

The spatiotemporal tau statistic: a review

Timothy M Pollington^{a,c,*}, Michael J Tildesley^b, T Déirdre Hollingsworth^{c,**}, Lloyd AC Chapman^{d,**}

^a*MathSys CDT, University of Warwick, UK*

^b*Zeeman Institute (SBIDER), School of Life Sciences and Mathematics Institute, University of Warwick, UK*

^c*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK*

^d*London School of Hygiene & Tropical Medicine, UK*

Abstract

Introduction The *tau statistic* is a recent second-order correlation function that can assess the magnitude and range of global spatiotemporal clustering from epidemiological data containing geolocations of individual cases and, usually, disease onset times. This is the first review of its use, and the aspects of its computation and presentation that could affect inferences drawn and bias estimates of the statistic.

Methods Using Google Scholar we searched papers or preprints that cited the papers that first defined/reformed the statistic. We tabulated their key characteristics to understand the statistic’s development since 2012.

Results Only half of the 16 studies found were considered to be using true tau statistics, but their inclusion in the review still provided important insights into their analysis motivations. All papers that used graphical hypothesis testing and parameter estimation used incorrect methods. There is a lack of clarity over how to choose the time-relatedness interval to relate cases and the distance band set, that are both 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. A gap was noticed in the estimators available to account for variable person–time at risk.

*Corresponding author: MathSys CDT, University of Warwick CV4 7AL, UK

**Equal contributions from TDH & LACC

Email address: timothy.pollington@gmail.com (Timothy M Pollington^{id})

Discussion Our review comprehensively covers current uses of the tau statistic for descriptive analysis, graphical hypothesis testing, and parameter estimation of spatiotemporal clustering. We also define a new estimator of the tau statistic for disease rates. For the tau statistic there are still open questions on its implementation which we hope this review inspires others to research.

Keywords: dependence, second-order, spacetime clustering, transmission, relative risk, global statistic

Contents

1	Introduction	5
1.1	Current statistics & tests for global clustering	6
1.1.1	Spatial-only or spatiotemporal tests	6
2	The tau statistic	8
2.1	A brief history	8
2.2	Tau statistic τ_{odds} (odds ratio estimator)	9
2.3	Tau statistic τ_{prev} (relative prevalence estimator)	9
2.4	A new tau statistic τ_{rate} (rate ratio estimator)	10
2.4.1	Motivation	10
2.4.2	The estimator	10
2.5	Statistical characteristics	11
3	Methods	13
3.1	Search strategy, selection & data extraction	13
4	Results	15
4.0.1	Source information	15
4.0.2	Statistic purpose	15
4.0.3	Disease spectrum & study location	15
4.0.4	Study length, region size & spatial/temporal resolution	16
4.0.5	Tau statistic estimators & bootstrapping	16
4.0.6	Case definitions & misclassification	17
4.0.7	Graphical presentation	18
4.0.8	Distance band choice	19
4.0.9	Variables used to define relatedness	20
4.0.10	Defining time-relatedness using serial intervals	21
4.0.11	Testing spatiotemporal clustering & estimating its range	22
4.0.12	External validity	23
5	Discussion	24
5.1	Recommendations for further quantitative research	25
6	Acknowledgements & funding sources	28
7	Competing interests	28

8 Contributions: CRedit statement	28
9 Copyright	29
10 Figures	30
10.1 Key characteristics of the reviewed papers	31

1. Introduction

Transmission of infection is a dynamic process in time and space. Infectious diseases spread because a pathogen is transmitted by ‘contact’ with ‘parent’ cases (where we use ‘contact’ in a loose sense to include transmission from the 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 both close in time *and* space. The additional distinction of a *spatiotemporal* infection process is because normally any case will only be infectious for a short period relative to the study length thus leaving a temporal signal coinciding with their spatial presence. We focus on infectious diseases which require some explicit consideration of the infection process to formulate pair-relatedness variables. However, the risk factors for non-infectious diseases like poverty, environment, family/cultural traits may also follow spatiotemporal processes, like those between parent and offspring cases which we seek to measure for infectious disease. This demonstrates how non-infectious risk factors that coincide with infectious transmission may interfere with the spatiotemporal signal measured from an infectious disease dataset.

Knox defines *clustering* as “a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance” [1]. This could be ‘hotspot clustering’ which is “any area within the study region of significant elevated risk” [2], or *global clustering* which is a general tendency to cluster across a study. We avoid using the term ‘spatial/spatiotemporal dependence’ as it is not clearly defined in the literature.

Increasing availability of accurate geolocation data in recent years has enabled better understanding of this process. Being able to detect global disease clustering and assess its spatial extent can inform decisions on infectious disease control that can make better use of limited public health resources. However, standard clustering statistics in this domain often consider the spatial dimension only. (§1.1). Here we review publications using the *tau statistic* [3, 4]—a recent global clustering statistic for infectious disease (§2). In this literature review we explain its general purpose and identify sources of bias in the statistic. In an upcoming paper we consider how the different aspects of implementation identified here may bias the tau statistic [5].

1.1. Current statistics & tests for global clustering

This review focuses on the tau statistic [3, 4], but we first present other statistics for assessing disease clustering to highlight the benefits of the tau statistic to newcomers. Ward summarises spatiotemporal methods for disease data [6] by those that are based on mechanistic modelling, like spatiotemporal kernel models [7], and those 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 and sensitivities of the spatial distributions of the underlying population at risk [6]. Alternatively, empirical measures can estimate global clustering of individual cases (first-order) (§1.1) or case pairs (second-order) (§1.1 & §2). Second-order measures are particularly appropriate for investigating the infection process *between* individuals since we typically assume that infection occurs from one parent case infecting one susceptible offspring.

1.1.1. Spatial-only or spatiotemporal tests

Cuzick & Edwards’ k -nearest neighbours test [8], Anderson & Titterton’s Integrated Squared Difference function [9] and Tango’s C [10] are tests for clustering that divide the data into cases and 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 and $m_2(\mathbf{x})$ for sampled controls” [10] randomly chosen from “individuals at risk in the study region” [10].

Cuzick & Edwards’ k -nearest neighbours test sums the number of case-case pairings within a certain range [8], which has similarities to the tau statistic.

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

$$a_{ij} = \mathbf{1}(\mathbf{x}_j \in k\text{-nearest neighbours of } \mathbf{x}_i)$$
(1)

Anderson & Titterton’s Integrated Squared Difference function ($\widehat{\text{ISD}}$) smooths the difference of non-parametric kernel density-estimated relative risks in cases and controls (\hat{m}_1, \hat{m}_2) at point \mathbf{x} across 2D space S [9, 10].

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

Tango's C imposes a parametric kernel in the $\widehat{\text{ISD}}$ (Equation 2), e.g. a step function for hotspot clusters or exponential decay for clinal clusters [10].

Spatiotemporal K-function initially developed for stationary point processes [11], it has strong connections to the tau statistic as it is mentioned in appendix of the first paper to define and use the tau statistic [3], hereafter referred to as the *Root* paper. Epidemiologically, its stationarity will never adequately explain a disease process and a constant intensity does not take account of population heterogeneity. Gabriel & Diggle's *inhomogeneous K function* extended it using a special class of inhomogeneous point processes [12] and is available through the `stpp` R package [13]. It requires a spatial case intensity estimate via kernel-based density estimation and a temporal estimate from time-series modelling [12] so the calculation can be lengthy.

* * *

The tau statistic is defined in section 2 and the review into its use described in section 3. We identify best practices, determine aspects of its estimation where it may be biased and make recommendations in sections 4 & 5.

2. The tau statistic

2.1. A brief history

The *tau statistic*¹ is a non-parametric global clustering statistic which evaluates the disease frequency (risk, odds or rate) within a certain annulus around an average case and compares it to the background measure (at any distance), so is always positive [3, 4]. “It measures the tendency of case pairs to spatially cluster while implicitly accounting for their likeliness of being transmission-related temporally, making it a *spatiotemporal* statistic” [5, 3, 4]. Occasionally, space and time are swapped to measure temporal clustering instead with transmission relations based on spatial proximity.

The tau statistic was first defined and applied in 2012 by Salje et al. [3]. In 2016 Lessler et al. described its context in the fields of spatial statistics and epidemiology, demonstrated robustness, formulated estimators for case-only or case & non-case data, and reformed formulae in the *Tau* paper [4]. Both these *foundation* papers have inspired a steady stream of papers applying the tau statistic or similar statistics. The code to calculate both $\hat{\tau}_{\text{odds}}$, $\hat{\tau}_{\text{prev}}$ estimators is available in the `IDSpatialStats` R package [15]. Since datasets with thousands of cases can take tens of hours to construct confidence intervals for, we have re-implemented $\hat{\tau}_{\text{odds}}$ & $\hat{\tau}_{\text{prev}}$ in the C language, providing a speed-up of up to 76 times [16]. For a dataset of a few hundred cases the point estimate and bootstrapped tau estimates can typically be obtained in seconds.

