

Infectious Disease Modelling

Michael Höhle

Department of Mathematics, Stockholm University, Sweden
hoehle@math.su.se

16 March 2015

This is an author-created preprint of a book chapter to appear in the Handbook on Spatial Epidemiology edited by Andrew Lawson, Sudipto Banerjee, Robert Haining and Lola Ugarte, CRC Press. The final version of this text is to be found in this book – once available the ISBN number of the book will be given. The preprint is available as http://www.math.su.se/~hoehle/pubs/Hoehle_SpaMethInfEpiModelling2015.pdf

1 Introduction

Infectious diseases impose a critical challenge to human, animal and plant health. Emerging and re-emerging pathogens – like SARS, influenza, hemorrhagic fever among humans, or foot and mouth disease and classical swine fever among animals – hit the news coverage with regular certainty. Zoonoses and host transmitted diseases underline how tight the connection is between human and animal diseases. While plant epidemics receive less immediate attention, they can severely impact crop yield or wipe out entire species. Unifying for the above epidemics is that they all represent realization of temporal processes. Why does the spatial dimension then matter for the modelling of epidemics? It depends very much on the aims of the analysis: Many relevant questions can be adequately answered by models considering the population as being homogeneous. However, in other situations heterogeneity is important, e.g., induced by age or spatial structure of the population. Spatially varying demographic and environmental factors could influence the disease transmission. Furthermore, having a spatial resolution allows the model to express spatial heterogeneity in the manifestation of the disease over time. This becomes particularly important when investigating the probability of fade-out or short-term predicting the location of new cases. This kind of analysis represents an important mathematical contribution aimed at understanding the dynamics of disease transmission and predicting the course of epidemics in order to, for example, assess control measures or determine the source of an epidemic. This chapter is about the *spatio-temporal* analysis of epidemic processes.

2 The Role of Space in Infectious Disease Epidemiology

The focus in this chapter will be on communicable microparasite infections (typically viral or bacterial diseases) among humans – though the application of mathematical modelling is equally immediate in animal (Dohoo et al., 2010) and plant epidemics (Madden et al., 2007).

Following Giasecke (2002), epidemiology is about 'the study of diseases and their determinants in populations'. The concepts of incidence and prevalence, known from chronic disease

epidemiology, also apply to infectious disease epidemiology. However, contrary to chronic disease modelling it is important to realize that in infectious disease epidemiology each case is also a risk factor and that not everyone is necessarily susceptible to a disease (e.g. immunity due to previous infections or vaccination). As a consequence, the study of infectious diseases is very much concerned with the study of interacting populations. Still, as treated in for example [Becker \(1989\)](#) or [Hens et al. \(2012\)](#), the principal of regression by linear models (LMs), generalized linear models (GLMs) or survival models applies to a variety of problems in infectious disease epidemiology. Hence, spatial extensions of such regression models, e.g. structured additive regression models ([Fahrmeir and Kneib, 2011](#)), also immediately apply to the context of infectious disease epidemiology. Examples are the use of spatially enhanced ecological Poisson regression models with CAR random effects to investigate the influence of socio-economic factors on the incidence of specific infectious diseases ([Wilking et al., 2012](#)) or the assessment of mumps outbreak risk based on serological data by using generalized additive models (GAMs) with 2D splines adjusting for location of individuals ([Abrams et al., 2014](#)). Another application of classical methods from spatial epidemiology is the use of spatial point process models for investigating putative sources for a foodborne outbreak ([Diggle and Rowlingson, 1994](#)).

In this chapter we will, however, restrict the focus to the use of spatio-temporal *transmission models*, that is dynamic models for the person-to-person spread of a disease in a well defined population. The use of such models has become increasingly popular in the epidemiological literature in order to assess risk of emerging pathogens or evaluate control measures. Even spatial aspects of such evaluations are now feasible due to the fact that data become spatially more refined and computer power allows for more complex models to be investigated. The modelling of measles and influenza are two examples where the impact of space has been especially thoroughly investigated ([Grenfell et al., 1995, 2001](#); [Viboud et al., 2006](#); [Eggo et al., 2011](#); [Gog et al., 2014](#)). As an illustration, we will look in detail at the spatio-temporal modelling of biweekly measles counts in 60 towns and cities in England and the UK, 1944-1966. [Figure 1](#) shows the locations of the 60 cities, including both very large and very small cities, and forming a subset of the data for 954 communities used in [Xia et al. \(2004\)](#). An illustration of the the time series for the three largest and three smallest cities in the data set can be found as part of [Fig. 5](#). The aim of such an analysis is to quantify the effect of demographics and seasonality on the dynamics, but also to investigate the role of spatial spread on extinction and re-introduction.

In spatial analyses of this kind the movement of populations plays an important role. From a historic perspective especially mobility has undergone dramatic changes within the last 200 years – for instance [Cliff and Haggett \(2004\)](#) talk about this as the ‘collapse of geographical space’: The time for travelling long distances has reduced immensely, which has resulted in populations mixing at increasingly higher rates. Whilst there has been an abundance of papers and animations illustrating global spread of emerging diseases, e.g. due to airline travelling, the core transmission of many ‘neglected’ diseases still occurs at short-range: in the household, in the kindergarten, at work. The role of space in transmission models is to be studied within this dissonance between long-distance and short-distance spread. A number of chapters and articles have already surveyed this field, e.g. [Isham \(2004\)](#), [Deardon et al. \(2015\)](#), [Riley et al. \(2014\)](#) and [Held and Paul \(2013\)](#). The emphasis in the present chapter is on metapopulation models and their likelihood based inference. It is structured as follows: [Section 3](#) consists of a primer on continuous-time and discrete time epidemic modelling – first for a homogeneous population and then for spatially coupled populations. [Section 4](#) then illustrates the application of discrete-time versions of such models to the 60

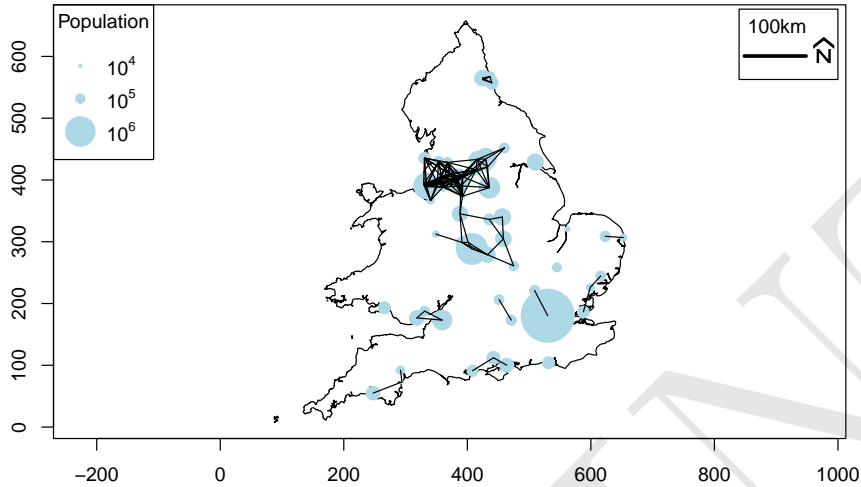


Figure 1: Location of the 60 cities in England and Wales (OSGB36 reference system) used in the measles modelling. The area of each circle is proportional to the city’s population and the connecting lines indicate immediate neighbourhood (distance ≤ 50 km).

cities measles data. A discussion in Section 5 ends the chapter.

