., Wang, L., Masse, B., White, R. G., Hayes, R. J., and Alary, M. (2009). Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *Lancet Inf. Dis.*, 9(2):118–129.

[31] Bracher, J. (2020). `hhh4addon` *R* package v0.0.0.0.9014. `doi.org/10.5281/zenodo.4696125`.

[32] Bracher, J. and Held, L. (2020). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *Int. J. Forecast.*

[33] Bulstra, C. A., Le Rutte, E. A., Malaviya, P., Hasker, E. C., Coffeng, L. E., Picado, A., Singh, O. P., Boelaert, M. C., de Vlas, S. J., and Sundar, S. (2018). Visceral leishmaniasis: Spatiotemporal heterogeneity and drivers underlying the hotspots in Muzaffarpur, Bihar, India. *PLoS Negl. Trop. Dis.*, 12(12):1–21.

[34] Burza, S., Croft, S., and Boelaert, M. (2018). Leishmaniasis. *Lancet*.

[35] Burza, S., Mahajan, R., Sanz, M. G., Sunyoto, T., Kumar, R., Mitra, G., and Lima, M. A. (2014a). HIV and Visceral Leishmaniasis Coinfection in Bihar, India: An Underrecognized and Underdiagnosed Threat Against Elimination. *Clin. Inf. Dis.*, 59(4):552–555.

[36] Burza, S., Mahajan, R., Sinha, P. K., van Griensven, J., Pandey, K., Lima, M. A., Sanz, M. G., Sunyoto, T., Kumar, S., Mitra, G., Kumar, R., Verma, N., and Das, P. (2014b). Visceral Leishmaniasis and HIV Co-infection in Bihar, India: Long-term Effectiveness and Treatment Outcomes with Liposomal Amphotericin B (AmBisome). *PLoS Negl. Trop. Dis.*, 8(8):1–12.

[37] Cameron, M. M., Acosta-Serrano, A., Bern, C., Boelaert, M., den Boer, M., Burza, S., Chapman, L. A. C., Chaskopoulou, A., Coleman, M., Courtenay, O., Croft, S., Das, P., Dilger, E., Foster, G., Garlapati, R., Haines, L., Harris, A., Hemingway, J., Hollingsworth, T. D., Jervis, S., Medley, G., Miles, M., Paine, M., Picado, A., Poché, R., Ready, P., Rogers, M., Rowland, M., Sundar, S., de Vlas, S. J., and Weetman, D. (2016). Understanding the transmission dynamics of Leishmania donovani to provide robust evidence for interventions to eliminate visceral leishmaniasis in Bihar, India. *Parasites & Vectors*, 9(1):25.

[38] Campbell, F., Strang, C., Ferguson, N., Cori, A., and Jombart, T. (2018). When are pathogen genome sequences informative of transmission events? *PLoS Pathog.*, 14(2):1–21.

[39] Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, 19:1141–1164.

[40] CDC (2021). Global Health - Neglected Tropical Diseases. `cdc.gov/globalhealth/ntd`. Accessed: 2021-11-25.

[41] CDC (2022). Parasite lifecycle image. `cdc.gov/dpdx/leishmaniasis/index.html`. Accessed: 2022-01-08.

[42] Chapman, L. A. C., Dyson, L., Courtenay, O., Bern, C., Medley, G. F., and Hollingsworth, T. D. (2015). Quantification of the natural history of visceral leishmaniasis and consequences for control. *Parasites & Vectors*, 8(521).

[43] Chapman, L. A. C., Morgan, A. L. K., Adams, E. R., Bern, C., Medley, G. F., and Hollingsworth, T. D. (2018). Age trends in asymptomatic and symptomatic Leishmania donovani infection in the Indian subcontinent: A review and analysis of data from diagnostic and epidemiological studies. *PLoS Negl. Trop. Dis.*, 12(12):e0006803.

[44] Chapman, L. A. C., Spencer, S. E. F., Pollington, T. M., Jewell, C. P., Mondal, D., Alvar, J., Hollingsworth, T. D., Cameron, M. M., Bern, C., and Medley, G. F. (2020). Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis. *PNAS*, 117(41):25742–25750.

[45] Chowdhury, R., Mondal, D., Chowdhury, V., Faria, S., Alvar, J., Nabi, S. G., Boelaert, M., and Dash, A. P. (2014). How far are we from visceral leishmaniasis elimination in Bangladesh? An assessment of epidemiological surveillance data. *PLoS Negl. Trop. Dis.*, 8(8):1–10.

[46] Coleman, M., Foster, G., Deb, R., and Srikantiah, S. (2017). Enhancing surveillance of the visceral leishmaniasis elimination programme in India. `worldleish2017.org/documentos/Abstracts/Book/WL6/2017.pdf`. Accessed: 2018-03-21.

[47] Cori, A., Cauchemez, S., Ferguson, N. M., Fraser, C., Dahlqwist, E., Demarsh, P. A., Jombart, T., Kamvar, Z. N., Lessler, J., Li, S., Polonsky, J. A., Stockwin, J., Thompson, R., and van Gaalen, R. (2020). `EpiEstim` v2.2-4 *R* package. `doi.org/10.5281/zenodo.3685977`.

[48] Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *Am. J. Epidemiol.*, 178(9):1505–1512.

[49] Cummings, P. (2009). The Relative Merits of Risk Ratios and Odds Ratios. *JAMA Pediatr.*, 163(5):438–445.

[50] Cunningham, J., Hasker, E., Das, P., El Safi, S., Goto, H., Mondal, D., Mbuchi, M., Mukhtar, M., Rabello, A., Rijal, S., Sundar, S., Wasunna, M., Adams, E., Menten, J., Peeling, R., Boelaert, M., and Network, W. V. L. L. (2012). A global comparative evaluation of commercial immunochromatographic rapid diagnostic tests for visceral leishmaniasis. *Clin Infect Dis*, 10(55):1312–9.

[51] Cuzick, J. and Edwards, R. (1990). Spatial Clustering for Inhomogeneous Populations. *J. Royal Stat. Soc. Ser. B*, 52(1):73–104.

[52] Czado, C., Gneiting, T., and Held, L. (2009). Predictive Model Assessment for Count Data. *Biom.*, 65(4):1254–1261.

[53] Das, A., Karthick, M., Dwivedi, S., Banerjee, I., Mahapatra, T., Srikantiah, S., and Chaudhuri, I. (2016a). Epidemiologic Correlates of Mortality among Symptomatic Visceral Leishmaniasis Cases: Findings from Situation Assessment in High Endemic Foci in India. *PLoS Negl. Trop. Dis.*, 10(11):1–12.

[54] Das, P., Matlashewski, G., Das, V. R., and Pandey, R. N. (2016b). *Final GCC Project Report Phase II.* RMRIMS. Patna (unpublished).

[55] Das, V. N. R., Pandey, R. N., Kumar, V., Pandey, K., Siddiqui, N. A., Verma, R. B., Matlashewski, G., and Das, P. (2016c). Repeated training of accredited social health activists (ASHAs) for improved detection of visceral leishmaniasis cases in Bihar, India. *Pathog. Glob. Health*, 110(1):33–35.

[56] Davidson, R. and MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econom. Rev.*, 19(1):55–68.

[57] Diggle, P. J. (2019). One-to-one discussion online about the provisional results of [165].

[58] Diggle, P. J., Chetwynd, A. G., Morris, S. E., and Häggkvist, R. (1995). Second-order analysis of space-time clustering. *Stat. Methods Med. Res.*, 4(2):124–136.

[59] Diggle, P. J., Kaimi, I., and Abellana, R. (2010). Partial-Likelihood Analysis of Spatio-Temporal Point-process Data. *Biometrics*, 66(2):347–354.

[60] Dunnington, D. (2017). `prettymapr` *R* package v0.2.2. `CRAN.R-project.org/package=prettymapr`.

[61] Dye, C. and Wolpert, D. M. (1988). Earthquakes, influenza and cycles of Indian kala-azar. *Trans. R. Soc. Trop. Med. Hyg.*, 82(6):843–850.

[62] Eddelbuettel, D. and François, R. (2011). `Rcpp` *R* package: Seamless R & C++ integration. *J. Stat. Softw.*, 40(8):1–18. `R` package v1.0.7.

[63] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.*, 7(1):1–26.

[64] Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.*, 82(397):171–185.

[65] Efron, B. and Narasimhan, B. (2020). The Automatic Construction of Bootstrap Confidence Intervals. *J. Comput. Graph. Stat.*, 29(3):608–619.

[66] Efron, B. and Narasimhan, B. (2021). `bcaboot` *R* package v0.2-3. `CRAN.R-project.org/package=bcaboot`.

[67] Efron, B. and Tibshirani, R. (1998). *An introduction to the bootstrap.* Boca Raton; Chapman & Hall/CRC, London.

[68] Elmojtaba, I. M., Mugisha, J. Y. T., and Hashim, M. H. A. (2013). Vaccination model for visceral leishmaniasis with infective immigrants. *Math. Models Methods Appl. Sci.*, 36(2):216–226.