There are some similarities between the tau statistic and earlier statistics which focused on areas of excess risk R : $R(\mathbf{x}) = \lambda(\mathbf{x})/g(\mathbf{x})$ where the numerator represented the case intensity at point \mathbf{x} in space S and denominator the “background effect” [17]. “Some information concerning the scale of clustering can also be obtained by this method” [18] on changing tolerance bounds to detect where clustering is strongest. The tau statistic’s functional form differs as the numerator’s distance band $[d_1, d_2)$ is nested within the denominator’s $(0, \infty)$ as we shall see in Equations 3-5.

¹This tau statistic is different from ‘Kendall’s tau statistic’ or ‘Kendall’s rank correlation coefficient’ which is a bivariate statistic for ordinal data [14].

2.2. Tau statistic τ_{odds} (odds ratio estimator)

The distance form of the tau statistic τ_{odds} is a ratio of the odds $\theta(d_1, d_2)$ of finding any case j which is related to any case i , within a half-closed² annulus $[d_1, d_2)$ around case i , versus the odds $\theta(0, \infty)$ of finding related cases over any distance separation ($d_{ij} \geq 0$) for N total cases

$$\hat{\tau}_{\text{odds}}(d_1, d_2) := \frac{\hat{\theta}(d_1, d_2)}{\hat{\theta}(0, \infty)} \quad (3)$$

$$\text{where } \hat{\theta}(d_1, d_2) = \frac{\sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbf{1}(z_{ij} = 1, d_1 \leq d_{ij} < d_2)}{\sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbf{1}(z_{ij} = 0, d_1 \leq d_{ij} < d_2)},$$

where $\mathbf{1}$ represents the indicator function, i.e. is equal to 1 when its argument(s) are all true and 0 otherwise. It is best described as “equivalent to ratios of multitype pair correlation functions” [4]. Values of $\tau > 1$ signify spatiotemporal clustering, $\tau = 1$ implies no clustering/inhibition and $0 < \tau < 1$ means inhibition. The odds $\hat{\theta}$ in Equation 3 is the ratio of the number of related case pairs within $[d_1, d_2)$ to the number of unrelated case pairs. The relatedness of a case pair z_{ij} is determined using temporal (close onset times t_i, t_j), serological (same serotypes) or genotype information (e.g. most recent common ancestor within a time difference of the earliest onset of the pair [19]) [4]. “Typically temporal relation is defined when case onset times are within a single serial interval of each other” [5]. This relatedness is a probable but not certain statement of direct transmission; still the tau statistic is able to recover a spatiotemporal signal in many of its applications. Sometimes an expanding disc is chosen so we set $d_1 = 0$, relabel $d = d_2$ and have $\tau(d)$ instead. τ_{odds} is similar to an odds ratio in that it is a ‘ratio of odds’ yet note how the numerator’s distance condition ($d_1 \leq d_{ij} < d_2$) is a subset of the denominator ($\forall d_{ij} \geq 0$), whereas traditionally an odds ratio is between two mutually exclusive conditions.

2.3. Tau statistic τ_{prev} (relative prevalence estimator)

With the additional data of non-case locations one can compute the prevalence $\hat{\pi}(d_1, d_2)$ of related case pairs within a certain annulus versus any case

²This corrects Lessler et al.’s appendices [4] that originally used an open interval. “It has been updated in their GitHub repository [15] following email communication on 6 December 2018” [5]

or non-case pairing, and thus the prevalence form of the tau statistic approximates a risk of onset [4]. The tau statistic then becomes the relative prevalence of related case pairs within an annulus versus at any distance from an average case i (Equation 4). Note that N now represents the number of cases and non-cases combined.

$$\hat{\tau}_{\text{prev}}(d_1, d_2) := \frac{\hat{\pi}(d_1, d_2)}{\hat{\pi}(0, \infty)} \quad (4)$$

$$\text{where } \hat{\pi}(d_1, d_2) = \frac{\sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbb{1}(z_{ij} = 1, d_1 \leq d_{ij} < d_2)}{\sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbb{1}(d_1 \leq d_{ij} < d_2)}$$

2.4. A new tau statistic τ_{rate} (rate ratio estimator)

2.4.1. Motivation

Closed populations are an unrealistic model of nature due to migration, births and deaths. For long studies it is inaccurate to consider all participants being exposed to infection risk for equal times. Epidemiologists take account of this varying time-at-risk through a rate statistic. Furthermore, for diseases that confer little immunity and so occur repeatedly in the same individual, the true burden of disease would be underestimated if only one case per person is counted. For example, in cholera epidemics in Bangladesh [20], O1 & O139 strains co-circulate yet infection from either does not confer cross-protection [21].

2.4.2. The estimator

A rate ratio estimator of the tau statistic τ_{rate} can be calculated for data consisting of cases and non-cases, with time-varying geolocations, study entry and exit times and onset and recovery times for each disease episode (Equation 5).

$$\tau_{\text{rate}}(d_1, d_2) := \frac{\lambda(d_1, d_2)}{\lambda(0, \infty)} \quad (5)$$

From first principles, the incidence rate λ is traditionally defined as the number of new events divided by the person-time-at-risk [22]. For this second-order statistic we counted not the onset of a case but the transmission event across pairs. Individual i is allowed to have zero to multiple disease episodes. Each single episode l for individual i will result in multiple probable pair episodes ($l \rightarrow m$), each of which are linked to the multiple episodes m of a particular individual j within $[d_1, d_2]$ of i and within a certain time difference ($t_m - t_l$). So for n_r people at risk with n total disease episodes during

the study period, λ is described by summing episodes in the numerator and person-time-at-risk in the denominator, unlike τ_{odds} & τ_{prev} which sum cases or cases & non-cases, respectively (Equation 6).

$$\lambda(d_1, d_2) = \frac{\sum_{l=1}^n \sum_{m=1, k_l \neq k_m}^n \mathbb{1}(z_{lm} = 1, d_1 \leq d_{lm} < d_2)}{\sum_{i=1}^{n_r} \sum_{j=1, j \neq i}^{n_r} \sum_{t=1}^{T_{\text{end}}} \mathbb{1}(Z_{ij}(t) = 1, d_1 \leq d_{ij}(t) < d_2)} \quad (6)$$

where $Z_{ij}(t) = \mathbb{1}([\text{inf. start}_i, \text{inf. end}_i] \cap [\text{susc. start}_j, \text{susc. end}_j] \cap [t]) \neq \{\emptyset\}$

and $k_l (= i), k_m (= j)$ denote the indices of the individual to which the episodes belong; the denominator describes the total pair time at risk and is the length of time each i, j pair could be potentially relatable in space and time; $[\text{inf. start}_i, \text{inf. end}_i]$ represents the infectious period of individual i while $[\text{susc. start}_j, \text{susc. end}_j]$ represents j 's period of susceptibility, given the immunising effects of previous infection (if appropriate) and time spent in different locations $\mathbf{x}_j(t)$ relative to other i 's (Fig. 1).

Of course for a self-immunising disease, a pair will only share one episode at most; τ_{rate} is still useful in this instance as different pair times at risk still need to be accounted for. Note that the alternate calculation of i 's person-time-at-risk due to j is not symmetric. This means τ_{rate} will take longer to compute than τ_{odds} or τ_{prev} as we cannot assume transmission pairs are undirected. A proof of concept for a real-world dataset with large migratory movements is needed to see if τ_{rate} provides a substantially different estimate to τ_{prev} or τ_{odds} in terms of $\tau(d)$ and the range of spatiotemporal clustering $[0, D]$.

2.5. Statistical characteristics

Lessler et al. have found the following through epidemic simulations [4]:

- The range of clustering is consistent whether computed from the relative prevalence τ_{prev} or odds ratio τ_{odds} estimators.
- Only one temporal, serotype or genotype relatedness metric is needed to infer related pairs. However 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 τ in this region.

- In addition to the K function that estimates the clustering range, the tau statistic gives the relative magnitude of disease risk, odds or rate versus the background; this could provide informative priors for later mechanistic Bayesian modelling.
- It is robust to population spatial heterogeneities: correctly identifies no clustering in a spatially-clustered population unlike the pair correlation function. It 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” [4].
- Diseases with an effective reproductive number (“the average number of people someone infected at time t can infect over their infectious lifespan” [23]) $R_e > 1$ will overestimate the clustering range while underestimating the magnitude of the τ in the true region of clustering.
- Edge corrections are unnecessary.

3. Methods

3.1. Search strategy, selection & data extraction

We collected publicly-available works on 10 January 2019 that have cited the Root or Tau paper including articles (full text or abstract), conference abstracts, books, preprints, theses and dissertations in any language. We used Google Scholar to find articles (referred to as set B) that cite set A (either the Root paper [3] or Tau paper [4]); excluding duplicates. We also looked for articles that cited set B, called set C: in case set C only referred to the closest paper of inspiration from set B rather than set A. We checked active forks from GitHub repositories of the `IDSpatialStats` package [24, 15]. We also searched google.com for webpages and blogs about the “tau statistic”, with disambiguation exclusions. We also announced our review to some of the previous paper authors (Salje, Lessler, Truelove & Cummings) to inquire about any work in their research groups which was soon due for submission. We only accepted those which actively used the statistic in their analyses; mere citations to mention a previous clustering result for that particular disease were disregarded.