3 Epidemic Modelling

Mathematical modeling of infectious diseases has become a key tool in order to understand, predict and control the spread of infections. The fundamental difference to chronic disease epidemiology is that the temporal aspect is paramount. The aim of epidemic modelling is thus to model the spread of a disease in a population made up of a (possibly large) integer number of individuals. To simplify the description of this population, it is common to use a compartmental approach to modelling – for instance in its simplest form the population is divided into classes of *susceptible*, *infective* and *recovered* individuals. Disease dynamics can then be characterized by a mathematical description of each individual’s transitions between compartments, subject to the state of the remaining individuals.

3.1 Continuous time modelling

A number of books, e.g. [Anderson and May \(1991\)](#), [Diekmann and Heesterbeek \(2000\)](#), [Keeling and Rohani \(2008\)](#), give an introduction to epidemic modelling using primarily deterministic models based on ordinary differential equations (ODEs) in the setting of the susceptible-infective-recovered (SIR) model and its extensions. Let $S(t)$, $I(t)$ and $R(t)$ denote the number at time t of susceptible, infective and recovered individuals, respectively. Then the dynamics in the basic *deterministic SIR model* in a population of fixed size can be expressed as in the seminal work by [Kermack and McKendrick \(1927\)](#):

$$\frac{dS(t)}{dt} = -\frac{\beta}{N}S(t)I(t), \quad \frac{dI(t)}{dt} = \frac{\beta}{N}S(t)I(t) - \gamma I(t), \quad \frac{dR(t)}{dt} = \gamma I(t),$$

where the parameter $\beta > 0$ is the *transmission rate* and $\gamma > 0$ describes the *removal rate*. The initial condition is given by $S(0)$, $I(0)$, which are known integers, and $R(0) = 0$. In

a population of fixed size $N = S(0) + I(0)$ the expression for $dR(t)/dt$ in the above ODE system is redundant because $R(t)$ is implicitly given as $N - S(t) - I(t)$.

ODE modelling implies an approximation of the integer sized population using continuous numbers and that the stochastic behaviour of an epidemic is sacrificed by looking at a deterministic average behavior. If the population under study is large enough, such approximations are reasonably valid to obtain a biological understanding. In small populations, however, stochasticity plays an important role for extinction, which cannot be ignored. Stochastic epidemic modelling is described e.g. in [Bailey \(1975\)](#), [Becker \(1989\)](#), [Daley and Gani \(1999\)](#) and [Andersson and Britton \(2000\)](#), who all rely heavily on the theory of stochastic processes. In its simplest form, the basic discrete-state *stochastic SIR model* can be described as a general birth and death process, where the event rates for infection and removal are given as follows:

Event	Rate
$(S(t), I(t)) \rightarrow (S(t) - 1, I(t) + 1)$	$\frac{\beta}{N} S(t) I(t)$
$(S(t), I(t)) \rightarrow (S(t), I(t) - 1)$	$\gamma I(t)$

Again, the development of $R(t)$ is implicitly given, because a fixed population of size $S(0) + I(0)$ is assumed. Notice also how the integer size of the population is now taken into account: Once $I(t) = 0$, the epidemic ceases. From a point process viewpoint the above specification corresponds to an assumption of piecewise constant conditional intensities for the process of infection, while the length of the infective period is given by independent and identically distributed exponential random variables. An important point is that the deterministic SIR model is not just modelling the expectation of the stochastic SIR model. As an illustration, [Renshaw \(1991, Chapter 10\)](#) shows, based on calculations in [Bailey \(1950\)](#), how the expected number of susceptibles $\mu(t) = E(S(t))$ in the stochastic SI model differs from the solution of $S(t)$ in the deterministic SI model. [Figure 2](#) shows the result for a population with $S(0) = 10$, $I(0) = 1$, $\beta = 11$ and $\gamma = 0$. One notices the differences between the deterministic and stochastic model.

Finding an analytic expression for $\mu(t)$ or – even better – the probability mass function (PMF) of $S(t)$ at a specific time point t is less easy already for the SIR model and intractable for most models. Instead, a numerical approach is to formulate a first order differential-difference equation describing the time evolution of $P((S(t), I(t)) = (x, y))$ for each possible (x, y) . Such an approach is known as a master equation approach and corresponds to a discrete state continuous time Markov jump process with the solution of the master equation obeying the Chapman-Kolmogorov equation. These ODE equations are then solved numerically. For large populations the problem can, however, become intractable to solve even numerically. For further details on such stochastic population modelling see, e.g., [Renshaw \(1991\)](#). An alternative to the above is to resort to Monte Carlo simulation of the stochastic epidemic process – a method which has become increasingly popular even for inferential purposes. For the basic SIR model one needs to simulate a discrete-state continuous-time stochastic process with piecewise constant conditional intensity functions (CIFs). Several algorithms exist for doing this, see e.g. [Wilkinson \(2006\)](#) for an overview, of which the algorithm by [Gillespie \(1977\)](#) is the best known. As an example, [Fig. 2](#) shows 10 realizations of the previous SI model and the resulting empirical probability distribution of $S(t)$ computed from 1000 simulations. Besides simplicity Monte Carlo simulations have the additional advantage of being very flexible. For example, it is easy to use the samples to compute pointwise 95% prediction intervals for $S(t)$ or to compute the final number of infected.

A further important difference between deterministic and stochastic modelling is the interpretation of the *basic reproduction number* R_0 , which describes the reproductive potential

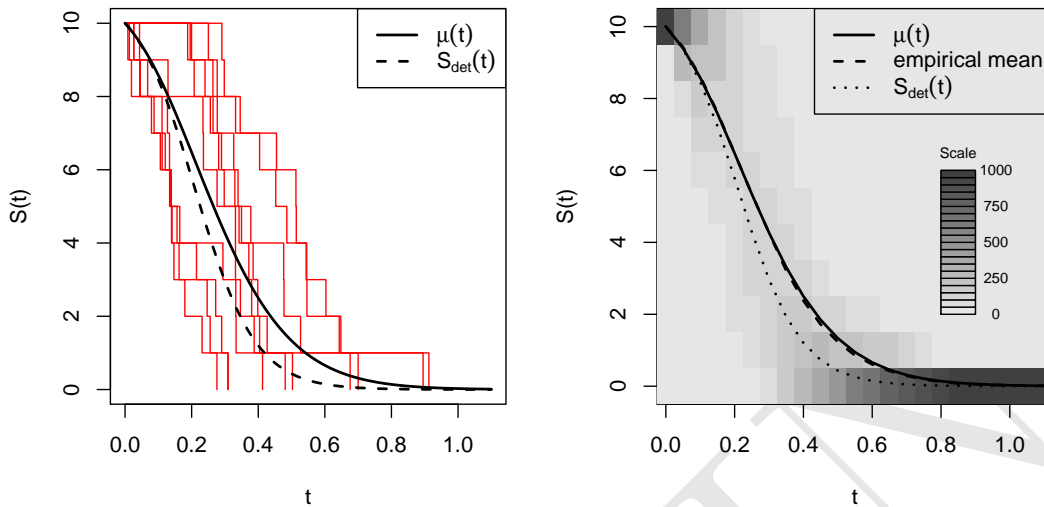


Figure 2: (left) $S(t)$ for 10 realizations of the SI model with parameters as described in the text. Also shown is $E(S(t))$ in the stochastic model and the solution $S(t)$ in the deterministic model. (right) Histograms of the $S(t)$ distribution based on 1000 realizations. The 23 histograms are computed for a grid between time 0.0 and 1.1 with a step size of 0.05. Also shown are the analytic mean $\mu(t)$ together with the 23 empirical means.

of an infectious disease and which is defined as the average number of secondary cases directly caused by an infectious case in an entirely susceptible population. In deterministic models a major outbreak can only occur if $R_0 > 1$. In stochastic models, if $R_0 > 1$, then a major outbreak will occur with a certain probability determined by the model parameters. See, e.g., [Andersson and Britton \(2000\)](#) or [Britton \(2010\)](#) for further details.