[69] EPSG (2006). EPSG:32646: Geographic 2D CRS for Bangladesh UTM zone 46N. `epsg.io/32646`. Accessed: 2021-18-02.

[70] Finger, F., Bertuzzo, E., Luquero, F. J., Naibei, N., Touré, B., Allan, M., Porten, K., Lessler, J., Rinaldo, A., and Azman, A. S. (2018). The potential impact of case-area targeted interventions in response to cholera outbreaks: A modeling study. *PLoS Med.*, 15(2):1–27.

[71] Fortunato, A. K., Glasser, C. P., Watson, J. A., Lu, Y., Rychtář, J., and Taylor, D. (2021). Mathematical modelling of the use of insecticide-treated nets for elimination of visceral leishmaniasis in Bihar, India. *R. Soc. Open Sci.*, 8(6).

[72] Fraser, C. (2007). Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLoS ONE*, 2(8).

[73] Gabriel, E. and Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Stat. Neerlandica*, 63(1):43–51.

[74] Gabriel, E., Diggle, P. J., Rowlingson, B., and Rodriguez-Cortes, F. J. (2018). stpp *R* package v2.0-3. CRAN.R-project.org/package=stpp.

[75] GADM (2015). India level 2 shapefiles v2.8. gadm.org.

[76] Garbuszus, J. M. and Jeworutzki, S. (2021). readstata13 *R* package v0.10.0. CRAN.R-project.org/package=readstata13.

[77] Giles, J. R., Salje, H., and Lessler, J. (2019). The IDSpatialStats R Package: Quantifying Spatial Dependence of Infectious Disease Spread. *R J.*, 11(2):308–327.

[78] Gillespie, S. and Pearson, R. (2002). *Principles and Practice of Clinical Parasitology*. John Wiley & Sons, Chichester.

[79] Giraud, E., Martin, O., Yakob, L., and Rogers, M. (2019). Quantifying Leishmania Metacyclic Promastigotes from Individual Sandfly Bites Reveals the Efficiency of Vector Transmission. *Commun. Biol.*, 2(1):25–28.

[80] Gorard, S. (2014). Confidence intervals, missing data and imputation: a salutary illustration. *Int. J. Res. Educ. Methodol.*, 5(3):693–698.

[81] Government of India (2015). Districtwise 2011 census population data. census2011.co.in/census/district. Accessed: 2018-03-15.

[82] Government of India (2017). Kala-azar monthly programme data. *State VBD Office (Patna)*.

[83] Grabowski, M. K., Lessler, J., Redd, A. D., Kagaayi, J., Laeyendecker, O., Ndyanabo, A., Nelson, M. I., Cummings, D. A., Bwanika, J. B., Mueller, A. C., Reynolds, S. J., Munshaw, S., Ray, S. C., Lutalo, T., Manucci, J., Tobian, A. A., Chang, L. W., Beyrer, C., Jennings, J. M., Nalugoda, F., Serwadda, D., Wawer, M. J., Quinn, T. C., and Gray, R. H. (2014). The Role of Viral Introductions in Sustaining Community-Based HIV Epidemics in Rural Uganda: Evidence from Spatial Clustering, Phylogenetics, and Egocentric Transmission Models. *PLoS Med.*, 11(3).

[84] Grantz, K. H., Rane, M. S., Salje, H., Glass, G. E., Schachterle, S. E., and Cummings, D. A. T. (2016). Disparities in influenza mortality and transmission related to sociodemographic factors within Chicago in the pandemic of 1918. *PNAS*, 113(48):13839–13844.

[85] Held, L. (2020). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction talk. fields.utoronto.ca/talks/Endemic-epidemic-models-discrete-time-serial-interval-distributions-infectious-disease. Advancing Knowledge About Spatial Modeling, Infectious Diseases, Environment & Health Conference hosted by the Fields Institute.

[86] Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Stat. Model.*, 5(3):187–199.

[87] Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13$^{\text{th}}$ Armitage lecture. *Stat. Med.*, 36(22):3443–3460.

[88] Henry, L. and Wickham, H. (2019). `purrr` *R* package v0.3.3: Functional programming tools. CRAN.R-project.org/package=purrr.

[89] Heymann, D. L. et al. (2008). *Control of Communicable Diseases Manual.* APHA, 19th edition.

[90] Hill, E. M., House, T., Dhingra, M. S., Kalpravidh, W., Morzaria, S., Osmani, M. G., Brum, E., Yamage, M., Kalam, M. A., Prosser, D. J., Takekawa, J. Y., Xiao, X., Gilbert, M., and Tildesley, M. J. (2018). The impact of surveillance and control on highly pathogenic avian influenza outbreaks in poultry in Dhaka division, Bangladesh. *PLoS. Comput. Biol.*, 14(9):1–27.

[91] Hirve, S., Boelaert, M., Matlashewski, G., Mondal, D., Arana, B., Kroeger, A., and Olliaro, P. (2016). Transmission Dynamics of Visceral Leishmaniasis in the Indian Subcontinent – A Systematic Literature Review. *PLOS Negl Trop Dis*, 10(2):345–353.

[92] Hirve, S., Kroeger, A., Matlashewski, G., Mondal, D., Banjara, R., Das, P., Be-nazir, A., Arana, B., and Olliaro, P. (2017). Towards elimination of visceral leishmaniasis in the Indian subcontinent—Translating research to practice to public health. *PLoS Negl. Trop. Dis.*, pages 1–25.

[93] Hirve, S., Singh, S. P., Kumar, N., Banjara, M. R., Das, P., Sundar, S., Rijal, S., Joshi, A., Kroeger, A., Varghese, B., Thakur, C. P., Huda, M. M., and Mondal, D. (2010). Effectiveness and Feasibility of Active and Passive Case Detection in the Visceral Leishmaniasis Elimination Initiative in India, Bangladesh, and Nepal. *Am. J. Trop. Med. Hyg.*, 83(3):507–511.

[94] Hoang Quoc, C., Salje, H., Rodriguez-Barraquer, I., In-Kyu, Y., Chau, N. V. V., Hung, N. T., Tuan, H. M., Lan, P. T., Willis, B., Nisalak, A., Kalayanarooj, S., Cummings, D. A., and Simmons, C. P. (2016). Synchrony of Dengue Incidence in Ho Chi Minh City and Bangkok. *PLoS Negl. Trop. Dis.*, 10(12):1–18.

[95] Huda, M. M., Hirve, S., Siddiqui, N. A., Malaviya, P., Banjara, M. R., Das, P., Kansal, S., Gurung, C. K., Naznin, E., Rijal, S., Arana, B., Kroeger, A., and Mondal, D. (2012). Active case detection in national visceral leishmaniasis elimination programs in Bangladesh, India, and Nepal: feasibility, performance and costs. *BMC Public Health*, 12(1).

[96] Islam, S., Kenah, E., Bhuiyan, M. A. A., Rahman, K. M., Goodhew, B., Ghalib, C. M., Zahid, M. M., Ozaki, M., Rahman, M. W., Haque, R., Luby, S. P., Maguire, J. H., Martin, D., and Bern, C. (2013). Clinical and immunological aspects of post-kala-azar dermal leishmaniasis in Bangladesh. *Am. J. Epidemiol.*, 89(2):345–353.

[97] Ismail, H. M., Kumar, V., Singh, R. P., Williams, C., Shivam, P., Ghosh, A., Deb, R., Foster, G. M., Hemingway, J., Coleman, M., Coleman, M., Das, P., and Paine, M. J. I. (2016). Development of a Simple Dipstick Assay for Operational Monitoring of DDT. *PLoS Negl. Trop. Dis.*, 10(1):1–14.

[98] January a.k.a. user @ztrewq (2017). Adding figure labels (A,B,C,. . . ) in the top left corner of the plotting region. logfc.wordpress.com/2017/03/15/adding-figure-labels-a-b-c-in-the-top-left-corner-of-the-plotting-region. Accessed: 2019-10-26.

[99] Jervis, S., Chapman, L. A. C., Dwivedi, S., Karthick, M., Das, A., Rutte, E. A. L., Courtenay, O., Medley, G. F., Banerjee, I., Mahapatra, T., Chaudhuri, I., Srikantiah, S., and Hollingsworth, T. D. (2017). Variations in visceral leishmaniasis burden, mortality and the pathway to care within Bihar, India. *Parasites & Vectors*, 10(601):1–17.

[100] Johannes Bracher (2021). hhh4addon: extending the functionality of surveillance:hhh4. github.com/jbracher/hhh4addon/blob/master/vignettes/hhh4addon.html. Accessed: 2020-11-03.

[101] Joshi, A., Narain, J. P., Prasittisuk, C., Bhatia, R., Hashim, G., Jorge, A., Banjara, M., and Kroeger, A. (2008). Can visceral leishmaniasis be eliminated from Asia? *J. Vector. Borne. Dis.*, 45(2):105–111.

[102] Kaminsky, J., Keegan, L. T., Metcalf, C. J. E., Lessler, J., and Lessler, J. (2019). Perfect counterfactuals for epidemic simulations. *Phil. Trans. R. Soc. B*, pages 1–7.