The remaining works were then read fully and we contacted the papers’ corresponding authors to clarify missing information; furthermore, following manuscript submission to *arXiv*, we gave them a ‘right-to-reply’ on 1 December 2019 to our commentary on their papers. This is some of the metadata extracted to summarise and find similarities, and ensure reviewing consistency:

- Disease
- Format of work (article, preprint, report etc)
- Country & setting
- Study type (cohort, cross-sectional, etc.)
- Sampling method for the data
- Calculation method of the tau statistic
- How they presented results of the tau statistic in text & graphics

This review is restricted to the use of the tau statistic, as consistently described in the Root & Tau papers only. Although we found papers which claimed but did not in fact use the tau statistic, this is not a critique of their analyses. We still considered papers claiming to use the tau statistic because we wanted to review a broad spectrum of analyses based on the authors' belief that it was a tau analysis, which is still relevant to this review.

4. Results

4.0.1. Source information

For works already online prior to journal publication we recorded its later journal version in the bibliography. Google Scholar found 16 papers (including the Root & Tau papers) [3, 4, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 19] that claimed to use the tau statistic in their analyses (Table 3); 61 papers that mentioned the Root and 6 the Tau paper without using the statistic were ignored. There was no active code, webpages nor blogs about the statistic. All were peer-reviewed articles or reports except one recent preprint [36]; all peer-reviewed works were from respected journals with a minimum recent impact factor of 2.8. The Root paper was published in June 2012 and 15 separate works followed in 2014(2)³; 2015(1); 2016(5); 2017(1); 2018(5) & 2019(1) (Table 3:col. 1). The Tau paper published in May 2016 saw 10 papers follow it. There were seven that cited both the Root & Tau papers [31, 32, 33, 34, 35, 36, 30] and a further seven that cited the Root only [37, 19, 25, 26, 27, 28, 29]. All papers had multiple authors and always involved Salje and/or Lessler.

4.0.2. Statistic purpose

The main use was determining clustering and its strength but two novel alternatives were: calibrating an Approximate Bayesian Computation model by adding τ as summary statistic that captured global clustering [33]; and empirically as a stopping criterion once a random labelling algorithm had reached a certain global clustering threshold [31] (Table 3:col. 5).

4.0.3. Disease spectrum & study location

The papers covered seven human diseases—chikungunya, cholera, HIV, influenza/influenza-like illnesses/upper respiratory illnesses, measles, pneumonia, and dengue which made the most appearances(8) (Table 3:col. 2). Analyses cover all populated continents except South America and Oceania. The tau statistic or related statistics has been used in settings where the region is a substantial landmass [31] or where there were spatial restrictions nearby or through their populations due to rivers [32, 33, 34, 3, 36, 19, 28, 29], walkways [37] or major roads [27] (Table 3:col. 3).

³Italicised numbers in round brackets indicate the number of papers.

4.0.4. Study length, region size & spatial/temporal resolution

The tau statistic is most commonly used in cohort studies, which ranged from 3 months to 5 years with median 15.5 months (Table 3:col. 6). Two studies used cross-sectional data.

The spatial resolution of field studies was often constrained by GPS receivers i.e. $\sim 10\text{m}$ (Table 3:col. 6). When relying on patient’s reported street address to extract a geolocation, large spatial errors of 110m-1km were calculated when validated with a household visit [3, 34]. Furthermore cases may be aggregated at a higher spatial level because of gridded population data [33] or too few cases during the study period [29]. The temporal resolution in days, weeks or months was ultimately constrained by the reporting system. For tau papers which explicitly reported it, temporal resolution was as follows: cholera 1 day(2); dengue 1 day(1), 1 month(2); measles 1 day(1).

One paper used data with a temporal resolution similar to the length of the serial interval [34] (Table 2) which is not ideal: as it could miss additional transmission pairs ($i \rightarrow j$ and $j \rightarrow k$) as conceivably within the mean 15-17 day incubation period, a case i may infect j and it infect a secondary case k , yet at monthly resolution only $i \rightarrow k$ would be observed.

4.0.5. Tau statistic estimators & bootstrapping

We consider only eight of the 16 papers that claimed to use the tau statistic actually used the form defined in the foundation papers [3, 4] (Table 3:col. 4). Not all papers used the same estimators as $\tau_{\text{odds}}, \tau_{\text{prev}}$ defined in the founds ratio estimator for case-only data (Equation 3) was the most common because studies typically collect the geolocation of only cases; the distance form appears in three of the 16 papers reviewed [32, 33, 34] with the lesser-known time form in two [32, 35].

The prevalence estimator (Equation 4) appeared in three papers [3, 26, 30]. Despite odds ratios and 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 effectively equal [38].

The rate estimator τ_{rate} we defined in §5 is yet to be used. We found one application of a rate-style risk ratio that varied with distance [37]. The choice of a rate made sense as the epidemiological unit was respiratory illness events—something a person could have repeatedly. However they did not explicitly account for variable times at risk, presumably because they assumed that all participants stayed throughout the study.

The time form $\tau(t_1, t_2)$ swaps space and time and defines case pairs as

related if they are within a specific distance band around a case. It is then evaluated across a set of time bands. Using a panel plot for different distance windows helps map out a “dynamic risk zone” [32] akin to a simpler representation of the 2D spacetime tau colour map [3]. We still consider this a tau statistic as spatiotemporal information is retained, just presented in a different way.

The number of bootstrap samples chosen had a wide range: 100(2), 500(6), 1000(4), 10,000(1) or unknown(2).

4.0.6. Case definitions & misclassification

In most situations the case definitions were of a clinical standard beyond those typically employed for surveillance (Table 3:col. 4). For diseases with a particularly fast progression from onset-to-death like cholera, people may die before reaching the hospital thus causing the data to be left-censored. There may be misclassification if the case definition of say cholera (acute watery diarrhoea at any age [21]) shares signs or symptoms with other pathogens like E.coli, shigella etc. Consideration is needed if the planned control is expected to prevent these related pathogens too and thus overestimate the potential reduction in the magnitude of tau at close distances. Finally it is unclear how the statistic would perform for an unknown disease with a broadly defined case definition e.g. in the initial stages of an epidemic.

There is potential misclassification of probably-related pairs if other contact activities occur by a different process than described by case dwelling proximity [37]. For instance in Levy et al. (a study of respiratory illness in military recruits) considered bed location in sleeping quarters as the spatial unit that would describe the infection process whereas activities like attending a crowded mess hall could be opportunities for infection.

The example of Grabowski et al. [26] challenges the assumption that using a more recent marker of infection (i.e. incident rather than prevalent HIV) will better identify probable transmission pairs thus leading to a stronger tau signal. The likely explanation for prevalent-incident case pairings showing higher relative risk within the household than incident-incident pairs is likely due to the low per act risk of HIV-1 infection for heterosexual vaginal sex in a developing country setting (0.08% [39]) and given the relatively short study (18 months) it would have been more likely to receive reports of prevalent-incident case pairs than incident-incident.

4.0.7. Graphical presentation

- The general use of two continuous lines to represent the upper and lower parts of a series of pointwise confidence intervals is unhelpful to the untrained reader. Instead plotting each point estimate with its own confidence bands like Salje et al. [30] encourages the reader to consider each in turn. However since this is a common reader mistake, a default warning in the caption may be required too.
- Most tau papers use a τ -versus-distance graph (and one τ -versus-time [32]) to show the magnitude of τ varying with distance.
- The convention is to plot $\tau(d_1, d_2)$ at the midpoint of the distance band like in Lessler et al. [4] i.e. $d = \frac{1}{2}(d_1 + d_2)$, but may be misinterpreted if not explained in the caption. The ideal default would plot the end of the distance band instead, unless the graph is used for parameter estimation purposes and then the midpoint makes sense.
- For within-household transmission the spatial aspect of the infection process is no longer modelled as household members have no spatial freedom to move as their house is modelled as a point. It may therefore be misleading to plot a line joining $[d = 0, \tau(d = 0)] \leftrightarrow [d_{2\text{nd}}, \tau(d_{2\text{nd}})]$ unless the first distance band includes non-zero distances i.e. $d \geq 0$.
- Plotting the tau axis on a log scale can aid identification of the curve's structure. However a log scale for the distance axis may affect accurate $\hat{\tau}(d) = 0$ determination.
- All point estimates should include envelopes unless multiple point estimates are displayed.
- Some plot too many tau lines on the same graph [32, 28]. This is discouraged for more than three envelopes—aligned panel plots are an alternative.
- The graph should cover the full extent of both bounds of the confidence interval. 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 caption should note the tau estimator, envelope type, number of bootstrap samples [40] and definition for time relatedness: since the graph’s shape is dependent on these values.
- The Root paper ([3]:Fig. 3) offers an advanced 2D colour plot where each pixel represents the tau estimate for a given distance and 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 serial interval is approximate. As well as the spatiotemporal signal of primary transmission, it can reveal seasonality (through repeated regular patterns in the temporal axis) and the immunising effect of each serotype [3]. However like spatiotemporal variograms the number of pairs that are separated by long spatial or temporal lags reduces, requiring caution near the plot’s extremities.