Other variations of the basic SI and SIR model could be, for example, the SI-Susceptible, SIR-Susceptible or S-Exposed-IR model. Further extensions consist of reflecting the protection due to maternal antibodies or vaccination by respective compartments in the models. Finally, the rates can be additionally modified to, e.g., reflect the import of infected from outside the population or demographics such as birth and death of individuals. See, e.g. [Keeling and Rohani \(2008, Chapter 2\)](#) for additional details.

3.2 A Spatial Extension: The Metapopulation Model

If interest is in enhancing the homogeneous SIR model with heterogeneity due to spatial aspects, one common modelling approach is to divide the overall population into a number of subpopulations – a so called *metapopulation model* ([Keeling et al., 2004, Chapter 17](#)). Here, one assumes that within each subpopulation the mixing is homogeneous whereas coupling between the subpopulations occurs by letting the force of infection contain contributions of the infectious from the other populations as well. Considering a total of K subpopulations the deterministic metapopulation SIR model looks as follows ([Keeling and Rohani, 2008,](#)

Chapter 7.2):

$$\begin{aligned}\frac{dS_k(t)}{dt} &= -\frac{\beta_k}{N_k} S_k(t) \left(\sum_{l=1}^K w_{kl} I_l(t) \right), \quad k = 1, \dots, K, \\ \frac{dI_k(t)}{dt} &= \frac{\beta_k}{N_k} S_k(t) \left(\sum_{l=1}^K w_{kl} I_l(t) \right) - \gamma I_k(t),\end{aligned}$$

where the weights w_{kl} quantify the impact of one infectious from population l on population k . Typically, these weights are scaled such that $w_{kk} = 1, k = 1, \dots, K$. Consequently, for a susceptible in k , w_{kl} represents the influence of an infectious in unit l relative to one in the same unit as the susceptible. The appropriate choice of weights depends on the modelled disease, its mode of transmission and the questions to be answered. They can even be subject to parametric modelling: For plant or animal diseases dispersal could for example be due to airborne spread which makes distance kernels such as exponential $w_{kl} \propto \exp(-\rho \text{dist}_{kl})$ or power law $w_{kl} \propto \text{dist}_{kl}^{-\rho}$ convenient choices. Here, dist_{kl} denotes the geographic distance between the population k and l . When restricting attention to directly transmitted human diseases the focus is instead on the movement of infectious individuals from population l to population k – not necessarily due to a permanent relocation of individuals, but more due to temporarily movement. This could, for instance in the case of seasonal influenza, be the commuting of individuals or in the case of emerging epidemics (e.g. SARS, swine influenza) due to long-distance airplane travelling. If commuter or airline data are available these can be used to determine the weights. Movement exhibits a strong age-dependence, though, which can make it difficult to extract the relevant information from such sources for childhood diseases. As a consequence, one might use distance based kernels as a proxy for mobility – possibly augmented by population sizes as in the *gravity model* (Erlander and Stewart, 1990; Xia et al., 2004). If interest is in short-term prediction of an emerging pathogen, long-distance travelling is an important concept to capture. However, the main bulk of infections for an established disease typically happens at a much smaller geographical scale as shown, for example, in recent re-analyses of the 1918 pandemic influenza (Eggo et al., 2011) and the 2009 swine influenza outbreak (Gog et al., 2014).

In analogy to the above deterministic metapopulation model, the system of rates of the continuous time stochastic SIR model can be modified accordingly to obtain a stochastic metapopulation model with unit specific (conditional) intensity function for infection events

$$\lambda_k(t) = \frac{\beta_k}{N_k} S_k(t) \left(\sum_{l=1}^K w_{kl} I_l(t) \right), \quad (1)$$

while the unit specific intensity function for recovery events is now $\gamma I_k(t)$ (assuming exponentially distributed infectious periods). One important insight is that the difference between deterministic and stochastic metapopulation models is increased as subpopulations become smaller. Studying the extinction and re-introduction of disease in such metapopulations is hence preferably conducted using stochastic metapopulation models, see e.g. Bjørnstad and Grenfell (2008). Two variants of the stochastic metapopulation model are of additional interest: The first variant is the so called Levins-type metapopulation model (Keeling and Rohani, 2008, Section 7.2.4), where one for each k is only interested in the probability of $I_k(t) > 0$ as a function of time. Such models have been used to study the arrival of the first disease cases in a city during a pandemic (Eggo et al., 2011; Gog et al., 2014). The second variant

is the individual model, i.e. when the number of considered subpopulations is equal to the population size $K = N$.

Both cases are instances of multivariate counting processes which can be consistently handled in the SIR-S framework using a so called two-component SIRS model, which additionally allows for immigration of disease case from external sources. This is particularly useful if the data contains multiple outbreaks. Following [Höhle \(2009\)](#) let $N_k(t), k = 1, \dots, K$, denote the counting process, which for unit k counts the number of changes from state susceptible to state infectious. The conditional intensity function given the history of all K processes up to, but not including, t is then given as:

$$\begin{aligned} \lambda_k(t) &= \exp(h_0(t) + \mathbf{z}_k(t)^T \boldsymbol{\beta}) + \sum_{j \in I(t)} \left\{ w(\text{dist}_{kj}) + \mathbf{v}_{kj}(t)^T \boldsymbol{\alpha}_e \right\} \\ &= Y_k(t) \cdot [\exp(h_0(t)) \exp(\mathbf{z}_k(t)^T \boldsymbol{\beta}) + \mathbf{x}_k(t)^T \boldsymbol{\alpha}], \end{aligned} \quad (2)$$

where $Y_k(t)$ is an indicator if unit k is susceptible, i.e. $Y_k(t) = \mathbb{1}_{k \in S(t)}$, while $S(t)$ and $I(t)$ now denotes the index set of all susceptibles and infectious, respectively. Furthermore, $w(\cdot)$ denotes a distance weighting kernel parametrised as a spline function while $\mathbf{z}_k(t)$ and $\mathbf{v}_{kj}(t)$ denote possibly time varying covariates affecting the introduction of new cases in the endemic and epidemic components, respectively, and $\mathbf{x}_k(t)$ denotes the combination of linear spline terms and epidemic covariates. Finally, $h_0(t)$ is the base-line rate of the endemic component. If import of new cases from external sources is not relevant the first component in the above can be left out (i.e. set to zero).