[103] Kassi, M., Kassi, M., Afghan, A., Rehman, R., and Kasi, P. (2008). Marring leishmaniasis: the stigmatization and the impact of cutaneous leishmaniasis in Pakistan and Afghanistan. *PLoS Negl Trop Dis*, 2(10).

[104] Knox, E. (1989). Detection of clusters. In Elliot, P., editor, *Methodology of enquiries into disease clustering*, pages 17–20, London. Small Area Health Statistics Unit, LSHTM.

[105] Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., and Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ*, 350(h2750):1–4.

[106] Kropko, J. and Harden, J. (2019). `coxed` *R* package v0.3.0: Duration-Based Quantities of Interest for the Cox Proportional Hazards Model. CRAN.R-project.org/package=coxed.

[107] Kumar, V., Kesari, S., Dinesh, D. S., Tiwari, A. K., Kumar, A. J., Kumar, R., Singh, V. P., and Das, P. (2009). A report on the indoor residual spraying (IRS) in the control of Phlebotomus argentipes, the vector of visceral leishmaniasis in Bihar (India): an initiative towards total elimination targeting 2015 (Series-1). *J. Vector. Borne. Dis.*, 46(3):225–9. ncbi.nlm.nih.gov/pubmed/19724087.

[108] Kumar, V., Mandal, R., Das, S., Kesari, S., Dinesh, S., Pandey, K., Das, V. R., Topno, R. K., Sharma, P., Dasgupta, R. K., and Das, P. (2020). Kala-azar elimination in a highly-endemic district of Bihar, India: A success story. *PLoS Negl. Trop. Dis.*, pages 1–27.

[109] Kumar, V., Mandal, R., Kesari, S., Das, S., and Das, P. (2017). Reaching the elimination target in the district of Vaishali, Bihar, India. worldleish2017.org/documentos/Abstracts/Book/WL6/2017.pdf.

[110] Lau, M. S., Dalziel, B. D., Funk, S., McClelland, A., Tiffany, A., Riley, S., Metcalf, C. J. E., and Grenfell, B. T. (2017). Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *PNAS*, 114(9):2337–2342.

[111] Lawson, A. B. (2006). Small Scale: Disease Clustering. In Lawson, A. B., editor, *Statistical Methods in Medical Research*, chapter 6, pages 111–141. John Wiley & Sons Ltd., second edition.

[112] Lawson, A. B. (2013). Small Scale: Putative Sources of Hazard. In Lawson, A. B., editor, *Statistical Methods in Spatial Epidemiology*, chapter 7, pages 143–187. John Wiley & Sons Ltd., second edition.

[113] Lawson, A. B. and Kulldorff, M. (1999). A Review of Cluster Detection Methods. In Lawson, A., Biggeri, A., Böhning, D., Emmanuel, L., Viel, J.-F., and Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 7, pages 99–110. John Wiley & Sons Ltd., Chichester, first edition.

[114] Le Rutte, E. A., Chapman, L. A. C., Coffeng, L. E., Ruiz-Postigo, J. A., Olliaro, P. L., Adams, E. R., Hasker, E. C., Boelaert, M. C., Hollingsworth, T. D., Medley, G. F., and de Vlas, S. J. (2018). Policy Recommendations From Transmission Modeling for the Elimination of Visceral Leishmaniasis in the Indian Subcontinent. *Clin. Inf. Dis.*

[115] Le Rutte, E. A., Coffeng, L. E., Malvolti, S., Kaye, P. M., and de Vlas, S. J. (2020). The potential impact of human visceral leishmaniasis vaccines on population incidence. *PLoS Negl. Trop. Dis.*, 14(7):1–13.

[116] Lessler, J. and Giles, J. (2018). `IDSpatialStats` *R* package v0.3.7 development version. `github.com/HopkinsIDD/IDSpatialStats`.

[117] Lessler, J., Salje, H., and Giles, J. (2018). `IDSpatialStats` *R* package v0.3.7 read-only CRAN mirror. `github.com/cran/IDSpatialStats`.

[118] Lessler, J., Salje, H., Grabowski, M. K., and Cummings, D. A. T. (2016). Measuring Spatial Dependence for Infectious Disease Epidemiology. *PLoS ONE*, 11(5):1–13.

[119] Levy, J. W., Bhoomiboonchoo, P., Simasathien, S., Salje, H., Huang, A., Rangsin, R., Jarman, R. G., Fernandez, S., Klungthong, C., Hussem, K., Gibbons, R. V., and Yoon, I.-K. (2015). Elevated transmission of upper respiratory illness among new recruits in military barracks in Thailand. *Influenza. Other Respir. Viruses*, 9(6):308–314.

[120] Loh, J. M. (2008). A valid and fast spatial bootstrap for correlation functions. *Astrophys. J.*, pages 726–734.

[121] Loh, J. M. and Stein, M. L. (2004). Bootstrapping a spatial point process. *Stat. Sin.*, 14(1):69–101.

[122] Lucas, T. C. D., Pollington, T. M., Davis, E. L., and Hollingsworth, T. D. (2020). Responsible modelling: Unit testing for infectious disease epidemiology. *Epidemics*, 33(10).

[123] Mack, E. A., Malizia, N., and Rey, S. J. (2012). Population shift bias in tests of space-time interaction. *Comput. Environ. Urban. Syst.*, 36(6):500–512.

[124] Mandal, R., Kesari, S., Kumar, V., and Das, P. (2018). Trends in spatio-temporal dynamics of visceral leishmaniasis cases in a highly-endemic focus of Bihar, India: an investigation based on GIS tools. *Parasites & Vectors*, 11(220):1–9.

[125] Martin Kulldorff and Information Management Services Inc. v10.0.1 (2021). Satscan™.

[126] MathSys (2021). Thesis Presentation and Submission. warwick.ac.uk/fac/sci/mathsys/people/studentintranet/thesissubmission. Accessed: 2021-12-01.

[127] Medley, G. (2017). Email discussion about disease progression arising from the cumulation of infectious bites.

[128] Meredith, M. and Kruschke, J. (2020). HDInterval *R* package v0.2.2. CRAN.R-project.org/package=HDInterval.

[129] Meschiari, S. (2015). latex2exp *R* package v0.5.0: Use latex expressions in plots. CRAN.R-project.org/package=latex2exp.

[130] Meyer, S., Held, L., and Höhle, M. (2017). Spatio-Temporal Analysis of Epidemic Phenomena Using the *R* package surveillance v1.19.1. *J. Stat. Softw.*, 77(11):1–55.

[131] Meyer, Sebastian and Held, Leonhard and Höhle, Michael (2021). hhh4: Endemic-epidemic modeling of areal count time series. cran.r-project.org/web/packages/surveillance/vignettes/hhh4_spacetime.pdf. Accessed: 2019-01-01.

[132] Microsoft and Weston, S. (2020). doParallel *R* package v1.0.16. CRAN.R-project.org/package=doParallel.

[133] Mondal, D., Bern, C., Ghosh, D., Rashid, M., Molina, R., Chowdhury, R., Nath, R., Ghosh, P., Chapman, L. A., Alim, A., Bilbe, G., and Alvar, J. (2019). Quantifying the infectiousness of post–kala-azar dermal leishmaniasis toward sand flies. *Clin. Inf. Dis.*, 69(2):251–258.

[134] Mooney, S. J., Knox, J., and Morabia, A. (2014). The Thompson-McFadden Commission and Joseph Goldberger: Contrasting 2 Historical Investigations of Pellagra in Cotton Mill Villages in South Carolina. *Am. J. Epidemiol.*, 180(3):235–244.

[135] Myllymäki, M. (2019). Global envelope tests for spatial processes and beyond. elsevier.com/events/conferences/spatial-statistics/programme/speakers-abstract#mari. *Spatial Statistics* conference talk, Sitges, Spain. Accessed: 2019-07-13.

[136] Myllymäki, M., Mrkvička, T., Grabarnik, P., Hahn, U., Kuronen, M., Rost, M., and Seijo, H. (2019). GET: *R* package v0.1-3. cran.r-project.org/web/packages/GET.

[137] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *J. Royal. Stat. Soc. Ser. B*, 79(2):381–404.

[138] Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261.

[139] Nightingale, E. S., Chapman, L. A., Srikantiah, S., Subramanian, S., Jambulingam, P., Bracher, J., Cameron, M. M., and Medley, G. F. (2020). A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India. *PLoS Negl. Trop. Dis.*, 14(7):1–21.

[140] NTD Modelling Consortium Visceral Leishmaniasis Group (2019). Insights from mathematical modelling and quantitative analysis on the proposed WHO 2030 targets for visceral leishmaniasis on the Indian subcontinent. *Gates Open Res.*, 3:1–12.

[141] NVBDCP (2017). Kala-azar Cases and Deaths in the Country since 2010. nvbdcp.gov.in/ka-cd.html, Accessed: 2018-02-15.

[142] NVBDCP (2018). Kala-azar situation in India. nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=467&lid=3750. Accessed: 2018-11-07.