4.0.8. Distance band choice

The distance from an average case i can be represented by a half-closed annulus with distance band $[d_1, d_2)$ or as an open disc $[d_1 = 0, d_2)$. The choice depends on the purpose of analysis. An annulus will give a more precise estimate closer to some ‘instantaneous’ τ , but conversely as narrower distance bands contain fewer pairs, τ will become more variable and lead to a τ -versus-distance graph that is spikier with an indiscernible trend. Alternatively an open disc conveys the cumulative risk up to a said distance d_2 for use by policymakers: it represents how fieldworkers operate i.e. up to a fixed distance from an index case $[0, d_2)$, rather than a complicated annulus shape. However open discs also smooth any intermediary spatiotemporal structure like village-to-village. Smoothing can be accentuated further by allowing distance bands to overlap as at least three papers do [4, 32, 31]⁴. Also as d_2 increases, annuli will cover more pairs so that the estimate’s variance changes with distance which is detrimental to the performance of global envelope tests [41] (a method that will soon be employed for graphical hypothesis testing in [5]). Setting bins with equal numbers of pairs may solve this.

The tau statistic may in theory be definable at a specific single distance lag and time lag from a case to describe an instantaneous relative measure of risk, however it could never be estimated for a real data set as we only

⁴we learned of this for [4] through their analysis code that they kindly shared with us

have a finite collection of points in spacetime, and apart from household transmission ($d = 0$), a given spacetime lag combination is likely to exist for one pair at most. We therefore have to settle for distance bands. This thought experiment demonstrates how the existence of a true tau statistic is not known unless a future statistical proof can show asymptotic convergence to a particular limit as the number of points tend to infinity or a distance band width tends to zero. It is also telling that even if we know the transmission tree the estimate is still dependent on the distance bands we choose. Minimisation of the mean squared error $((\hat{\tau} - \tau)^2 + \text{Var}(\hat{\tau}))$ is tempting but even the ‘true’ τ is dependent on the set of distance bands. In conclusion this highlights a problem for internal validity, as for the same dataset we can arrive at a non-unique tau estimate.

4.0.9. Variables used to define relatedness

‘Location & time’ are the common variables(5) used to identify probable-related transmission pairs (Tables 1 & 3:col. 4).

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

Table 1: Epidemiological variables used in the papers’ statistics to describe transmission-relatedness of pairs. *presentation, admission or onset time of the case. **most recent common ancestor.

Some authors purported use of a ‘tau statistic’ lacked a temporal element [31] thus reducing it to a spatial statistic. Either the papers used the phi statistic ϕ (a related statistic concerned with spatiotemporal interaction [3]) [29, 25, 36], or risk [37, 28, 19, 31] or odds ratios [27]. For instance, for $\pi(d)$ (the numerator of τ_{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 τ was the proportion of pairs within d rather than the proportion of sick pairs with d compared to all pairs. Similarly others make τ the ratio between seroconverted and all individuals [28, 19] or cases and non-cases [27] rather than between the risk/odds of finding a case within a distance versus at all distances.

Grabowski et al. [26] is a unique example of the tau statistic where no temporal, geno nor serotype information is needed to link pairs—through an implicit temporal relation. 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 between prevalent and incident cases can be formed. This may be a useful workaround if explicit onset data is unavailable for your study. All authors use case or virus pairs to represent the transmission chain, except Grantz et al. [25] 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 and access to healthcare. However practically deaths may be the only available variable from the initial assessment of an outbreak of an unknown cause.

4.0.10. Defining time-relatedness using serial intervals

The aim is to represent pairs involved in primary transmission i.e. a single, direct transmission event between parent case i and offspring case j so it has been common to choose a time-relatedness interval with length equal to a single serial interval (Table 3:col. 4). We compare the time intervals chosen against published serial intervals (Table 2). Conceptually there is sufficient reason to believe τ is sensitive to the choice of the $[t_i = T_1, t_j = T_2]$ interval, but if so then to what extent and how to find the ‘optimal’ $[T_1, T_2]$? It is not solely about maximising specificity as for some diseases, altering T_1, T_2 to minimise co-primary and secondary transmission, may result in nothing left of the primary transmission peak. Alternatively a poor choice could contaminate the primary transmission signal with other indirectly-related transmission chains like coprimaries or the primary cases k of j . Additionally we do not yet know how the effect of transmission contamination biases the true spatiotemporal signal of primary transmission.

It is common for studies to use a particular interval without reference to the source, except Azman et al.’s cholera study [32]. As a caution to future tau statistic users, the reliability of published incubation period parameters is poor e.g. a sample of respiratory viral infections found half did not cite the source [42]. It is not just the length of the interval that is of interest but the start and endpoints (T_1, T_2) too—papers commonly use $T_1 = 0$ and set T_2 to the mean serial interval (Table 2).

Azman et al. [32] are nuanced in their $[T_1, T_2]$ selection for interpretative purposes. Initially they chose $[0, 5\text{d}]$ —sensible as cholera can have an incu-

bation period as short as a few hours [21]. However they switch to $[1, 5d]$ to show the elevated risk in cases that they could avert i.e. it is unrealistic to be expected to respond to the reported onset of case i , to mitigate a same-day onset of j .

Disease	Serial interval chosen	Published source
Cholera	$[0, 5d](2)$	median 5d, range 1-11d [43, 44]
	$[0, 4d](1)$	
	$[1, 4d](1)$	
	$[0, 5d], \dots, [25d, 30d](1)$	
Dengue	Same month $[0, 0mo](1)$	mean 15-17d [45]
	$[1, 3mo](1)$	
	$[3, 4-30mo](1)$	
Measles	$[0, 2wk](1)$	mean 11.7d [46], 14.9d [47]

Table 2: Serial intervals featuring in reviewed articles (paper frequencies in round brackets) compared with values from published sources. Papers choosing variable times [30] or model-informed times [35] have been excluded.

4.0.11. Testing spatiotemporal clustering & estimating its range

It is common for authors to test the evidence against no spatiotemporal clustering using visual inspection of a τ -versus-distance graph, which is a *graphical hypothesis test*. As detailed in [5], all papers incorrectly estimated this range in two ways and incorrectly simultaneously established the significance of clustering:

- i) most construct bootstrapped estimates around the point estimate to form a *central envelope* with a particular upper and lower bound according to a series of pointwise confidence intervals; they chose the endpoint of the clustering range as where the lower bound of the central envelope touched $\tau = 1$.
- ii) one paper [3] randomly permuted the time marks t across all cases (with points (x, y, t)) to simulate a process with no spatiotemporal clustering. An envelope was constructed about these simulations that straddled $\tau = 1$ to form a *null envelope* to simulate $H_0 : \tau = 1$; where the point estimate touches the upper bound marks the endpoint. Again the upper and lower bounds are defined by a series of pointwise confidence intervals.

In [5], we propose corrections for the hypothesis test of no clustering $H_0 : (\tau = 1)$ using global envelope tests [48] and estimation of the range of clustering as the horizontal set of points where the bootstrapped simulations $\hat{\tau}^*$ intersect $\tau = 1$ intersection points where the bootstrapped simulation $\hat{\tau}^*(d) = 1$ (as kindly suggested by Peter Diggle in a Skype conversation on 22 October 2019). The latter also provides a measure of precision for the clustering range, unavailable under the existing methods.

In essence the methods of the reviewed papers are incorrect as they:

- mix graphical hypothesis testing (which can only give a binary answer of accept/reject no spatiotemporal clustering i.e. $H_0 : \tau = 1$) with parameter estimation. Nearly all authors determine the range when the lower bound of the confidence interval touches 95%. Azman et al. [32] takes 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 we do not have a theoretically-informed correction factor.
- Pointwise confidence intervals are common to describe the uncertainty in $\hat{\tau}$ however many authors [3, 26, 4, 34, 32] incorrectly use them for hypothesis testing to assert “statistically significant” [3, 26] results: it is incorrect to scan the graph and search at multiple points along d where the bound of τ or null envelope is first crossed and then declare that as the clustering endpoint. Since multiple pointwise CIs are compared with $\tau = 1$, during this inspection it amounts to a series of multiple hypothesis tests which inflates the chance a true null hypothesis is rejected (type I error).
- Their method also cannot estimate the uncertainty of the clustering range parameter D estimated.

All papers use two-sided confidence intervals. This is sensible as immunising effects could cause inhibition at close distances. Although for a well-studied disease with known localised clustering, there may be reason to choose a one-tailed test apriori.

4.0.12. External validity

The tau statistic has been well tested in a range of infectious diseases exploring person-to-person and vector transmitted diseases, with short to

medium serial intervals and different markers of case relatedness (Table 3:cols. 2 & 4). The study settings have ranged from urban, peri-urban to rural setting at different population densities (Table 3:col. 3).