3.3 Fitting SIR Models to Data

From a statistician's point of view, parameter inference in epidemic models appears to receive only marginal attention in the medical literature. One reason might be that little or no data are available and hence parameters are 'guesstimated' from literature studies or expert knowledge. Another reason is that inference often boils down to extracting information about parameter values from a *single* realization of the stochastic epidemic process. For deterministic models, estimation might also appear less of a statistical issue. Finally, estimation is complicated by the epidemic process only being partially observable: The number of susceptibles to begin with as well as the time point of infection of each case with might be unknown. As a consequence, parameter estimation in ODE based SIR models has, typically, been done using least square type estimation based on observable quantities. Only recently, the models have been extended to non-Gaussian observational components enabling count-data likelihood based statistical inference, see e.g. [Pitzer et al. \(2009\)](#),

$$\mathbb{E}(Y_{t,k}) = \int_t^{t+1} \left\{ -\frac{dS_k(u; \boldsymbol{\theta})}{du} \right\} du.$$

where $Y_{t,k}$ is the observed number of *new* infections within time interval t , which is assumed to follow a count data PMF f such as the Poisson or negative binomial distribution with expectation $\mathbb{E}(Y_{t,k})$. This setup allows the estimation of the model parameters $\boldsymbol{\psi}$ in a likelihood framework by the loglikelihood function $l(\boldsymbol{\psi}) = \sum_{t=1}^T \sum_{k=1}^K \log f(y_{t,k}; \boldsymbol{\psi})$, when assuming observations are independent given the model. However, a residual analysis often shows remaining auto-correlation, which means any quantification of estimation uncertainty based on asymptotic theory using the Hessian of the loglikelihood tends to be overly optimistic. As a consequence, the confidence intervals might be too narrow. A novel two-step approach

to improve on this shortcoming within the above framework is treated in [Weidemann et al. \(2014\)](#).

An advantage of stochastic continuous-time SIR modelling is that it allows for a quantification of the uncertainty of the estimates, even though the estimate is based on a single process realization. The work of [Becker \(1989\)](#) and the second part of [Andersson and Britton \(2000\)](#) are some of the few books dedicated to this task. If the epidemic process $(S(t), I(t))'$ is completely observed over the interval $(0, T]$, where T is the entire duration of the epidemic, the resulting data of the epidemic are given as $\{(t_i, y_i, k_i), i = 1, \dots, n\}$, where n is the number of infections in the population during the epidemic, t_i denotes the time of infection of individual i , y_i is the length of the individuals' infectious period and $k_i \in \{1, \dots, K\}$ is the unit it belongs to. Further assuming that the PDF of the infectious period is exponential $f_{I \rightarrow R}(y) = \gamma \exp(-\gamma y)$ we obtain the loglikelihood for the parameter vector $\psi = (\beta, \gamma)'$ as

$$l(\psi) = \sum_{i=1}^n \log f_{I \rightarrow R}(y_i) + \sum_{i=1}^n \log \lambda_{k_i}(t_i) - \int_0^T \sum_{k=1}^K \lambda_k(u) du, \quad (3)$$

where $\lambda_k(t_i)$ is defined as in (1) and evaluated at the time just prior to t_i . Note that in (3) the CIFs have to be integrated over time, however, for the simple SIR model the CIF is a piecewise constant function between events and hence integration is tractable. [Höhle \(2009\)](#) develops these likelihood equations further using counting process notation for the two-component SIR model (2), whereas [Lawson and Leimich \(2000\)](#), [Diggle \(2006\)](#) and [Scheel et al. \(2007\)](#) contain accounts and examples of a partial-likelihood approach for spatial SIR type model inference including covariates.

In applications the times of infection t_i of infected individuals would typically be unknown. One way to make inference tractable is to assume that the duration of the infectious period is a constant, say, $\mu_{I \rightarrow R}$ (known or to be estimated). Furthermore, the initial number of susceptibles might also be unknown and might require estimation. See [Becker \(1989\)](#) for details, which also covers a discrete-time approximation covered in the next section. More recently, [Gibson and Renshaw \(1998\)](#), [O'Neill and Roberts \(1999\)](#), [O'Neill and Becker \(2001\)](#), [Neal and Roberts \(2004\)](#) and [Höhle et al. \(2005\)](#) use a Bayesian data augmentation approach using Markov Chain Monte Carlo to impute the missing infection times while simultaneously performing Bayesian parameter inference for the S(E)IR and metapopulation S(E)IR model. Model diagnostics can in the likelihood context be performed using a graphical assessment of residuals and forward simulation ([Höhle, 2009](#)); for the Bayesian models one can use additional posterior predictive checks and latent residuals ([Lau et al., 2014](#)).

3.4 Discrete time models

Up to now, focus has been on continuous time epidemic modeling. However, data are usually only available at much coarser time scales: weekly or daily reporting is usual in public health surveillance. If individual data is available, this observational situation can be handled by considering the observed event times as being for the event times interval censored. However, when looking at large populations or routinely collected data, data is typically provided in aggregated form without access to the individual data. It is thus necessary to consider alternative ways of casting the continuous time stochastic SIR approach into a discrete time framework. Time series analysis is one such approach, providing a synthesis between complex stochastic modeling and available data.

In order to derive such a model, we consider a sufficiently small time interval $[t, t + \delta t)$. We could then – as an approximation – assume constant conditional intensity functions of our

SIR model in $[t, t + \delta t)$. By definition, let these intensities be equal to the intensities at the left time point of the interval, i.e. at time t . This implies that all individuals are independent for the duration of the interval. Looking at one susceptible individual, its probability of escaping infection during $[t, t + \delta t)$ is then equal to $\exp(-\beta I(t) \cdot \delta t)$. Denoting by $C_{[t, t + \delta t)}$ the number of newly infected and by $D_{[t, t + \delta t)}$ the number of recoveries in the interval $[t, t + \delta t)$, we obtain

$$C_{[t, t + \delta t)} \sim \text{Bin} \left(S(t), 1 - \exp(-\beta I(t) \cdot \delta t) \right), \quad (4)$$

$$D_{[t, t + \delta t)} \sim \text{Bin} \left(I(t), 1 - \exp(-\gamma \cdot \delta t) \right). \quad (5)$$

The state at time $t + \delta t$ is then given by $S(t + \delta t) = S(t) - C_{[t, t + \delta t)}$, $I(t + \delta t) = I(t) + C_{[t, t + \delta t)} - D_{[t, t + \delta t)}$. Now changing notation to discrete time with discrete time subscript t denoting the time $t + \delta t$ and subscript $t - 1$ the time t we write:

$$S_t = S_{t-1} - C_t, \quad I_t = I_{t-1} + C_t - D_t, \quad (6)$$

for $t = 1, 2, \dots$ and with $S_0 = S(0), I_0 = I(0)$. Consequently, the discrete quantities C_t, D_t , etc. now replace the continuous ones in (4) and (5). Such models are known as chain binomial models (Becker, 1989), if one assumes that $D_t = I_t$, i.e. the time scale is chosen such that all infective individuals recover after one time step. In such models one time step can be seen as one generation time (Daley and Gani, 1999, Chapter 4). For large S_t and I_t the binomial distributions can be further approximated by Poisson distributions: by a first order Taylor expansion of the $1 - \exp(-x)$ terms one obtains

$$C_t \sim \text{Po} \left(\beta S_{t-1} I_{t-1} \cdot \delta t \right), \quad (7)$$