[143] Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). `fields` *R* package v9.9: Tools for spatial data. doi.org/10.5065/D6W957CT.

[144] Oesterle, H. (1992). *Statistische Reanalyse einer Masernepidemie 1861 in Hagelloch*. PhD thesis, Eberhard-Karls-Universitäat Tübingen.

[145] of Statistics, B. B. (2014). Population & housing census 2011. national volume-2: Union statistics. Technical report. 203.112.218.65:8008/WebTestApplication/userfiles/Image/National%20Reports/Union%20Statistics.pdf.

[146] Ooms, J. (2020). `rtools` *R* package v4.0.0: build base r & packages with compiled code. cran.r-project.org/bin/windows/Rtools.

[147] Oxford code review network (2021). Optimising a pair statistic in C++/Rcpp. github.com/OxfordCodeReviewNet/forum/issues/15. Accessed: 2021-07-29.

[148] Patole, S., Burza, S., and Varghese, G. M. (2014). Multiple relapses of visceral leishmaniasis in a patient with HIV in India: A treatment challenge. *Int. J. Inf. Dis.*, 25:204–206.

[149] Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat. Med.*, 30(10):1118–1136.

[150] Paul, M. and Meyer, S. (2016). hhh4: An endemic-epidemic modelling framework for infectious disease counts. cran.r-project.org/web/packages/surveillance/vignettes/hhh4.pdf. Accessed: 2019-06-16.

[151] Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. `sp` *R* package v1.4-5. *R News*, 5(2):9–13.

[152] Perry, D., Dixon, K., Garlapati, R., Gendernalik, A., Poché, D., and Poché, R. (2013). Visceral Leishmaniasis Prevalence and Associated Risk Factors in the Saran District of Bihar, India, from 2009 to July of 2011. *Am. J. Trop. Med. Hyg.*, 88(4):778–784.

[153] Pfeilsticker, A. (1863). *Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse*. PhD thesis, Eberhard-Karls-Universität Tübingen.

[154] Picado, A., Dash, A. P., Bhattacharya, S., and Boelaert, M. (2012). Vector control interventions for visceral leishmaniasis elimination initiative in South Asia, 2005–2010. *Indian J. Med. Res.*, 136(1):22–31.

[155] Picado, A., Ostyn, B., Rijal, S., Sundar, S., Singh, S. P., Chappuis, F., Das, M. L., Khanal, B., Gidwani, K., Hasker, E., Dujardin, J. C., Vanlerberghe, V., Menten, J., Coosemans, M., and Boelaert, M. (2015). Long-lasting Insecticidal Nets to Prevent Visceral Leishmaniasis in the Indian Subcontinent; Methodological Lessons Learned from a Cluster Randomised Controlled Trial. *PLoS Negl. Trop. Dis.*, 9(4):4–11.

[156] Picado, A., Singh, S. P., Rijal, S., Sundar, S., Ostyn, B., Chappuis, F., Uranw, S., Gidwani, K., Khanal, B., Rai, M., Paudel, I. S., Das, M. L., Kumar, R., Srivastava, P., Dujardin, J. C., Vanlerberghe, V., Andersen, E. W., Davies, C. R., and Boelaert, M. (2010). Longlasting insecticidal nets for prevention of Leishmania donovani infection in India and Nepal: paired cluster randomised trial. *BMJ*, 341:8.

[157] Plate, T. and Heiberger, R. (2016). `abind` *R* package v1.4-5: Combine multidimensional arrays. `CRAN.R-project.org/package=abind`.

[158] Poché, D. M., Torres-Poché, Z., Garlapati, R., Clarke, T., and Poché, R. M. (2018). Short-term movement of Phlebotomus argentipes (Diptera: Psychodidae) in a visceral leishmaniasis-endemic village in Bihar, India. *J. Vector. Ecol.*, 43(2):285–292.

[159] Pollington, T. M. (2019). Tau statistic speedup v1.1.1. `doi.org/10.5281/zenodo.3460744`.

[160] Pollington, T. M. (2021). Code analysis for Impact of intensified control strategies on incidence of visceral leishmaniasis in a highly endemic district of Bihar, India: an interrupted time series analysis: Post-review v1.3. `doi.org/10.5281/zenodo.5701378`.

[161] Pollington, T. M., Chapman, L. A. C., Bern, C., and Hollingsworth, T. D. (2016). Spatial modelling of a visceral leishmaniasis epidemic in Fulbaria, Bangladesh during 2002–10. Master's thesis, MathSys CDT, Maths Institute, University of Warwick CV4 7AL, UK. Unpublished.

[162] Pollington, T. M., Tildesley, M. J., Hollingsworth, T., and Chapman, L. A. C. (2019a). Epidemics$^7$ conference poster: Use global envelope tests not pointwise CIs, for graphical hypothesis tests of spatiotemporal clustering and tau statistic $\tau$. `doi.org/10.13140/RG.2.2.28985.52322`. Accessed: 2019-12-05.

[163] Pollington, T. M., Tildesley, M. J., Hollingsworth, T., and Chapman, L. A. C. (2019b). The spatiotemporal tau statistic: a review. `arxiv.org/abs/1911.11476`.

[164] Pollington, T. M., Tildesley, M. J., Hollingsworth, T. D., and Chapman, L. A. C. (2020a). Code repository for Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic. `doi.org/10.5281/zenodo.4906452`. Accessed: 2021-06-07.

[165] Pollington, T. M., Tildesley, M. J., Hollingsworth, T. D., and Chapman, L. A. C. (2020b). Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic. *Spat. Stat.*

[166] Pollington, Timothy M (2020). Pull request: Functional changes for graphical hypothesis testing and clustering range estimation. `github.com/HopkinsIDD/IDSpatialStats/pull/2`. Accessed: 2020-03-27.

[167] Porta, M. (2008). *A Dictionary of Epidemiology, Fifth Edition: Edited by Miquel Porta*. Oxford University Press, New York, fifth edition.

[168] R (2020). *R* v4.1.1: An Environment for Statistical Computing.

[169] R (2021). `R` v4.1.0-foss-2021a: A language and environment for statistical computing. `R-project.org`.

[170] Rahman, K. M., Islam, S., Rahman, M. W. M., Kenah, E., Galive, C. M., Zahid, M. M., Maguire, J., Haque, R., Luby, S. P., and Bern, C. (2010). Increasing incidence of post–kala-azar dermal leishmaniasis in a population-based study in Bangladesh. *Clin. Inf. Dis.*, 50:2002–2005.

[171] Ready, P. (2014). Epidemiology of visceral leishmaniasis. *Clin Epidemiol*, (6):147–154.

[172] Rehman, N. A., Salje, H., Kraemer, M. U., Subramanian, L., Saif, U., and Chunara, R. (2020). Quantifying the localized relationship between vector containment activities and dengue incidence in a real-world setting: A spatial and time series modelling analysis based on geo-located data from Pakistan. *PLoS Negl. Trop. Dis.*, 14(5):1–22.

[173] Rehman, N. A., Salje, H., Kraemer, M. U. G., Subramanian, L., Cauchemez, S., Saif, U., and Chunara, R. (2018). Quantifying the impact of dengue containment activities using high-resolution observational data. *bioRxiv*. Accessed: 2019-06-16.

[174] Reich, N. G., Perl, T. M., Cummings, D. A., and Lessler, J. (2011). Visualizing Clinical Evidence: Citation Networks for the Incubation Periods of Respiratory Viral Infections. *PLoS ONE*, 6(4).

[175] Revolution Analytics and Weston, S. (2020). `iterators` *R* package v1.0.13: Provides iterator construct. `CRAN.R-project.org/package=iterators`.

[176] Rock, K. S., Quinnell, R. J., Medley, G. F., and Courtenay, O. (2016). *Progress in the Mathematical Modelling of Visceral Leishmaniasis*. Elsevier.

[177] RStudio (2019). IDE for *R* v1.4.17.17. `rstudio.com`.

[178] Salje, H., Cauchemez, S., Alera, M. T., Rodriguez-Barraquer, I., Thaisomboonsuk, B., Srikiatkhachorn, A., Lago, C. B., Villa, D., Klungthong, C., Tac-An, I. A., Fernandez, S., Velasco, J. M., Roque Vito G., J., Nisalak, A., Macareo, L. R., Levy, J. W., Cummings, D., and Yoon, I.-K. (2016a). Reconstruction of 60 Years of Chikungunya Epidemiology in the Philippines Demonstrates Episodic and Focal Transmission. *J. Inf. Dis.*, 213(4):604–610.

[179] Salje, H., Cummings, D. A., and Lessler, J. (2016b). Estimating infectious disease transmission distances using the overall distribution of cases. *Epidemics*, 17:10–18.

[180] Salje, H., Cummings, D. A. T., Rodriguez-Barraquer, I., Katzelnick, L. C., Lessler, J., Klungthong, C., Thaisomboonsuk, B., Nisalak, A., Weg, A., Ellison, D., Macareo, L., Yoon, I.-K., Jarman, R., Thomas, S., Rothman, A. L., Endy, T., and Cauchemez, S. (2018). Reconstruction of antibody dynamics and infection histories to evaluate dengue risk. *Nature*, 557(7707):719–723.