The tau statistic has not yet been applied to diseases like leprosy or visceral leishmaniasis whose highly variable incubation periods would increase the uncertainty in the clustering range [21].

5. Discussion

Clustering is an important characteristic to shed light on infection dynamics and can inform disease control or academic study. The tau statistic has been applied for this purpose to disease datasets which contain the location of cases (and possibly non-cases) and some variables to link probable transmission pairs by temporal, serological or genotypic attributes.

This review surveyed a number of papers which claimed to analyse disease clustering using the tau statistic. We only considered half of the 16 papers reviewed to truly use the tau statistic as originally defined by the foundation papers [3, 4]; this was either because a temporal element was lacking or their formulae were better described as a risk/odds ratio or phi statistic. Lessler et al.’s analysis demonstrated robustness of the statistic Lessler et al. [4]. However, we caution readers that they cannot necessarily expect these same benefits or caveats to apply for statistics outside the τ_{odds} or τ_{risk} definition. Despite this, this review has been richer for their inclusion through learning about the authors’ analysis motivations. This review has also uncovered examples of good practice or ways in which the statistic has been redefined:

- defining the lower interval T_1 of the time-relatedness interval $[T_1, T_2]$ to equal the expected field response time to avoid overestimating the elevated risk that may be averted
- outcome variable of death rather than case of disease
- plotting a ‘distance lag’-vs-‘time lag’ 2D colour plot, where each pixel represents a tau value
- a time form of the tau statistic $\tau(t_1, t_2)$
- considering ‘pre-study’ (prevalent) cases and ‘during study’ (incident) cases as a proxy to define time-relatedness on if onset times are unavailable

Knowledge about the tau statistic has been concentrated in the medium of journal articles and further limited to papers written by authors of the foundation papers—Salje & Lessler. Yet this statistic could be very useful to infectious disease modellers, field epidemiologists and policy makers, particularly given its implementation in the freely available `IDSpatialStats` R package. To further boost adoption, we plan to provide `R Markdown` tutorials of the tau statistic on open access training hubs like RECONlearn.org.

We hope this review has given readers an appreciation of the tau statistic with caveats on its use. Like any statistic the skilled epidemiologist should still be aware of standard concepts like case definitions to avoid misclassification. On graphing the results we have mentioned a few standard practices that can present the data objectively to avoid misleading the reader. Depending on the graphical purpose of the τ -vs-distance graph, open annuli for control policy questions; or open discs to investigate fine spatiotemporal structure may be appropriate, respectively. Furthermore as a spatiotemporal statistic one must consider the data’s spatial resolution, temporal resolution relative to the mean serial interval, and if the space and time variables are likely to represent the actual infection process.

5.1. Recommendations for further quantitative research

Following this review we recommend the following aspects of the tau statistic’s implementation are further investigated to assess their bias:

- It is currently unclear how best to choose the distance band set to reduce both bias and variance in the tau statistic and whether equidistant or equi-number bins should compose them.
- If using time-relatedness to link cases then how to choose the interval $[T_1, T_2]$ given a known serial interval for the disease.
- Differences in health status or treatment-seeking/healthcare could change the disease latent period or infectious period, respectively. Would this require a reappraisal of the time-relatedness interval over the course of the study?
- The number of bootstrap samples ranges over 100 fold but the importance of this implementation parameter is unknown.

- Test the new rate ratio estimator τ_{rate} on a dataset containing geolocations of cases and non-cases and the times of onset and recovery of disease episodes. The chosen setting should be where individuals have variable times-at-risk of disease due to seasonal migration or staggered study entry/exit; ideally the disease would be one that confers little protection following exposure so that multiple episodes are observable. Assess differences in the estimator compared to τ_{prev} . Given the scarcity of existing good quality data of this kind, a study may to be prospectively designed.
- Investigate the use of the tau statistic as a (global) spatial summary statistic for Approximate Bayesian Computation. To what extent does the tau statistic help with computation accuracy and efficiency and how should it be weighted relative to other summary statistics used by the algorithm?
- Immunity from disease exposure had a large biasing effect on the estimation of the mean transmission distance of simulated epidemics [49]. It is therefore sensible to assess tau statistic performance for immunising (SIR-style)⁵ and non-immunising (SIS-style) diseases.
- Test the validity of the tau statistic for diseases with highly variable incubation periods
- Is the tau statistic prone like other spatial statistics to population shift bias over time?
- Although it has been shown to robustly extract a clustering signal for a 1% fraction of a simulated dataset [4], what is the minimum number of cases for the tau statistic to perform reliably?
- What is the theoretical formulation of the tau statistic and can mathematical analysis of its statistical properties bring us new insights?
- Considering that the tau statistic uses symmetric shapes like discs or annuli that assume an isotropic disease process, how would anisotropies

⁵SIR is a compartmental model describing the progression of individuals from Susceptible, through Infected, to Recovered states of a disease. SIS diseases oscillate between S and I as individuals do not gain protection following infection.

in transmission e.g. along road networks, land relief or wind affect tau estimates?

- Impact of spatial aggregation of misspecification versus actual infection location and seasonality, on the tau statistic.

Some of the above aspects will be covered in an upcoming quantitative study applied to a measles dataset [5].

* * *

Control programmes have already been informed by the tau statistic so applying improvements on its implementation and further research will safeguard future health decisions based on it.

6. Acknowledgements & funding sources

We kindly thank:

- Lessler & Salje who openly answered questions on their work. We would like to thank Truelove who replied to our questions by email.
- Peter Diggle for explaining proper methods for graphical hypothesis testing and methods for estimating the clustering range.

TMP, LACC & TDH gratefully acknowledge funding of the NTD Modelling Consortium by the Bill & Melinda Gates Foundation (BMGF) (grant N^o OPP1184344), and LACC acknowledges funding of the SPEAK India consortium by BMGF (grant N^o OPP1183986). Views, opinions, assumptions or any other information set out in this article should not be attributed to BMGF or any person connected with them.

TMP's PhD was supported by the Engineering & Physical Sciences Research Council, Medical Research Council and University of Warwick (grant N^o EP/L015374/1). TMP would like to thank Big Data Institute for hosting him during this review under the supervision of Déirdre Hollingsworth.

All funders had no role in the study design, collection, analysis, interpretation of data, writing of the report, or decision to submit the manuscript for publication.

7. Competing interests

All authors declare no competing interests.

8. Contributions: CRediT statement

TMP: Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft & editing, Visualisation
MJT: Conceptualisation, Writing - review & editing, Supervision
PJD: Methodology (see §6)
TDH: Conceptualisation, Writing - review & editing, Supervision, Funding acquisition
LACC: Conceptualisation, Writing - review & editing, Supervision.

9. Copyright

This article is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives Works 4.0 International Licence (CCBY-NC-ND4.0). Anyone can copy and distribute this article unchanged and unedited but only for non-commercial purposes, provided the user gives credit by providing this article's DOI and a link to the licence (creativecommons.org/licenses/by-nc-nd/4.0). The use of this material by others does not imply endorsement by the authors.

10. Figures

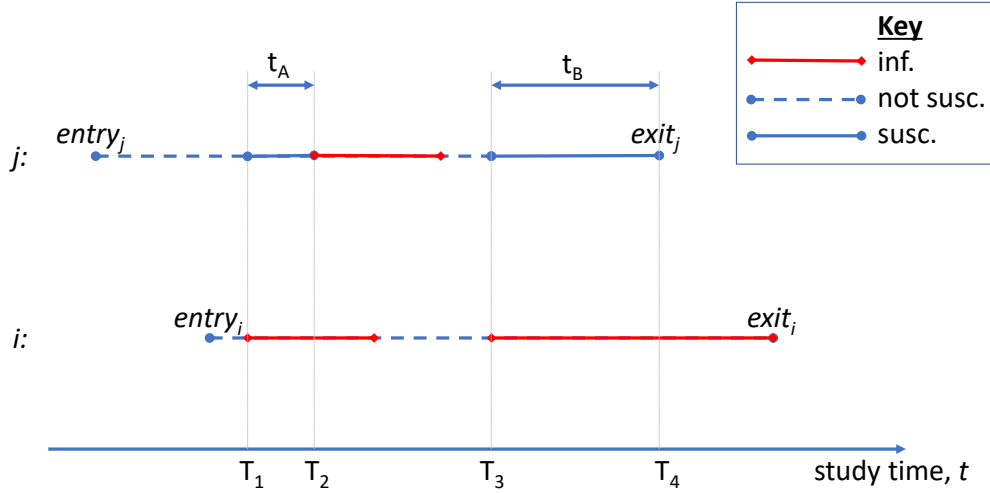


Figure 1: Describing the calculation of the person-time-at-risk ($t_A + t_B$) of individual j , due to i , by accounting for j 's location with respect to i , j 's susceptibility to infection (which for some diseases may depend on previous exposure) and i 's infectious period. j is born and enters the study and stays within $[d_1, d_2]$ of i 's eventual entry. j is considered to be exposed to risk of i during $[T_1, T_2]$ as j loses maternal immunity at $t = T_1$, thus becoming susceptible and is spatiotemporally related to i . At $t = T_2$, j becomes infected and is no longer susceptible. One potentially-related episode pair $i \rightarrow j$ is counted. j recovers but is exposed again to risk from i during $[T_3, T_4]$ when i becomes infectious again (in this example previous exposure does not confer protection). j leaves the study, and does not experience a second episode w.r.t. i (at least not when enrolled on the study).