$$D_t \sim \text{Po} \left(\gamma I_{t-1} \cdot \delta t \right). \quad (8)$$

Note that these approximations no longer ensure that $C_t \leq S_{t-1}$ and $D_t \leq I_{t-1}$. If this is a practical concern, one can instead use right truncated Poisson distributions fulfilling the conditions. Altogether, we have transformed the continuous time stochastic model into a discrete time model. If $(S_t, I_t)'$ is known at each point in time $t = 1, \dots, T$, estimates for β and γ can be found using maximum likelihood approaches based on (7)-(8). For example, Becker (1989) shows how the above equations can be used to fit a homogeneous SIR model to data using generalized linear model (GLM) software: (8) can be represented as a log-link Poisson GLM with offset $\log(S_{t-1} I_{t-1} \delta t)$ and intercept $\log(\beta)$ or an identity link GLM with covariate $S_{t-1} I_{t-1} \delta t$ and no intercept. For the stochastic metapopulation SIR model with known weights the idea can similarly be extended by jointly modelling $C_{t,k}$ and $D_{t,k}$, $1 \leq t \leq T$, $1 \leq k \leq K$ by an appropriate Poisson GLM with either log-link or identity link with offset, intercept and potential covariates derived from (1). Furthermore, Klinkenberg et al. (2002) uses the above equations to fit a spatial grid SIR model using numerical maximization of the binomial likelihood.

3.5 Time Series SIR model

The model given by (7)-(8) corresponds to a simple version of the *time series SIR model* (TSIR model) initially proposed in Finkenstädt and Grenfell (2000) and since extended (Finkenstädt et al., 2002; Bjørnstad et al., 2002). For the TSIR model it is assumed that $D_t = I_t$, i.e. as in a chain binomial model, and $\delta t = 1$. As an example, one time unit corresponds to a biweekly scale when modelling measles. In addition, the TSIR model contains additional flexibility

beyond the simple SIR model, e.g. by taking population demographics into account, where births provide new susceptibles and immigration provides an influx of infectives allowing the re-introduction of cases into a population where the disease was already extinct. Further extensions are a time-varying transmission rate, e.g. due to school closings, and a modification of the multiplication term $I(t)$ in the transmission rate to $I(t)^\alpha$, which allows for spatial substructures and other forms of heterogeneity in the population (Liu et al., 1987). Xia et al. (2004) extends the model to a multivariate (and hence spatial) time series model, where the transmission between populations is based on a gravity model.

Below we present a slightly modified version of this *multivariate TSIR model* (mTSIR) given in Xia et al. (2004). Let $I_{t,k}$ and $S_{t,k}$ be the number of infectious and susceptibles in region k at time t , where $1 \leq k \leq K$ and $1 \leq t \leq T$. For $t = 2, \dots, T$ the model is now defined by

$$\begin{aligned}
 I_{t,k} | \lambda_{t,k}, \iota_{t,k} &\sim \text{NB}(\lambda_{t,k}, I_{t-1,k}), \text{ where} \\
 \lambda_{t,k} | \iota_{t,k} &= \frac{1}{N_{t-1,k}} \beta(t) S_{t-1,k} (I_{t-1,k} + \iota_{t,k})^\alpha, \\
 \iota_{t,k} &\sim \text{Ga}(m_{t,k}, 1), \text{ and} \\
 m_{t,k} &= \theta N_{t-1,k}^{\tau_1} \sum_{j \neq k} d_{j,k}^{-\rho} I_{t-1,j}^{\tau_2}.
 \end{aligned} \tag{9}$$

Here, $\text{NB}(\mu, c)$ denotes the negative binomial distribution with expectation μ and clumping parameter c , i.e. the variance is $\mu + \mu^2/c$. For example, Bailey (1964, Chapter 8) shows that the number of offspring generated by a pure birth process with an according rate and starting with a population of c individuals has exactly this distribution. Note, however, that the model in (9) is a double-stochastic model, because the rate itself is a random variable due to the random influx of new infectives which is gamma distributed. This implies that there is a small twist in case $I_{t-1,k} = 0$, because (9) would then imply that $I_{t,k} | (I_{t-1,k} = 0) \equiv 0$, which is not as intended. The reason is that the birth process motivation is not directly applicable this situation – instead we will assume that $I_{t,k} | I_{t-1,k} = 0$ is just $\text{Po}(\lambda_t)$ distributed. Another approach is to assume that the clumping parameter in (9) is equal to $I_{t-1,k} + \iota_{t,k}$ (Morton and Finkenstädt, 2005; Bjørnstad and Grenfell, 2008). In practice, if $I_{t-1,k}$ is large the negative binomial effectively reduces to the Poisson distribution. The additional complexity given by the negative binomial is thus somewhat artificial – it might be worthwhile to let the clumping parameter vary more freely instead.

Finally, $\beta(t) \geq 0$ is a periodic function with period one year and parametrized by the parameter vector β . For biweekly measles data this could for example be a 26-parameter function or a harmonic seasonal forcing function, i.e.

$$\beta(t) = \beta_{t \bmod 26+1} \quad \text{or} \quad \beta(t) = \beta_1(1 + \beta_2 \cos(2\pi t/26)).$$

The number of susceptibles in the TSIR model is given by the recursion

$$S_{t,k} = S_{t-1,k} - I_{t,k} + B_{t,k} \quad \text{for } t = 2, \dots, T, \tag{10}$$

where $B_{t,k}$ denotes the birth rate in region k at time t . One challenge when using the TSIR model is, however, that $S_{t,k}$ is only partially observable. Untangling the recursion yields that

$$S_{t,k} = S_{1,k} - \sum_{u=2}^t I_{u,k} + \sum_{u=2}^t B_{u,k}, \quad t = 2, \dots, T,$$

where $S_{1,k}$ is usually unknown but needs to be such that $0 < S_{t,k} \leq N_{t,k} - I_{t,k}$ for all $t = 2, \dots, T$. As a consequence, either all unknown $S_{1,k}$'s need to enter the analysis as K additional parameters or one has just a single extra parameter κ together with the coarse assumption that $S_{1,k} = \kappa N_{1,k}$. An alternative is to use a pre-processing step to determine it: Conditional on $S_{1,k}$ and the observed $I_{t,k}$ one can compute the resulting $S_{t,k}$'s. Inspired by the univariate procedure in [Finkenstädt et al. \(2002\)](#) one then considers $\iota_{t,k}$ within the unit to be a time series varying around its mean and uses a Taylor expansions up to order three to derive

$$\log I_{t,k} = \log \beta_t + \alpha \log(I_{t-1,k}) + c_1 I_{t-1,k}^{-1} + c_2 I_{t-1,k}^{-2} + c_3 I_{t-1,k}^{-3} + \log S_{t,k} + \epsilon_{t,k}, \quad (11)$$

where the $\epsilon_{t,k}$'s are zero-mean random variables. Altogether, conditionally on $S_{1,k}$ expression (11) represents a linear regression model which can be fitted using maximum likelihood. The resulting profile loglikelihood in $S_{1,k}$ can then be optimized for each unit separately.