[181] Salje, H., Lessler, J., Berry, I. M., Melendrez, M. C., Endy, T., Kalayanarooj, S., A-Nuegoonpipat, A., Chanama, S., Sangkijporn, S., Klungthong, C., Thaisomboonsuk, B., Nisalak, A., Gibbons, R. V., Iamsirithaworn, S., Macareo, L. R., Yoon, I.-K., Sangarsang, A., Jarman, R. G., and Cummings, D. A. (2017). Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science*, 355(6331):1302–1306.

[182] Salje, H., Lessler, J., Endy, T. P., Curriero, F. C., Gibbons, R. V., Nisalak, A., Nimmannitya, S., Kalayanarooj, S., Jarman, R. G., Thomas, S. J., Burke, D. S., and Cummings, D. A. T. (2012). Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. *PNAS*, 109(24):9535–9538.

[183] Salje, H., Lessler, J., Paul, K. K., Azman, A. S., Rahman, M. W., Rahman, M., Cummings, D., Gurley, E. S., and Cauchemez, S. (2016c). How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. *PNAS*, 113(47):13420–13425.

[184] Seto, E., Knapp, F., Zhong, B., and Yang, C. (2007). The use of a vest equipped with a global positioning system to assess water-contact patterns associated with schistosomiasis. *Geospat. Health.*, 1(2).

[185] Simpson, G. and Mayer-Hasselwander (1986). Bootstrap sampling: applications in gamma-ray astronomy. *Astron. Astrophys.*, pages 340–348.

[186] Sinha, P., Bimal, S., Pandey, K., Singh, S., Ranjan, A., Kumar, N., Lal, C., Barman, S., Verma, R., Jeyakumar, A., Das, P., Bhattacharya, M., Sur, D., and Bhattacharya, S. (2008). A community-based, comparative evaluation of direct agglutination and rK39 strip tests in the early detection of subclinical Leishmania donovani infection. *Ann Trop Med Parasitol*, 2(102):119–125.

[187] Smith, C. M., Le Comber, S. C., Fry, H., Bull, M., Leach, S., and Hayward, A. C. (2015). Spatial methods for infectious disease outbreak investigations: Systematic literature review. *Eurosurveillance*, 20(39):1–21.

[188] Srivastava, P., Gidwani, K., Picado, A., Van der Auwera, G., Tiwary, P., Ostyn, B., Dujardin, J., Boelaert, M., and Sundar, S. (2013). Molecular and serological markers of Leishmania donovani infection in healthy individuals from endemic areas of Bihar, India. *Trop Med Int Health.*

[189] Succo, T., Noël, H., Nikolay, B., Maquart, M., Cochet, A., Leparc-Goffart, I., Catelinois, O., Salje, H., Pelat, C., de Crouy-Chanel, P., de Valk, H., Cauchemez, S., and Rousseau, C. (2018). Dengue serosurvey after a 2-month long outbreak in Nîmes, France, 2015: was there more than met the eye? *Eurosurveillance*, 23(23).

[190] Sundar, S. (2019). Transmission dynamics of visceral leishmaniasis in India: Role of asymptomatically infected individuals in American Society of Tropical Medicine and Hygiene: 68th Annual Meeting Abstract Book. astmh.org/ASTMH/media/2019-Annual-Meeting/ASTMH-2019-Abstract-Book.pdf.

[191] Tango, T. (1999). Comparison of General Tests for Spatial Clustering. In Lawson, A., Biggeri, A., Böhning, D., Emmanuel, L., Viel, J.-F., and Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 8, pages 111–117. John Wiley & Sons Ltd., Chichester, first edition.

[192] Tango, T. (2021). Spatial scan statistics can be dangerous. *Stat. Methods Med. Res.*, 30(1):75–86.

[193] Topno, R., Das, V., Ranjan, A., Pandey, K., Singh, D., Kumar, N., Siddiqui, N., Singh, V., Kesari, S., Kumar, N., Bimal, S., Kumar, A., Meena, C., Kumar, R., and Das, P. (2010). Asymptomatic infection with visceral leishmaniasis in a disease-endemic area in Bihar, India. *Am J Trop Med Hyg*, pages 502–506.

[194] Truelove, S. A., Graham, M., Moss, W. J., Metcalf, C. J. E., Ferrari, M. J., and Lessler, J. (2019). Characterizing the impact of spatial clustering of susceptibility for measles elimination. *Vaccine*, 37(5):732–741.

[195] Tsybakov, A. (2009). *Introduction to nonparametric estimation.* Springer.

[196] Vink, M. A., Bootsma, M. C. J., and Wallinga, J. (2014). Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis. *Am. J. Epidemiol.*, 180(9):865–875.

[197] Ward, M. P. (2007). Spatio-temporal analysis of infectious disease outbreaks in veterinary medicine: clusters, hotspots and foci. *Vet. Ital.*, 43(3):559–70.

[198] Weil, A. A., Khan, A. I., Chowdhury, F., LaRocque, R. C., Faruque, A. S. G., Ryan, E. T., Calderwood, S. B., Qadri, F., and Harris, J. B. (2009). Clinical Outcomes in Household Contacts of Patients with Cholera in Bangladesh. *Clin. Inf. Dis.*, 49(10):1473–1479.

[199] WHO (2010). Control of the Leishmaniases. Technical report, WHO, Geneva.

[200] WHO (2010). Monitoring and evaluation tool kit for indoor residual spraying. who.int/tdr/publications/documents/irs_toolkit.pdf, Accessed: 2018-10-09.

[201] WHO (2012). Accelerating work to overcome the global impact of Neglected Tropical Diseases - A roadmap for implementation. who.int/neglected_diseases/NTD_RoadMap_2012_Fullversion.pdf. Accessed: 2019-02-01.

[202] WHO (2015). Kala-azar elimination programme - report of a WHO consultation of partners. apps.who.int/iris/bitstream/handle/10665/185042/9789241509497_eng.pdf. Accessed: 2019-02-01.

[203] WHO (2017a). Leishmaniasis country profile - 2015, Bangladesh. irycis.org/media/upload/arxius/country_profiles/LEISHMANIASIS_CP_BGD_2015.pdf, Accessed: 2022-07-03.

[204] WHO (2017b). Leishmaniasis country profile - 2015, India. irycis.org/media/upload/arxius/country_profiles/LEISHMANIASIS_CP_IND_2015.pdf, Accessed: 2022-07-03.

[205] WHO (2017c). Status of endemicity of visceral leishmaniasis worldwide, 2015. who.int/leishmaniasis/burden/Status_of_endemicity_of_VL_worldwide_2015_with_imported_cases.pdf. Accessed: 2018-03-15.

[206] WHO (2020a). Ending the Neglect to Attain the Sustainable Development Goals: A Road Map for Neglected Tropical Diseases 2021–2030. who.int/neglected_diseases/resources/who-ucn-ntd-2020.01/en. Accessed: 2020-10-30.

[207] WHO (2020b). Independent Assessment Of Kala-azar Elimination Programme India. Technical report, WHO SEARO, New Delhi.

[208] WHO (2022). Status of endemicity of visceral leishmaniasis worldwide, 2015. who.int/news-room/fact-sheets/detail/leishmaniasis. Accessed: 2022-07-09.

[209] Wickham, H. (2007). Reshaping data with the `reshape` *R* package v1.4.4. *J. Stat. Softw.*, 21(12):1–20.

[210] Wickham, H. (2011). `testthat`: Get started with testing. *R J.*, 3:5–10.

[211] Wickham, H. (2016). *`ggplot2` R package v3.3.5: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[212] Wickham, H. (2018). `scales` *R* package v1.1.1: Scale functions for visualization. CRAN.R-project.org/package=scales.

[213] Wickham, H., Hester, J., and Chang, W. (2021). `devtools` *R* package v2.4.2. CRAN.R-project.org/package=devtools.

[214] Xie, Y. (2016). `bookdown` *R* package v0.2.4. bookdown.org/yihui/bookdown.

[215] Zijlstra, E. and Alvar, J. (2012). *The Post Kala-azar Dermal Leishmaniasis (PKDL): Atlas A Manual for Health Workers*. Geneva, Switzerland.