10.1. Key characteristics of the reviewed papers

Reviewed papers. Root ^R /Tau ^T paper cited?	Human disease & case defn.	Location rural ^r /urban ^u	Statistic(●) & epidemiological unit(o)	Purpose & stated findings	Study type ⁶ & scale	N ^o cases, people or deaths	Sampling method
ROOT PAPER Salje et al. [3] 2012	Dengue RT-PCR ⁷ (incl. serology)	Bangkok ^u , TH	<ul style="list-style-type: none"> • τ (prevalence, distance) (case only)⁸ • ϕ (distance & time)⁹ o case with serotype, admission date, address, time ($t_j - t_i \leq 1-3\text{mo} $, or $\in [3,4-30\text{mo}] \Rightarrow z_{ij} = 1$)	Spatial clustering of same-serotype cases within 1 km.	TS 5yr $\sim 1,569\text{km}^2$	1,912 geocoded	hospital, children
Grabowski et al. ^R [26] 2014	HIV ¹⁰ confirmed by serology/western blot	Rakai district ^r , UG	<ul style="list-style-type: none"> • τ (prevalence, distance) o case/non-case pair, hhd GPS11, serostatus (every 12-18mo) 	Spatial clustering of seropositive individuals from the hhd. level up to 250m but not at the community level	C 19mo $\sim 3,352\text{km}^2$	14,594 people (70% of censused popn.), 8,899 hhlds. 8,156/8,899 geocoded hhlds., 12.2% HIV seroprevalence and incidence 1.2/100pyrs	community, 15-49yrs
Bhoombhoonchoo et al. ^R [29] 2014	Dengue confirmed RT-PCR and IgM/IgG ¹² serology	Kamphaeng Phet province ^r , TH	<ul style="list-style-type: none"> • ϕ (distance) o cases, village-level GPS, time ($t_j - t_i \leq 30\text{d} \Rightarrow z_{ij} = 1$) 	Spatiotemporal clustering of cases within 1 mo and living in the same village.	TS 14yr $\sim 8,608\text{km}^2$	4,768 (93% of all cases)	hospital, from villages with ≥ 40 cases

⁷Reverse-Transcription Polymerase Chain Reaction

⁸Tau statistic, see §2 for detailed information on estimators

⁹Phi statistic measures spatiotemporal interaction [3]

¹⁰Human Immunodeficiency Virus

¹¹Global Positioning System

¹²Immunoglobulin M & G antibodies

Reviewed papers, Root ^R /Tau ^T paper cited?	Human disease & case defn.	Location rural ^r /urban ^u	Statistic(●) & epidemiological unit(o)	Purpose & stated findings	Study type ⁶ & scale	N ^o cases, people or deaths	Sampling method
Levy et al. ^R [37] 2015	URI/ILI ¹³ /influenza confirmed by RT-PCR and multiplex PCR	Military barracks ^r , TH	<ul style="list-style-type: none"> risk ratio(events, distance)¹⁴ <ul style="list-style-type: none"> 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 11w 1 sleeping quarter	77 ILI/URI events, 122 recruits	20-31yr male recruits, Pre-existing TB or immunosuppression excl.
Salje et al. ^R [28] 2016	Chikungunya confirmed by febrile RT-PCR +	Cebu City ^u , PH	<ul style="list-style-type: none"> risk ratio(fixed distance window)¹⁴ <ul style="list-style-type: none"> seroconversion 1-4)(12mo apart),hhld. location 	Spatial dependence of seroconversion $\leq 230m$ — rationale for focal interventions.	C 1yr $\sim 315km^2$	~ 106 seroconversions of 851 people	community, randomly sampled, $\geq 6mo$ age, only one selected per hhld.
TAU PAPER Lessler et al. ^R [4] 2016	Dengue, measles, HIV,	Data re-use [3, 26] & Hageloch ^r , DE	<ul style="list-style-type: none"> τ(prevalence, distance)¹⁶/\dots/\dots onset date ($t_j - t_i \leq 2w \Rightarrow z_{ij} = 1$) 	Reformed the use of τ w.r.t. formulae and use of case and non-case data.	\dots / \dots / \dots / 3mo $\sim 0.06km^2$	\dots / \dots / 188	\dots / \dots / community, children from case homes
Salje et al. ^{RT} [30] 2016	Chikungunya $\sim 48\%$ confirmed by IgM serology	Palpara ^r , BG	<ul style="list-style-type: none"> τ(prevalence, distance) <ul style="list-style-type: none"> 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 6mo $\sim 0.6km^2$	1,933 individuals, 460 hhlds, 175 confirmed	community, every hhld in outbreak village
Grantz et al. ^R [25] 2016	Influenza/pneumonia reported by Chicago D.o.H.	Chicago ^u , US	<ul style="list-style-type: none"> ϕ(distance) <ul style="list-style-type: none"> 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 7w $\sim 606km^2$	7,971 deaths	community, routine data

¹³Upper Respiratory Illness or Influenza-Like Illness

¹⁴Reported by authors as a tau statistic

¹⁵Dengue Virus

¹⁶“ \dots ” = Re-use of data mentioned elsewhere in table, see disease or location featured in the second or third columns of this row.

Reviewed papers. Root ^R /Tau ^T paper cited?	Human disease & case defn.	Location rural ^r /urban ^u	Statistic(●) & epidemiological unit(o)	Purpose & stated findings	Study type ⁶ & scale	N ⁶ cases, events, people or deaths	Sampling method
Hoang Quoc et al. ^{RT} [34] 2016	Dengue confirmed by RT-PCR	Ho Chi Minh City ^u , VN & [3]	<ul style="list-style-type: none"> • τ (odds, distance) <ul style="list-style-type: none"> o 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 4yr ~2061km ² / ...	1,444 with serology & geolocated/...	hospital, precise geolocations & geolocated
Salje et al. ^R [19] 2017	Dengue confirmed by RT-PCR or IgM/IgG serology	TH ^{r,u}	<ul style="list-style-type: none"> • risk ratio(prevalence, distance)¹⁷ <ul style="list-style-type: none"> o case pair or virus pair, admission date($t_j - t_i \leq 6$mo$\Rightarrow z_{ij} = 1$), serotype(DENV1-4), hhld. GPS, MRCA18 date ($g_j - g_i \leq 6$mo, or $\in [6$mo,2yr), [5,10yr) from sequencing data) 	Virus pair spatiotemporal clustering ≤ 5 km & 6mo of MRCA.	RC 16yr ~513,120km ²	17,931 (=640+17,24)	hospital, children or young teenagers where serotype is known
Finger et al. ^{RT} [33] 2018	Cholera acute watery diarrhoea + any age	N'Djamena ^u TD	<ul style="list-style-type: none"> • τ (odds, distance). Dis-tance windows were constrained by "spatial discretisation of the model domain". <ul style="list-style-type: none"> o case pair, hhld. GPS, onset date($t_j - t_i \leq 5$d$\Rightarrow z_{ij} = 1$ 	τ calibrated a simulation model (in equal parts with a spatially explicit individual-based stochastic model) to test different intervention scenarios.	TS 7mo 166km ²	1,585 geolocated (of 4,352)	hospital, ~1/2 cases geolocated (confirmed by home visit)
Salje et al. ^{RT} [35] 2018	Dengue virus isolation + serological evidence	Kamphaeng Phet province ^r , TH	<ul style="list-style-type: none"> • τ (odds, time) <ul style="list-style-type: none"> • odds ratio(place, fixed time windows)¹⁴ <ul style="list-style-type: none"> o case pair, serotype(DENV1-4), school, augmented model infection time (assume symptoms' median incubation period - 7d; undetecteds' infection-to-titre rise = 11d) 	Model diagnostic on inferred undetected subclinical infections—augmented infections shared the temporal clustering (specific to serotype and place) as symptomatic infections.	C 5yr ~98km ²	3,451 with fever symptoms	school 8-11yr age, blood sampled every 3mo, excl. if migration plans within 12mo or thalassaemia.

¹⁷The authors also analysed the spatial relationship of proportions of case pairs falling ill within 6 months and coming from the same transmission chain at different distances. However as a proportion ranges between 0 and 1 we have not included it here as it is not comparable with the positive real τ .