As already mentioned, one problem with the above model formulation, is that the update given in (9) does not ensure that $I_{t,k} \leq S_{t-1,k} + B_{t-1,k}$. In other words $S_{t,k} \geq 0$ is not explicitly ensured by the model. When fitting the model to data this is unproblematic, because the $S_{t,k}$'s are computed by a separate pre-processing step. However, when simulating from the model and hence computing $S_{t,k}$ in each step from (10) it becomes clear that these can become negative if $I_{t,k}$ is large enough. To ensure validity of the model we thus right-truncate the $\text{NB}(\lambda_{t,k}, I_{t-1,k})$ at $S_{t-1,k}$.

3.6 Fitting the mTSIR Model to Data

In the work of [Xia et al. \(2004\)](#) a heuristic optimization criterion is applied, which consists both of a one-step ahead squared prediction error assessment and a criterion aimed at minimizing the absolute difference between the observed and predicted cross-correlation at lag zero between the time series of a unit and the time series of London. The 26 seasonal parameters β_t are explicitly fixed at the values obtained in [Finkenstädt and Grenfell \(2000\)](#). However, from [Xia et al. \(2004\)](#) it is not entirely clear how the point prediction of $\hat{I}_{t,k}$ is defined. Instead, we proceed here using a marginal likelihood oriented approach for inference in the mTSIR model. We do so by computing the marginal distribution of $I_{t,k}$ given $\mathbf{I}_{t-1} = \mathbf{i}_{t-1}$ as follows

$$f_m(i_{t,k} | \mathbf{i}_{t-1}; \boldsymbol{\psi}) = \int_0^\infty f(i_{t,k} | \iota_{t,k}, \mathbf{i}_{t-1}; \boldsymbol{\psi}) f(\iota_{t,k} | \mathbf{i}_{t-1}; \boldsymbol{\psi}) d\iota_{t,k},$$

where the two integrated densities refer to the PMF of the truncated negative binomial and the PDF of the gamma distribution, respectively. The former is computed from the ordinary negative binomial PMF as $f_{\text{NB}}(i_{t,k}) / F_{\text{NB}}(S_{t-1,k})$ with F denoting the cumulative distribution function. Based on the above the marginal loglikelihood for the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}', \alpha, \theta, \tau_1, \tau_2, \rho)'$ is then given as

$$l(\boldsymbol{\psi}) = \sum_{t=2}^T \sum_{k=1}^K \log f_m(i_{t,k} | \mathbf{i}_{t-1}; \boldsymbol{\psi}). \quad (12)$$

This expression is then optimized numerically for $\boldsymbol{\psi}$ in order to find the maximum likelihood estimator (MLE). In our R implementation ([R Core Team, 2014](#)) of the model we use a BFGS-method, as implemented in the function `optim`, for optimization while using the function `gauss.quad.prob` to perform Gaussian quadrature based on orthogonal Laguerre polynomials

for handling the integration over the Gamma densities. In recent work [Jandarov et al. \(2014\)](#) present a more complex Bayesian approach including an evaluation of the required integrals by MCMC ([Smyth, 1998](#)). Note that the quadrature strategy can be numerically difficult in practice, because the the marginal density for some parameter configurations to be evaluated is so small that – with the given floating point precision – it becomes zero. This then causes problems when taking the logarithm. A simplification to allow for estimation is to replace the stochastic $\nu_{t,k}$ component in (9) by its expectation $m_{t,k}$ – as also done in [Jandarov et al. \(2014\)](#) – at the cost of reducing the variability of the model.

3.7 Endemic-Epidemic Modelling

Inspired by the SIR and mTSIR model, [Held et al. \(2005\)](#) presented a multivariate count data time series model for routine surveillance data, which does not require the number of susceptibles to be available. The formal inspiration for the model was the spatial branching process with immigration, which means that observation time and generation time have to correspond. In a series of successive papers the modelling was subsequently extended such that it now constitutes a powerful and flexible regression approach for multivariate count data time series ([Paul et al., 2008](#); [Paul and Held, 2011](#); [Held and Paul, 2012](#); [Meyer and Held, 2014](#)). From a spatial statistics perspective this even includes the use of CAR type random effects ([Besag et al., 1991](#)) in the time series modelling. The fundamental idea is to divide the infection dynamics into two components: an *endemic component* handles the influx of new infections from external sources and an *epidemic component* covers the contagious nature by letting the expected number of transmissions be a function of the lag-one number of infections. The resulting so called HHH model is given by:

$$\begin{aligned}
 Y_{t,k} &\sim \text{NB}(\mu_{t,k}, \gamma_k), \quad t = 2, \dots, T, \quad k = 1, \dots, K, \\
 \mu_{t,k} &= N_{t,k} \nu_{t,k} + \lambda_{t,k} Y_{t-1,k} + \phi_{t,k} \sum_{j \neq k} w_{jk} Y_{t-1,j}.
 \end{aligned}
 \tag{13}$$

In the above, $Y_{t,k}$ denotes the number of new cases in unit k at time t which is assumed to follow a negative binomial distribution with expectation $\mu_{t,k}$ and region specific clumping parameter γ_k . As before, $N_{t,k}$ denotes the corresponding population in region k at time t and the w_{jk} are known weights describing the impact of cases in unit j on unit k . This can, for example, be population flux data such as airline passenger data ([Paul et al., 2008](#)) or neighbourhood indicators such as $w_{jk} \propto I(d_{jk} = 1)$ or a power-law weight $w_{jk} \propto d_{jk}^{-\rho}$, where in the later two examples d_{jk} denotes the graph-based distance between units j and k in the neighbourhood graph while ρ is a parameter to estimate ([Meyer and Held, 2014](#)). Furthermore, $\nu_{t,k}$, $\lambda_{t,k}$ and $\phi_{t,k}$ are linear predictors covering the endemic component as well as the within and between unit auto-regressive behaviour, respectively,

$$\begin{aligned}
 \log(\nu_{t,k}) &= \alpha^{(\nu)} + b_k^{(\nu)} + \mathbf{z}_{t,k}^{(\nu)\prime} \boldsymbol{\beta}^{(\nu)}, \\
 \log(\lambda_{t,k}) &= \alpha^{(\lambda)} + b_k^{(\lambda)} + \mathbf{z}_{t,k}^{(\lambda)\prime} \boldsymbol{\beta}^{(\lambda)}, \\
 \log(\phi_{t,k}) &= \alpha^{(\phi)} + b_k^{(\phi)} + \mathbf{z}_{t,k}^{(\phi)\prime} \boldsymbol{\beta}^{(\phi)}.
 \end{aligned}$$

In each case, $\alpha^{(\cdot)}$ denotes the overall intercept, $b_k^{(\cdot)}$ a unit-specific random intercept and $\mathbf{z}_{t,k}^{(\cdot)}$ represents a length $p^{(\cdot)}$ vector of possibly time-varying predictor specific covariates with associated parameter vector $\boldsymbol{\beta}^{(\cdot)}$. The use of covariates allows for a flexible modelling of, e.g., secular trends and concurrent processes influencing the disease dynamics (temperature, occurrence

of other diseases, vaccination, etc.). The vector of random effects per unit $(b_k^{(\nu)}, b_k^{(\lambda)}, b_k^{(\phi)})'$ can either be iid normal with predictor specific variance, trivariate normal with a correlation matrix (to be estimated) or of CAR type for each predictor. The HHH model allows the absence of one or several of the model components, e.g., lack of endemic component, and also supports the use of offsets in the linear predictors, which is especially convenient to represent population influences as we shall see subsequently. The HHH model has already been successfully used to describe influenza, measles and meningococcal dynamics, see for example Herzog et al. (2011) or Geilhufe et al. (2014).