APPENDIX A

# Literature review of the ‘tau’ clustering statistic & proposition of a new rate estimator (Chapter 2)

## A.1 Key characteristics of the reviewed papers

| **Reviewed papers. ROOT$^{(1)}$/ REFORMING$^{(2)}$ paper cited?** | **Human disease & case defn.** | **Location rural$^r$/ urban$^u$** | **Statistic (● ) & epidemiological unit (○ )** | **Purpose & stated findings** | **Study type[1] & scale** | **№ cases, events, people or deaths** | **Sampling method** |
|---|---|---|---|---|---|---|---|
| **ROOT PAPER** (:=$^{(1)}$) Salje et al. [182]<br>2012 | Dengue<br>RT-PCR[2] (incl. serology) | Bangkok$^u$, TH | ● $\tau$(prevalence, distance) (case only)[3]<br>● $\phi$(distance & time)[4]<br>○ case with serotype, admission date, address, time ($t_j - t_i \leq$[1–3mo], or $\in$[3,4–30mo]$\Rightarrow z_{ij} = 1$) | First defined the $\tau$ (& $\phi$) statistics. Spatial clustering of same-serotype cases within 1km. | TS<br>5yr<br>~1,569km$^2$ | 1,912 geocoded | hospital, children |
| Grabowski et al.$^{(1)}$ [83]<br>2014 | HIV[5]<br>confirmed by serology/western blot | Rakai district$^r$, UG | ● $\tau$(prevalence, distance)<br>○ case/non-case pair, hhld. GPS[6], serostatus (every 12–18mo) | Spatial clustering of seropositive individuals from the hhld. level up to 250m but not at the community level | C<br>19mo<br>~3,352km$^2$ | 14,594 people (70% of censused popn.), 8,899 hhlds. 8,156/8,899 geocoded hhlds., 12·2% HIV seroprevalence & incidence 1·2/100pyrs | community, 15–49yrs |
| Bhoomiboonchoo et al.$^{(1)}$ [20]<br>2014 | Dengue<br>confirmed by RT-PCR & IgM/IgG[7] serology | Kamphaeng Phet province$^r$, TH | ● $\phi$(distance)<br>○ cases, village-level GPS, time ($t_j - t_i \leq$30d$\Rightarrow z_{ij} = 1$) | Spatiotemporal clustering of cases within 1mo & living in the same village. | TS<br>14yr<br>~8,608km$^2$ | 4,768 (93% of all cases) | hospital, from villages with $\geq$ 40 cases |

[2] Reverse-Transcription Polymerase Chain Reaction

[3] Tau statistic, see §2.3 for detailed information on estimators

[4] Phi statistic measures spatiotemporal interaction [182]

[5] Human Immunodeficiency Virus

[6] Global Positioning System

[7] Immunoglobulin M & G antibodies

| **Reviewed papers. ROOT$^{(1)}$/ REFORMING$^{(2)}$ paper cited?** | **Human disease & case defn.** | **Location rural$^{r}$/ urban$^{u}$** | **Statistic (● ) & epidemiological unit (○ )** | **Purpose & stated findings** | **Study type[1] & scale** | **№ cases, events, people or deaths** | **Sampling method** |
|---|---|---|---|---|---|---|---|
| Levy et al.$^{(1)}$ [119]<br>2015 | URI/ILI[8]/ influenza confirmed by influenza RT-PCR & multiplex PCR | Military barracks$^{r}$, TH | ● risk ratio (events, distance)[9]<br>○ case events, bed location, presentation time ($t_j - t_i \leq$1w$\Rightarrow z_{ij} = 1$) | Non-significant clustering of cases up to 5m. | C<br>11w<br>1 sleeping quarter | 77 ILI/URI events, 122 recruits | 20–31yr male recruits. Pre-existing TB or immunosuppression excl. |
| Salje et al.$^{(1)}$ [178]<br>2016 | Chikungunya confirmed by febrile + RT-PCR | Cebu City$^{u}$, PH | ● risk ratio (fixed distance window)[9]<br>○ seroconversion event (DENV[10] 1–4)(12mo apart), hhld. location | Spatial dependence of seroconversion $\leq$ 230m—rationale for focal interventions. | C<br>1yr<br>$\sim$315km$^2$ | $\sim$106 seroconversions of 851 people | community, randomly sampled, $\geq$6mo age, only one selected per hhld. |
| **REFORMING PAPER** (:=$^{(2)}$) Lessler et al.$^{(1)}$ [118]<br>2016 | Dengue, HIV, measles | Data re-use [83, 182] & Hagelloch$^{r}$, DE | . . .[11]/. . . /<br>● $\tau$(prevalence, distance)<br>○ onset date ($t_j - t_i \leq$2w$\Rightarrow z_{ij} = 1$) | Reformed the use of $\tau$ w.r.t. formulae and 'case & non-case' data. | . . . /. . . /<br>3mo<br>$\sim$0·06km$^2$ | . . . /. . . /<br>188 | . . . /. . . /<br>community, children from case homes |
| Salje et al.$^{(1,2)}$ [183]<br>2016 | Chikungunya $\sim$48% confirmed by IgM serology | Palpara$^{r}$, BG | ● $\tau$(prevalence, distance)<br>○ case/non-case pair, onset date (variable generation time, mean 14d), hhld. GPS | Used to test the sensitivity of global clustering by different transmission kernel sizes of a simulated epidemic. | XS<br>6mo<br>$\sim$0·6km$^2$ | 1,933 individuals, 460 hhlds., 175 confirmed | community, every hhld. in outbreak village |
| Grantz et al.$^{(1)}$ [84]<br>2016 | Influenza/ pneumonia reported by Chicago D.o.H. | Chicago$^{u}$, US | ● $\phi$(distance)<br>○ case death pair, death date ($t_j - t_i \leq$1w$\Rightarrow z_{ij} = 1$) | Spatial clustering of mortality at the census-tract level. | TS<br>7w<br>$\sim$606km$^2$ | 7,971 deaths | community, routine data |
| Hoang Quoc et al.$^{(1,2)}$ [94]<br>2016 | Dengue confirmed by RT-PCR | Ho Chi Minh City$^{u}$, VN & [182] | ● $\tau$(odds, distance)<br>○ case pair, serology (DENV1–4), address, admission date($t_j - t_i = 0$)mo$\Rightarrow$ $z_{ij} = 1$), | Small-scale spatial clustering of cases $<$ 500m | C<br>4yr<br>$\sim$2,061km$^2$/ . . . | 1,444 with serology & geolocated/. . . | hospital, imprecise geolocations dropped |

[8] Upper Respiratory Illness or Influenza-Like Illness

[9] Reported by authors as a tau statistic

[10] Dengue Virus

[11] ". . . /" = Re-use of data mentioned elsewhere in this Table, see disease or location featured in the second or third columns of this row.

| **Reviewed papers. ROOT$^{(1)}$/ REFORMING$^{(2)}$ paper cited?** | **Human disease & case defn.** | **Location rural$^{r}$/ urban$^{u}$** | **Statistic (● ) & epidemiological unit (○ )** | **Purpose & stated findings** | **Study type[1] & scale** | **№ cases, events, people or deaths** | **Sampling method** |
|---|---|---|---|---|---|---|---|
| Salje et al.$^{(1)}$ [181]<br>2017 | Dengue<br>confirmed by RT-PCR or IgM/IgG serology | TH$^{ru}$ | ● risk ratio (prevalence, distance)[12]<br>○ case pair or virus pair, admission date ($t_j - t_i$ ≤6mo⇒ $z_{ij} = 1$), serotype (DENV1–4), hhld. GPS, MRCA[13] date ($g_j - g_i$ ≤6mo, or ∈[6mo,2yr), [5,10yr) from sequencing data) | Virus pair spatiotemporal clustering ≤5km & 6mo of MRCA. | RC<br>16yr<br>∼513,120km$^2$ | 17,931<br>(= 640 + 17,291) | hospital,<br>children or young teenagers where serotype is known |
| Finger et al.$^{(1,2)}$ [70]<br>2018 | Cholera<br>acute watery diarrhoea + any age | N'Djamena$^{u}$, TD | ● $\tau$(odds, distance). Distance windows were constrained by "spatial discretisation of the model domain".<br>○ case pair, hhld. GPS, onset date ($t_j - t_i$ ≤5d⇒ $z_{ij} = 1$ | $\tau$ calibrated a simulation model (in equal parts with a spatially explicit individual-based stochastic model) to test different intervention scenarios. | TS<br>7mo<br>166km$^2$ | 1,585 geolocated (of 4,352) | hospital,<br>∼ 1/2 cases geolocated (confirmed by home visit) |
| Salje et al.$^{(1,2)}$ [180]<br>2018 | Dengue<br>virus isolation + serological evidence | Kamphaeng Phet province$^{r}$, TH | ● $\tau$(odds, time)<br>● odds ratio (place, fixed time windows)[9]<br>○ case pair, serotype (DENV1–4), school, augmented model infection time (assume symptomatics' median IP - 7d; undetecteds' infection-to-titre rise = 11d) | Model diagnostic on inferred undetected subclinical infections—augmented infections shared the temporal clustering (specific to serotype & place) as symptomatic infections. | C<br>5yr<br>∼98km$^2$ | 3,451<br>with fever symptoms | school<br>8–11yr age, blood sampled every 3mo, excl. if migration plans within 12mo or thalassaemia. |
| Succo et al.$^{(1)}$ [189]<br>2018 | Dengue<br>anti-DENV IgM & IgG +ve + febrile + body temp ≥38$^{o}$C + not another condition | Nîmes$^{u}$, FR | ● odds ratio (fixed distance window)[14]<br>○ case/non-case pair, hhld. GPS, hhld. ID (to differentiate same bldg. but different hhld.) | Spatial clustering of case vs. non-case pairs detected at the hhld. level but no further. | XS<br>15d<br>∼0·6km$^2$ | 1,431 people, 512 hhlds., prev. 0·4% | community, residing ≥ 4mo, ≥ 2yr age |
| Rehman et al.$^{(1,2)}$ [172]<br>2018 | Dengue<br>confirmed case | Rawalpindi$^{u}$ & Lahore$^{u}$, PK | ● $\phi$(distance & time)[9]<br>○ case, hhld. GPS, onset date ($t_j - t_i$ ≤30d⇒ $z_{ij} = 1$ | $\phi$ statistic compares interaction of cases in a matched intervention/control study design. | TS<br>4 & 6yr<br>259km$^2$ & 1,772km$^2$ | 7,890 & 2,998 | community & hospital |

[12] The authors also analysed the spatial relationship of proportions of case pairs falling ill within 6mo & coming from the same transmission chain at different distances. However, as a proportion ranges 0–1, it is not included here as it is not comparable with the positive real $\tau$.