¹⁸Most Recent Common Ancestor

Reviewed papers. Root ^R /Tau ^T paper cited?	Human disease & case defn.	Location rural ^r /urban ^u	Statistic(●) & epidemiological unit(o)	Purpose & stated findings	Study type ⁶ & scale	N ^e cases, people or deaths	Sampling method
Succo et al. ^R [27] 2018	Dengue anti-DENV IgM and IgG +ve + febrile + body temp $\geq 38^{\circ}\text{C}$ + not another condition	Nimes ^u , FR	<ul style="list-style-type: none"> odds ratio (fixed distance window)¹⁹ 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 15d $\sim 0.6\text{km}^2$	1431 people, 512 hhlds, prev. 0.4%	community, residing $\geq 4\text{mo}$, $\geq 2\text{yr}$ age
Rehman et al. ^{R,T} [36] 2018	Dengue confirmed case	Rawalpindi ^u & Lahore ^u , PK	<ul style="list-style-type: none"> ϕ(distance & time)¹⁴ case, hhld GPS, onset date($t_j - t_i \leq 30\text{d} \Rightarrow z_{ij} = 1$) 	ϕ statistic compares interaction of cases in a matched intervention/control study design.	TS 4 & 6yr 259km^2 & $1,772\text{km}^2$,	7,890 & 2,998	community and hospital
Azman et al. ^{R,T} [32] 2018	Cholera acute watery diarrhoea + any age	[33] & Kalamie ^u , CD	<ul style="list-style-type: none"> τ(odds, distance) τ(odds, time) case, hhld GPS, presentation date($t_j - t_i \in [0,4\text{d}], [1,4\text{d}], [0,5\text{d}], \dots, [25\text{d}, 30\text{d}] \Rightarrow z_{ij} = 1$) 	Rationale for targeted intervention: $\leq 100\text{m}$, $\leq 1\text{w}$ of index case presenting/TS 12mo $\sim 64\text{km}^2$	1,692/4,359 & 1,077/1,146 (geolocated/all)/hospital, all cases geolocated
Truelove et al. ^{R,T} [31] 2019	Measles	TZ ^{ru}	<ul style="list-style-type: none"> risk ratio(prevalence of vacc. status, distance), sample-weighted for clusters¹⁴ time-relatedness is swapped for vacc. status unvacc. proportions of DHS²⁰ 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 ?yr 900km^2	100,000 individuals	community, residences randomly distributed in $30 \times 30\text{km}^2$, vacc. status clustered by random swapping algorithm until empirical reached.

¹⁹Reported as a relative risk in the main text, but as an odds ratio in the Supplementary material
²⁰Demographic Health Survey

References

- [1] E. Knox, Detection of clusters, in: P. Elliot (Ed.), *Methodology of enquiries into disease clustering*, Small Area Health Statistics Unit, LSHTM, London, 1989, pp. 17–20.
- [2] A. B. Lawson, M. Kulldorff, A Review of Cluster Detection Methods, in: A. Lawson, A. Biggeri, D. Böhning, L. Emmanuel, J.-F. Viel, R. Bertollini (Eds.), *Disease Mapping and Risk Assessment for Public Health*, first ed., John Wiley & Sons Ltd., Chichester, 1999, pp. 99–110.
- [3] H. Salje, J. Lessler, T. P. Endy, F. C. Curriero, R. V. Gibbons, A. Nisalak, S. Nimmannitya, S. Kalayanarooj, R. G. Jarman, S. J. Thomas, D. S. Burke, D. A. T. Cummings, Revealing the microscale spatial signature of dengue transmission and immunity in an urban population, *PNAS* 109 (2012) 9535–9538. doi:10.1073/pnas.1120621109.
- [4] J. Lessler, H. Salje, M. K. Grabowski, D. A. T. Cummings, Measuring Spatial Dependence for Infectious Disease Epidemiology, *PLoS ONE* 11 (2016) 1–13. doi:10.1371/journal.pone.0155249.
- [5] T. M. Pollington, M. J. Tildesley, T. D. Hollingsworth, L. A. C. Chapman, Measuring spatiotemporal disease clustering with the tau statistic (2019). URL: <http://arxiv.org/abs/1911.08022>. arXiv:1911.08022.
- [6] M. P. Ward, Spatio-temporal analysis of infectious disease outbreaks in veterinary medicine: clusters, hotspots and foci., *Vet. Ital.* 43 (2007) 559–70. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20422535>.
- [7] L. A. Chapman, C. P. Jewell, S. E. Spencer, L. Pellis, S. Datta, R. Chowdhury, C. Bern, G. F. Medley, T. D. Hollingsworth, The role of case proximity in transmission of visceral leishmaniasis in a highly endemic village in Bangladesh, *PLoS Neglected Trop. Dis.* 12 (2018) 1–29. doi:10.1371/journal.pntd.0006453.
- [8] J. Cuzick, R. Edwards, Spatial Clustering for Inhomogeneous Populations, *J. Royal Stat. Soc. Ser. B* 52 (1990) 73–104. doi:10.1111/j.2517-6161.1990.tb01773.x.

- [9] N. H. Anderson, D. M. Titterington, Some Methods for Investigating Spatial Clustering, with Epidemiological Applications, *J. Royal Stat. Soc. Ser. A* 160 (1997) 87–105. doi:10.1111/1467-985X.00047.
- [10] T. Tango, Comparison of General Tests for Spatial Clustering, in: A. Lawson, A. Biggeri, D. Böhning, L. Emmanuel, J.-F. Viel, R. Bertollini (Eds.), *Disease Mapping and Risk Assessment for Public Health*, first ed., John Wiley & Sons Ltd., Chichester, 1999, pp. 111–117.
- [11] P. J. Diggle, A. G. Chetwynd, S. E. Morris, R. Häggkvist, Second-order analysis of space-time clustering, *Stat. Methods Med. Res.* 4 (1995) 124–136. doi:10.1177/096228029500400203.
- [12] E. Gabriel, P. J. Diggle, Second-order analysis of inhomogeneous spatio-temporal point process data, *Stat. Neerlandica* 63 (2009) 43–51. doi:10.1111/j.1467-9574.2008.00407.x.
- [13] E. Gabriel, P. J. Diggle, B. Rowlingson, F. J. Rodriguez-Cortes, *stpp v2.0-3: Space-Time Point Pattern Simulation, Visualisation and Analysis*, 2018. URL: <https://CRAN.R-project.org/package=stpp>.
- [14] J. M. Bland, *An Introduction to Medical Statistics*, Oxford medical publications, third ed., OUP Oxford, New York, USA, 2000. URL: <https://books.google.co.uk/books?id=J-F6mwEACAAJ>.
- [15] J. Lessler, J. Giles, *IDSpatialStats R package development version*, 2018. URL: <https://github.com/HopkinsIDD/IDSpatialStats>.
- [16] T. M. Pollington, *Tau statistic speedup*, 2019. URL: <https://github.com/t-pollington/tau-statistic-speedup>. doi:10.5281/zenodo.3460744.
- [17] A. B. Lawson, Small Scale: Putative Sources of Hazard, in: A. B. Lawson (Ed.), *Statistical Methods in Spatial Epidemiology*, second ed., John Wiley & Sons Ltd., 2013, pp. 143–187. doi:10.1002/9780470035771.ch7.
- [18] A. B. Lawson, Small Scale : Disease Clustering, in: A. B. Lawson (Ed.), *Statistical Methods in Medical Research*, second ed., John Wiley & Sons Ltd., 2006, pp. 111–141. doi:10.1002/9780470035771.ch6.