3.7.1 Linking the HHH and mTSIR model

As an example of the flexibility of HHH modelling we consider it here as a framework for representing a simplified version of the mTSIR model including a gravity model based flux of infectives: Replacing the stochastic $\nu_{t,k}$ component in (9) by its expectation and assuming $\alpha = \tau_2 = 1$ then yields

$$\lambda_{t,k}^{\text{mTSIR}} = \mu_{t,k}^{\text{HHH}} = \frac{1}{N_{t,k}} \beta(t) S_{t-1,k} \left(I_{t-1,k} + \theta N_{t-1,k}^{\tau_1} \sum_{j \neq k} d_{j,k}^{-\rho} I_{t-1,j} \right).$$

Ignoring the difference between $N_{t-1,k}$ and $N_{t,k}$, this corresponds to a HHH model without endemic component, i.e. $\nu_{t,k} \equiv 0$, and

$$\begin{aligned} \log(\lambda_{t,k}) &= \log(S_{t-1,k}/N_{t,k}) + \log \beta(t), \\ \log(\phi_{t,k}) &= \log(S_{t-1,k}) + (\tau_1 - 1) \log(N_{t,k}) + \log(\theta) + \log \beta(t), \end{aligned}$$

together with weights $w_{jk} \propto d_{jk}^{-\rho}$, where d_{jk} is a integer-discretised version (e.g. in steps of 50 km) of the Euclidean distance between units j and i . Note that in both predictors, the first term is an offset and that the number of susceptibles has been included as a time and unit specific covariate. Even if no data on the number of susceptibles is available, good proxies can be established. For example Herzog et al. (2011) perform a spatio-temporal analysis using the county specific vaccination rate against measles as a proxy for the availability of susceptibles and hence as the result of the modeling obtain a quantification of the effectiveness of the vaccination. Note also that the clumping parameter of the negative binomial is different in the two models: in the HHH model it is time constant but varies freely, whereas in the mTSIR model it is time varying but fixed to be the previous number of infectives.

3.8 Fitting HHH Models to Data

Likelihood based inference for the HHH model is performed by maximizing the corresponding (marginal) log-likelihood function, which is similar to (12), with f_m being the (marginalized) PMF of the negative binomial model given by (13). Altogether, a variant of the penalized quasi-likelihood approach discussed in Breslow and Clayton (1993) is used, if the predictors contain random effects. Model selection is performed by AIC or – in case of random effects – using proper-scoring rules for count data based on one-step ahead forecasts (Czado et al., 2009; Paul and Held, 2011); see the respective HHH papers for the inferential details.

4 Measles Dynamics in the Pre-Vaccination Area

In this section we analyze the 1944-1966 biweekly England and Wales measles data already presented in Sect. 2. These pre-vaccination data have been analysed univariately in Finken-

[städt and Grenfell \(2000\)](#) and [Finkenstädt et al. \(2002\)](#) and are available for download from the internet ([Grenfell, 2006](#)). The present analysis is an attempt to analyse the data in spatio-temporal fashion. In a pre-processing step we determine $S_{1,k}$ for each of the 60 series as follows: the linear model (11) is fitted for a grid of $S_{1,k}$ values and the resulting log-likelihood is used to determine $\hat{S}_{1,k}$. Subsequently, the recursion in (10) is used to compute the time series of susceptibles in unit k given $\hat{S}_{1,k}$. Figure 3 shows this exemplarily for the time series of London.

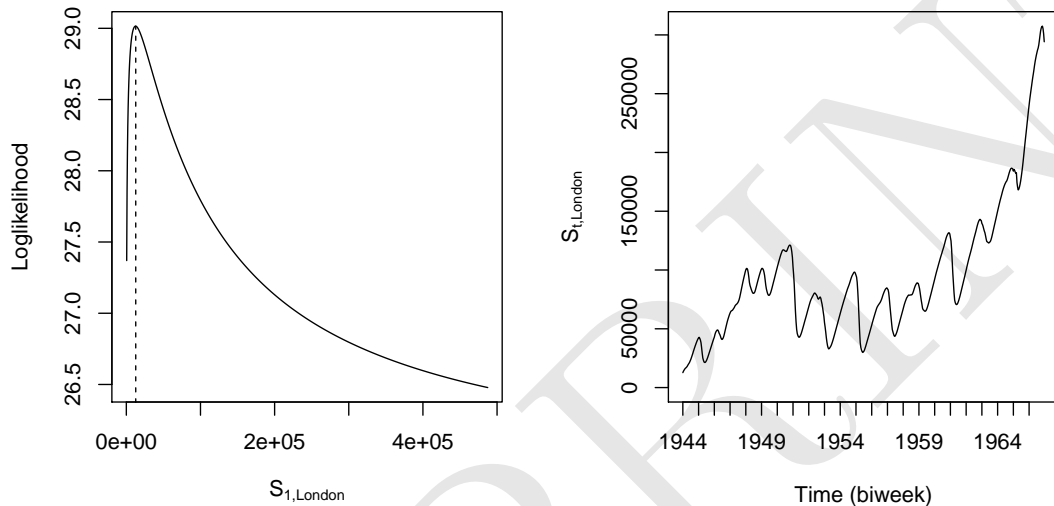


Figure 3: (left) Loglikelihood of the linear model as a function of $S_{1,k}$ for London. (right) Resulting time series of susceptibles obtained from the $S_{1,k}$ maximizing the loglikelihood.

4.1 Results of the mTSIR Model

We use the individually reconstructed series of susceptibles for each city to fit the mTSIR model to the multivariate time series of the 60 cities. Fitting the model based on the likelihood proves to be a complicated matter: We therefore follow the strategy of [Xia et al. \(2004\)](#) by fixing $\beta(t)$ to the 26 parameters found in [Finkenstädt and Grenfell \(2000\)](#) and fixing $\alpha = 0.97$. Table 1 contains the ML results in a model where the $\iota_{t,k}$ variables are replaced with their expectation. The table also contains 95% Wald confidence intervals (CI) based on the observed inverse Fisher information.

	Estimate	2.5%	97.5%
τ_1	2.24	2.24	2.24
τ_2	1.10	1.10	1.10
ρ	$3.05 \cdot 10^{-2}$	$2.94 \cdot 10^{-2}$	$3.16 \cdot 10^{-2}$
θ	$1.79 \cdot 10^{-16}$	$1.77 \cdot 10^{-16}$	$1.81 \cdot 10^{-16}$

Table 1: Parameter estimates and 95% CIs for the mTSIR model.

Altogether, the results differ quite markedly from the results reported in [Xia et al. \(2004\)](#), where the estimates are $\hat{\tau}_1 \approx 1$, $\hat{\tau}_2 \approx 1.5$, $\hat{\rho} \approx 1$ and $\hat{\theta} \approx 4.6 \cdot 10^{-9}$. Several explanations for the differences exist: different estimation approaches and different data were used. Altogether,

it appears that the full mTSIR model appears too flexible for the available reduced data set containing only 60 cities. Still, the one-step ahead 95% predictive distributions obtained by plug-in of the MLE show that the model enables too little variation: a substantial number of observations lay outside the 95% predictive intervals obtained by computing the 2.5% and 97.5% quantile of the truncated negative-binomial in (9). Fig. 4 shows this exemplarily for the London time series. One explanation might be the direct use of $I_{t-1,k}$ as clumping parameter for the negative binomial distribution in (9) – it appears more intuitive that the variance with increasing $I_{t-1,k}$ should increase instead of converging to the Poisson variance as implied by the model.