[13] Most Recent Common Ancestor

[14] Reported as a relative risk in their main text, but as an odds ratio in their supplementary material

| **Reviewed papers. ROOT(1)/ REFORMING(2) paper cited?** | **Human disease & case defn.** | **Location rural$^r$/ urban$^u$** | **Statistic (● ) & epidemiological unit (○ )** | **Purpose & stated findings** | **Study type[1] & scale** | **№ cases, events, people or deaths** | **Sampling method** |
|---|---|---|---|---|---|---|---|
| Azman et al.(1,2) [8]<br>2018 | Cholera<br>acute watery diarrhoea + any age | [70] & Kalemie$^u$, CD | • $\tau$(odds, distance)<br>• $\tau$(odds, time)<br>○ case, hhld. GPS, presentation date ($t_j - t_i$ ∈[0,4d],[1,4d],[0,5d],. . . ,[25d,30d]⇒ $z_{ij} = 1$ | Rationale for targeted intervention:≤ 100m,≤ 1w of index case presenting | . . . /<br>TS<br>12mo<br>∼64km$^2$ | 1,692/4,359<br>&<br>1,077/1,146 (geolocated/all) | . . . /<br>hospital,<br>all cases geolocated |
| Truelove et al.(1,2) [194]<br>2019 | Measles | TZ$^{ru}$ | • risk ratio (prevalence of vacc. status, distance), sample-weighted for clusters[9]<br>○ time-relatedness is swapped for vacc. status<br>unvacc. proportions of DHS[15] clusters, DHS cluster GPS, cluster sampling weights, numbers per cluster | Calibration tool to produce a synthetic population with a clustering of unvacc. that matched the empirical value from DHS surveys. | S<br>?yr<br>900km$^2$ | 100,000 individuals | community, residences randomly distributed in 30x30km$^2$, vacc. status clustered by random swapping algorithm until empirical $\tau$ reached. |

[15] Demographic Health Survey

APPENDIX B

# Developments in statistical inference when assessing spatiotemporal disease clustering with the tau statistic (Chapter 3)

## B.1 Extended notes on *Methods* section (§3.2)

### B.1.1 Computation methods

The `spatstat` library [11] was used for useful spatial functions, `purrr` for resampling [88], `fields` for image plots [143] and `latex2exp` & `scales` for graph notation [129, 212] and the code of ‘January’ (2017) for figure labelling. The `IDSpatialStats::get.tau()` and `get.tau.bootstrap()` functions were optimised by re-implementing them in $C$, which sped up $\tau_{\text{odds}}$ calculations by ∼29 times [159]. For consistency Lessler et al.’s overlapping distance band set was used throughout, *i.e.* $\underline{\Delta} = \big\{[0, 10), [0, 12), [0,14), \ldots, [0, 50), [2, 52), [4, 54), \ldots, [74, 124)\big\}$.

### B.1.2 Invalidation of the CI for the endpoint of spatiotemporal clustering

The CI for the endpoint of spatiotemporal clustering $\hat{D}$ is easily invalidated if not all $\hat{\tau}^*$ simulations intersect $\tau = 1$ within the distance band set $\underline{\Delta}$. Caution is needed as the simulations $\underline{D}$ on which the uncertainty in $\hat{D}$ is calculated, are not a random sample of the population of simulations $\underline{\hat{\tau}}^*$, which is a vital prerequisite for CI construction, as those that crossed $\tau = 1$ from above were selectively chosen and those that start at or below $\tau = 1$, or above it but never reached $\tau = 1$ ignored. Computing CIs at a 95% confidence level on any random sample with a small 5% dropout can substantially decrease the effective confidence level [80]. This selection bias is also $\underline{\Delta}$-dependent since if one chooses a large enough $\underline{\Delta}$, one may find that simulations that start above $\tau = 1$ eventually cross $\tau = 1$ and then contribute to the CI. Although this bias cannot be accounted for, the proportion of simulations used to construct the CIs are reported and the distance range extended as computation time permits to limit this bias.

### B.1.3 Estimating the startpoint of spatiotemporal inhibition

If inhibition was present at greater distances, estimating its range was ignored as it was irrelevant. However, if the reader wishes, it can be estimated using a similar algorithm as for estimating clustering at shorter distances, in which one instead captures simulation lines that *exit* the global envelope lower bound into $\tau < 1$ for increasing $d$.

### B.1.4 Advantages & caveats of MMPSB

The schema (Eqns. 3.3–3.7) is more robust than the original Loh & Stein method (Fig. B.1) when cases $i^*$ have no time-unrelated cases to pair with in their local distance band, *i.e.* $m_{i^*}(d_l, d_m, k = 0) = 0$ in Eqn. 3.2 causes infinite values for $\theta_{i*}(d_l, d_m)$, or `NaN` values when $m_{i^*}(d_l, d_m, k = 1) = 0$; the MMPSB simply characterises these null events as zeroes and their addition in Eqns. 3.4 & 3.5 separately protects the rest of the calculation. Alternative remedies that were attempted on Loh & Stein's approach such as dropping these contributions or merging contiguous distance bands proved fruitless—the envelope diverged greatly for short distances and was biased above for larger distances, and only 77·3% of simulations contributed to the CI compared to 100% for MMPSB (Fig. B.1). Dropping these inconvenient $i^*$ cases removes important spatial information which the tau bootstrap estimator is sensitive to in Eqn. 3.2.

This revised method solves the numerical challenges but is not exactly the Loh & Stein method as the tau estimate is indirectly obtained via calculation of the spatially bootstrapped odds $\theta^*$. Hence, it is unclear if the validation of their results automatically transfers to this modified form. It is also assumed that the mean of the bootstrap distribution of local mark functions asymptotically approximates the (global) tau statistic, as Loh & Stein only provided experimental evidence to support this [120, 121].