- [19] H. Salje, J. Lessler, I. M. Berry, M. C. Melendrez, T. Endy, S. Kalayana-rooj, A. A-Nuegoonpipat, S. Chanama, S. Sangkijporn, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, R. V. Gibbons, S. Iamsirithaworn, L. R. Macareo, I.-K. Yoon, A. Sangarsang, R. G. Jarman, D. A. Cummings, Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size, *Science* 355 (2017) 1302–1306. doi:10.1126/science.aaj9384.
- [20] M. Alam, N. A. Hasan, A. Sadique, N. A. Bhuiyan, K. U. Ahmed, S. Nusrin, G. B. Nair, A. K. Siddique, R. B. Sack, D. A. Sack, A. Huq, R. R. Colwell, Seasonal Cholera Caused by *Vibrio cholerae* Serogroups O1 and O139 in the Coastal Aquatic Environment of Bangladesh, *Appl. Environ. Microbiol.* 72 (2006) 4096–4104. doi:10.1128/aem.00066-06.
- [21] et al Heymann, *Control of Communicable Diseases Manual*, 19th ed., APHA, 2008. doi:10.1086/605668.
- [22] M. Porta, *A Dictionary of Epidemiology*, Fifth Edition: Edited by Miquel Porta, fifth ed., Oxford University Press, New York, 2008.
- [23] C. Fraser, Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic, *PLoS ONE* (2007). doi:10.1371/journal.pone.0000758.
- [24] J. Lessler, *IDSpatialStats v0.3.7 R package read-only CRAN mirror*, 2018. URL: <https://github.com/cran/IDSpatialStats>.
- [25] K. H. Grantz, M. S. Rane, H. Salje, G. E. Glass, S. E. Schachterle, D. A. T. Cummings, Disparities in influenza mortality and transmission related to sociodemographic factors within Chicago in the pandemic of 1918, *PNAS* 113 (2016) 13839–13844. doi:10.1073/pnas.1612838113.
- [26] M. K. Grabowski, J. Lessler, A. D. Redd, J. Kagaayi, O. Laeyendecker, A. Ndyanabo, M. I. Nelson, D. A. Cummings, J. B. Bwanika, A. C. Mueller, S. J. Reynolds, S. Munshaw, S. C. Ray, T. Lutalo, J. Manucci, A. A. Tobian, L. W. Chang, C. Beyrer, J. M. Jennings, F. Nalugoda, D. Serwadda, M. J. Wawer, T. C. Quinn, R. H. Gray, 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 (2014). doi:10.1371/journal.pmed.1001610.
- [27] T. Succo, H. Noël, B. Nikolay, M. Maquart, A. Cochet, I. Leparc-Goffart, O. Catelinois, H. Salje, C. Pelat, P. de Crouy-Chanel, H. de Valk, S. Cauchemez, C. Rousseau, Dengue serosurvey after a 2-month long outbreak in Nîmes, France, 2015: was there more than met the eye?, *Eurosurveillance* 23 (2018). doi:10.2807/1560-7917.ES.2018.23.23.1700482.
- [28] H. Salje, S. Cauchemez, M. T. Alera, I. Rodriguez-Barraquer, B. Thaisomboonsuk, A. Srikiatkachorn, C. B. Lago, D. Villa, C. Klungthong, I. A. Tac-An, S. Fernandez, J. M. Velasco, J. Roque Vito G., A. Nisalak, L. R. Macareo, J. W. Levy, D. Cummings, I.-K. Yoon, Reconstruction of 60 Years of Chikungunya Epidemiology in the Philippines Demonstrates Episodic and Focal Transmission, *J. Inf. Dis.* 213 (2016) 604–610. doi:10.1093/infdis/jiv470.
- [29] P. Bhooniboonchoo, R. V. Gibbons, A. Huang, I.-K. Yoon, D. Budhari, A. Nisalak, N. Chansatiporn, M. Thipayamongkolgul, S. Kalanarooj, T. Endy, A. L. Rothman, A. Srikiatkachorn, S. Green, M. P. Mammen, D. A. Cummings, H. Salje, The Spatial Dynamics of Dengue Virus in Kamphaeng Phet, Thailand, *PLoS Neglected Trop. Dis.* 8 (2014) 6–11. doi:10.1371/journal.pntd.0003138.
- [30] H. Salje, J. Lessler, K. K. Paul, A. S. Azman, M. W. Rahman, M. Rahman, D. Cummings, E. S. Gurley, S. Cauchemez, How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study, *PNAS* 113 (2016) 13420–13425. doi:10.1073/pnas.1611391113.
- [31] S. A. Truelove, M. Graham, W. J. Moss, C. J. E. Metcalf, M. J. Ferrari, J. Lessler, Characterizing the impact of spatial clustering of susceptibility for measles elimination, *Vaccine* 37 (2019) 732–741. doi:10.1016/j.vaccine.2018.12.012.
- [32] A. S. Azman, F. J. Luquero, H. Salje, N. N. Mbaïbardoum, N. Adalbert, M. Ali, E. Bertuzzo, F. Finger, B. Toure, L. A. Massing, R. Ramazani,

- B. Saga, M. Allan, D. Olson, J. Leglise, K. Porten, J. Lessler, Micro-Hotspots of Risk in Urban Cholera Epidemics, *J. Inf. Dis.* 218 (2018) 1164–1168. doi:10.1093/infdis/jiy283.
- [33] F. Finger, E. Bertuzzo, F. J. Luquero, N. Naibei, B. Touré, M. Allan, K. Porten, J. Lessler, A. Rinaldo, A. S. Azman, The potential impact of case-area targeted interventions in response to cholera outbreaks: A modeling study, *PLoS Med.* 15 (2018) 1–27. doi:10.1371/journal.pmed.1002509.
- [34] C. Hoang Quoc, H. Salje, I. Rodriguez-Barraquer, Y. In-Kyu, N. V. V. Chau, N. T. Hung, H. M. Tuan, P. T. Lan, B. Willis, A. Nisalak, S. Kalayanaroj, D. A. Cummings, C. P. Simmons, Synchrony of Dengue Incidence in Ho Chi Minh City and Bangkok, *PLoS Neglected Trop. Dis.* 10 (2016) 1–18. doi:10.1371/journal.pntd.0005188.
- [35] H. Salje, D. A. T. Cummings, I. Rodriguez-Barraquer, L. C. Katzelnick, J. Lessler, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, A. Weg, D. Ellison, L. Macareo, I.-K. Yoon, R. Jarman, S. Thomas, A. L. Rothman, T. Endy, S. Cauchemez, Reconstruction of antibody dynamics and infection histories to evaluate dengue risk, *Nature* 557 (2018) 719–723. doi:10.1038/s41586-018-0157-4.
- [36] N. A. Rehman, H. Salje, M. U. G. Kraemer, L. Subramanian, S. Cauchemez, U. Saif, R. Chunara, Quantifying the impact of dengue containment activities using high-resolution observational data, *bioRxiv* (2018). doi:10.1101/401653.
- [37] J. W. Levy, P. Bhooniboonchoo, S. Simasathien, H. Salje, A. Huang, R. Rangsin, R. G. Jarman, S. Fernandez, C. Klungthong, K. Hussem, R. V. Gibbons, I.-K. Yoon, Elevated transmission of upper respiratory illness among new recruits in military barracks in Thailand, *Influenza Respir. Viruses* 9 (2015) 308–314. doi:10.1111/irv.12345.
- [38] P. Cummings, The Relative Merits of Risk Ratios and Odds Ratios, *JAMA Pediatr.* 163 (2009) 438–445. doi:10.1001/archpediatrics.2009.31.
- [39] M. C. Boily, R. F. Baggaley, L. Wang, B. Masse, R. G. White, R. J. Hayes, M. Alary, Heterosexual risk of HIV-1 infection per sexual act:

- systematic review and meta-analysis of observational studies, *Lancet Inf. Dis.* 9 (2009) 118–129. doi:10.1016/S1473-3099(09)70021-0.
- [40] G. Simpson, Mayer-Hasselwander, Bootstrap sampling: applications in gamma-ray astronomy, *Astron. Astrophys.* (1986) 340–348.
- [41] A. Baddeley, E. Rubak, R. Turner, *Spatial Point Patterns: Methodology and Applications with R*, first ed., CRC Press/Taylor & Francis, Boca Raton, 2015. doi:10.1201/b19708.
- [42] N. G. Reich, T. M. Perl, D. A. Cummings, J. Lessler, Visualizing Clinical Evidence: Citation Networks for the Incubation Periods of Respiratory Viral Infections, *PLoS ONE* 6 (2011). doi:10.1371/journal.pone.0019496.
- [43] A. A. Weil, A. I. Khan, F. Chowdhury, R. C. LaRocque, A. S. G. Faruque, E. T. Ryan, S. B. Calderwood, F. Qadri, J. B. Harris, Clinical Outcomes in Household Contacts of Patients with Cholera in Bangladesh, *Clin. Inf. Dis.* 49 (2009) 1473–1479. doi:10.1086/644779.
- [44] A. S. Azman, J. Rumunu, A. Abubakar, H. West, I. Ciglenecki, T. Helderma, J. F. Wamala, R. de la Rosa Vázquez, W. Perea, D. A. Sack, D. Legros, S. Martin, J. Lessler, F. J. Luquero, Population-Level Effect of Cholera Vaccine on Displaced Populations, South Sudan, 2014, *Emerg. Infect. Dis.* 22 (2016) 1067–1070. doi:10.3201/eid2206.151592.
- [45] J. Aldstadt, I.-k. Yoon, D. Tannitisupawong, R. G. Jarman, S. J. Thomas, R. V. Gibbons, A. Uppapong, S. Iamsirithaworn, A. L. Rothman, T. W. Scott, T. Endy, 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 (2012) 1076–1085. doi:10.1111/j.1365-3156.2012.03040.x.
- [46] M. A. Vink, M. C. J. Bootsma, J. Wallinga, Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis, *Am. J. Epidemiol.* 180 (2014) 865–875. doi:10.1093/aje/kwu209.
- [47] A. Cori, N. M. Ferguson, C. Fraser, S. Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, *Am. J. Epidemiol.* 178 (2013) 1505–1512. doi:10.1093/aje/kwt133.

- [48] M. Myllymäki, M. Myllymäki, T. Tom, T. Mrkvička, P. Grabarnik, H. Seijo, U. Hahn, Global envelope tests for spatial processes, *J. Royal Stat. Soc. Ser. B* 79 (2017) 381–404. doi:10.1111/rssb.12172.
- [49] H. Salje, D. A. Cummings, J. Lessler, Estimating infectious disease transmission distances using the overall distribution of cases, *Epidemics* 17 (2016) 10–18. doi:10.1016/j.epidem.2016.10.001.