## B.2 Additional figure

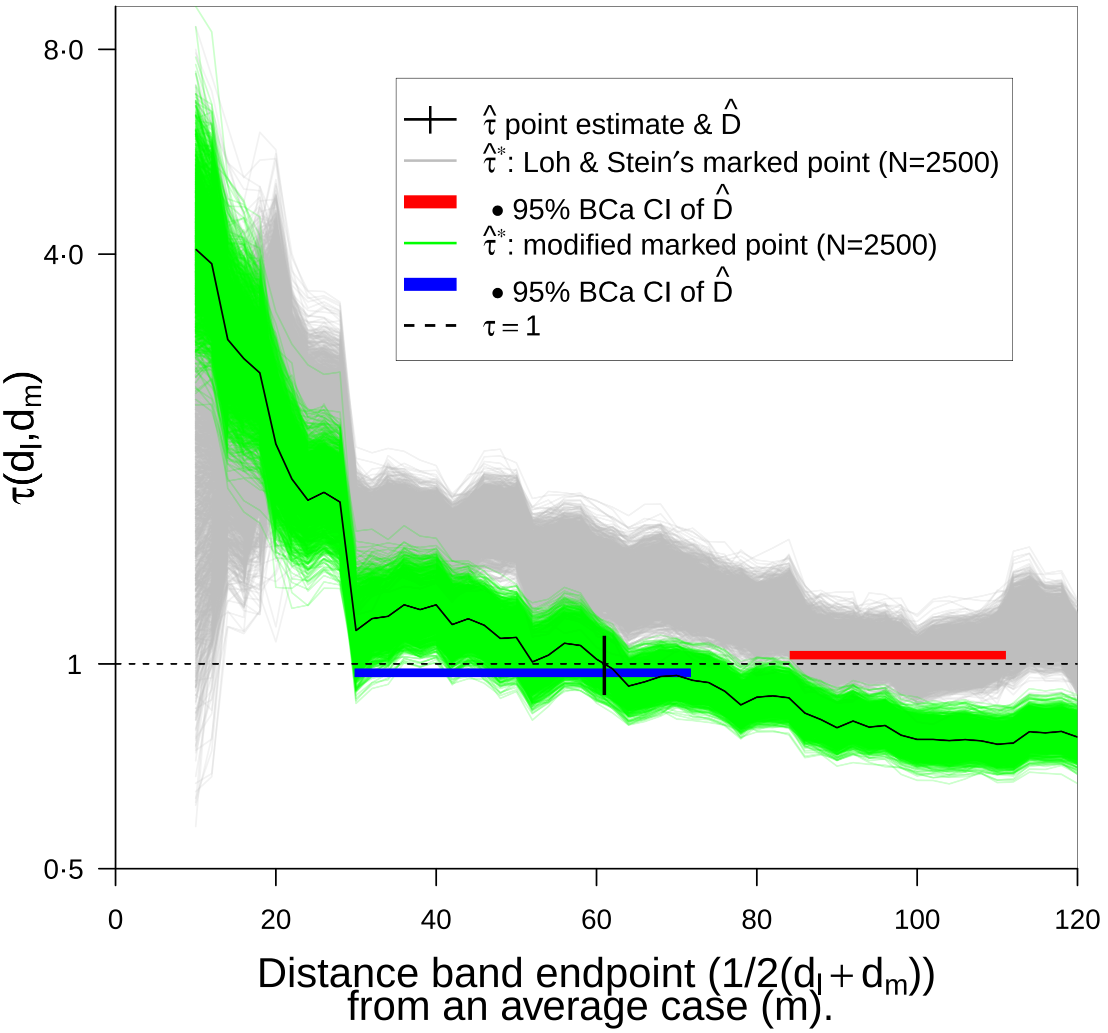


Figure B.1: **MMPSB sampling compared with the original Loh & Stein MPSB for the tau statistic**. The latter's envelope $\underline{\hat{\tau}}^*$ poorly covers $\hat{\tau}$ at short distances and leads to over-bias in $\hat{\tau}$ at large distances; note that only 77·3% of tau spatial bootstrap simulations $\underline{\hat{\tau}}^*$ contribute to the MPSB BCa CI compared to 100% for MMPSB. Distance band set as in Figure 3.6, $N$ =2500.

APPENDIX C

# Spatiotemporal clustering with variable exposure times: analysis using a new tau-rate estimator (Chapter 4)

## C.1 PRIME-NTD summary table: How the quality & relevance of modelling is communicated to stakeholders

| **Principle** | **What has been done to satisfy this?** | **Thesis location** |
|---|---|---|
| 1. Stakeholder engagement | These results shall be shared at the forthcoming WorldLEISH[7] Leishmaniasis conference. | n/a |
| 2. Complete model documentation | The calculation of the statistic is extensively described in the main chapter. Formulae are explicitly given with textual explanation.<br>Analysis code is available at: github.com/t-pollington/taurate | §2.3, §4.2.1 & §4.2.3 |
| 3. Complete description of data used | Data source & processing described. | §4.2.1 & §4.2.2 |
| 4. Communicating uncertainty | • Transparency in model assumptions<br>• Statistic's limitations extensively described | §4.2.1<br>§4.4 |
| | **Presenting hypothesis tests/ uncertainties**<br>• Number of Monte Carlo samples used in global envelope tests provided, as $p$-value conditional on this.<br>• 95% BCa CIs stated alongside parameters | In all respective figure captions |
| 5. Testable model outcomes | Analysis code is provided but unfortunately the data contains personal information so the analysis is not reproducible. There is not currently data nor forthcoming studies to validate these results. Surveillance in Bangladesh is far more limited during these lower endemicity times compared to during the 2002–2010 study. Although India may have more active control activities with its IRS programme, it is delivered over a standard radius & duration and so it is not possible to explore 'dose effect' at different radii that could test these estimates. | §4.2.2 |

Table C.1: **PRIME-NTD summary**. How the quality & relevance of modelling is communicated to stakeholders.

# APPENDIX D

# Impact of intensified control on VL: an ITSA (Chapter 5)

## D.1 Additional figure

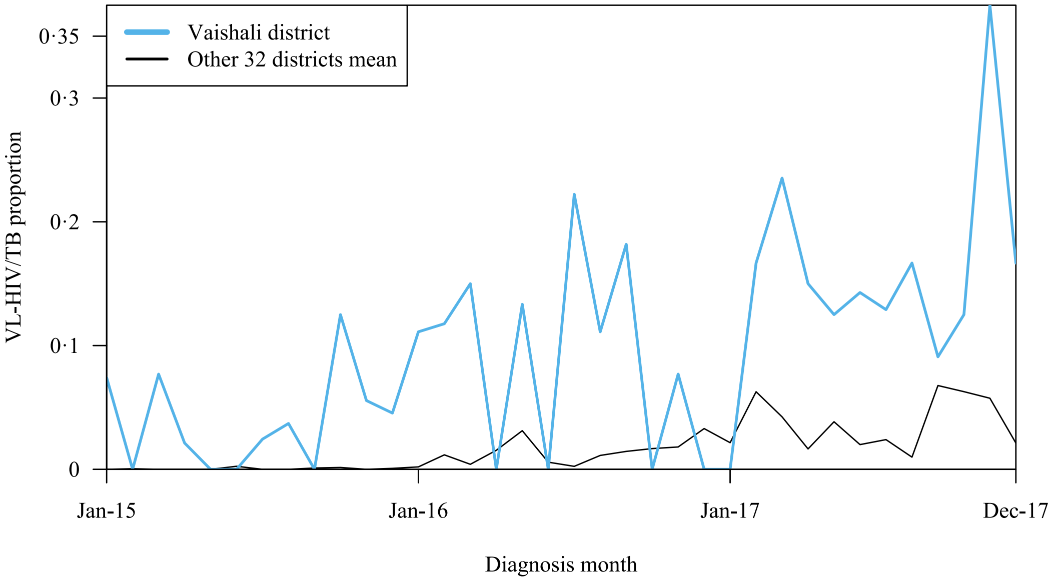


Figure D.1: **VL-HIV/TB case proportions out of all VL cases**. VL-HIV data from 2015–2017 and additional VL-HIV/TB data since 2017.

## D.2 Source of the data & PRIME-NTD summary

This routine surveillance data originated from the Kala-azar Notification Registry as part of the National Public Health Reporting System maintained by the Office of the Additional Director-cum-State Programme Officer, NVBDCP (Patna). The raw data was inputted electronically and checked for completeness, consistency & data entry errors. Any errors were resolved by the State Programme

Office & Nodal Officer of the NVBDCP. The cleaned data was also cross-validated with the NVBDCP's national data repository. This anonymised data aggregated by month & district (admin level 2) was shared with RMRIMS and so was non-personal & non-identifiable since age, sex & village location was not provided. New cases continued to be reported through the usual health system and collated by the NVBDCP; thus, this was a secondary data analysis.

| Principle | What has been done to satisfy this? | Thesis location |
|---|---|---|
| 1. Stakeholder engagement | These results will be shared at the forthcoming WorldLEISH[7] Leishmaniasis conference and with SpeakINDIA. | n/a |
| 2. Complete model documentation | The model structure is extensively described in the chapter.<br>Formulae are explicitly given with textual explanation. A compartmental model diagram was inappropriate for this statistical model.<br>Analysis code: github.com/t-pollington/ITSA | §5.2.3.2–5.2.4 & §5.2.3.1.1–5.2.4.1 |
| 3. Complete description of data used | Data source & processing described. | §5.2.1 & Appendix D.2 |
| 4. Communicating uncertainty | • Transparency in model assumptions | §5.2.3.2.1 |
| | • Model limitations extensively described | §5.4.1 & §5.5 |
| | **Presenting hypothesis tests/uncertainties** | |
| | • Number of Monte Carlo samples used in permutation test provided, as *p*-value conditional. | §5.2.3.2.1 |
| | • Standard error given for all model parameters. Combined parameters' point estimates provided (*e.g.* $\exp(\alpha^{(\lambda)}_{\text{other}} + \alpha^{(\lambda)}_{\text{Vaishali}})$ as it improves results interpretability for the reader) | Table 1 |
| | • Intervention effect: cases averted reported with uncertainty (IQR) | |
| | • CIs in parameter estimates & intervention effect. | |
| | • IQR of cases averted presented alongside point estimate (median). | Abstract |
| | **Sensitivity analysis** | |
| | • of model parameters to changes to the intervention start month | §5.2.3.2 & §5.3.4 |
| | • of OD distribution on outputs for RQ1 & RQ2 | §5.2.5 |
| | **Simulation uncertainty** | |
| | • Model comparison in models' ability to predict one-month-ahead was expressed as a *p*-value. | §5.2.3.2.1 |
| | • Fanplots: visually expressed simulation uncertainty around predicted values. | Fig. 5.14 |
| 5. Testable model outcomes | Code provided to show analysis is reproducible. Government data not shareable so reproducibility not demonstrable. Synthetic data unavailable. | §5.2.1 |
| | Requirements of *new* data listed to elucidate intervention success. | §5.1.4 |

Table D.1: **PRIME-NTD summary**. How the quality & relevance of modelling is communicated to stakeholders.