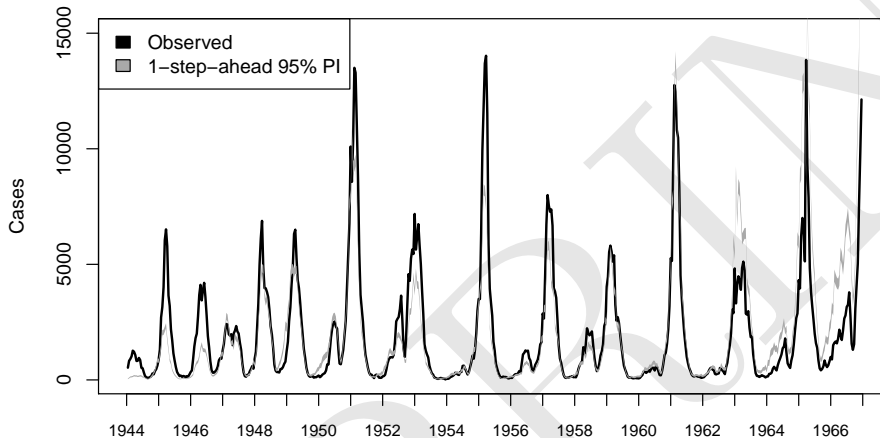


Figure 4: One-step ahead 95% predictive intervals for the London time series obtained by plug-in of the MLE in the mTSIR model. Also shown is the actual observed time series.

4.2 Results of the HHH modelling

Instead of trying to improve the mTSIR model manually, we do so by performing this investigation within the HHH model. As initial HHH model capturing seasonality in both the endemic and epidemic components we consider the following model:

$$\begin{aligned}\log(\nu_{t,k}) &= \log(N_{t,k}) + \alpha^{(\nu)} + \beta_1^{(\nu)} \sin\left(\frac{2\pi t}{26}\right) + \beta_2^{(\nu)} \cos\left(\frac{2\pi t}{26}\right), \\ \log(\lambda_{t,k}) &= \alpha^{(\lambda)} + \beta^{(\lambda)} \log(N_{t,k}) + \beta_1^{(\lambda)} \sin\left(\frac{2\pi t}{26}\right) + \beta_2^{(\lambda)} \cos\left(\frac{2\pi t}{26}\right), \\ \log(\phi_{t,k}) &= \alpha^{(\phi)} + \beta^{(\phi)} \log(N_{t,k}) + \beta_1^{(\phi)} \sin\left(\frac{2\pi t}{26}\right) + \beta_2^{(\phi)} \cos\left(\frac{2\pi t}{26}\right),\end{aligned}$$

and with weights $w_{jk} = I(0 < \text{dist}_{jk} \leq 50\text{km})$, where dist_{jk} denotes the geographic distance between cities j and k in km. As a first step in our model selection strategy we compare this model to one using weights $w_{jk} = d_{jk}^{-\rho}$, which corresponds to a power-law distance relationship with $d_{jk} = \lceil \text{dist}_{jk} / 50\text{km} \rceil$. Based on AIC such a power-law distance kernel is preferred (AIC $3.172 \cdot 10^5$ vs. $3.189 \cdot 10^5$) and we hence proceed analysing the power-law version. Figure 5 shows the model fit decomposed into the three components exemplarily for the three largest and three smallest cities. We observe that the endemic and neighbourhood components only play a small role in the measles transmission. However, neither excluding the endemic component (AIC $3.1981 \cdot 10^5$) nor excluding the neighbourhood component (AIC $3.1977 \cdot 10^5$) provides

a better AIC. Furthermore, Fig. 6 shows the weight quantifying neighbourhood interaction

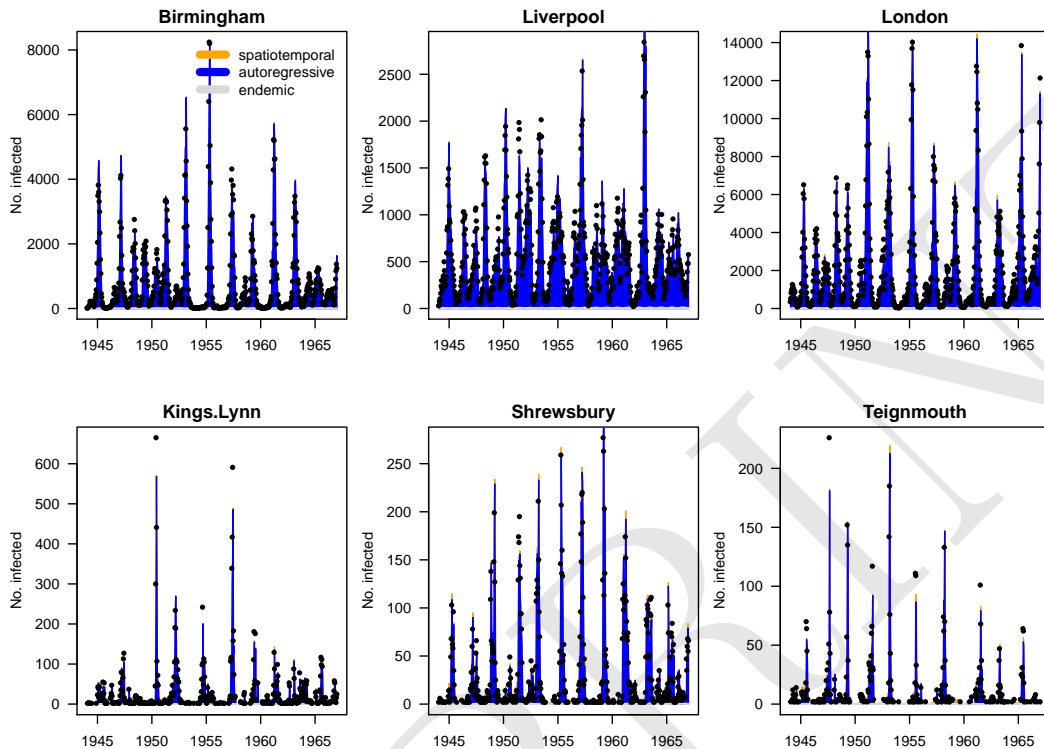


Figure 5: Time series of counts for the three largest (top row) and three smallest (bottom row) cities. Also shown are the model predicted expectations decomposed into the three components: endemic, within city and from outside cities.

as a function of distance (in steps of 50km). The right panel of the figure contrasts the powerlaw model with a model containing individual coefficients for lags 1-4; this gives some indication that the distance influence might be stronger at short lags than implied by the powerlaw model. Also, the AIC of this more flexible model is slightly better ($AIC\ 3.198 \cdot 10^5$). However, there is not enough information in the data to fit individual coefficients for lags 5 and higher.

The seasonal component in the two components is best illustrated graphically, c.f. Fig. 7. We note that the seasonality of the dominating within city transmission component has a shape very similar to what was found in previous work, i.e. with a lower transmission during the summer time. On the other hand, during summer time imports from external sources appear more likely. The interpretation of the neighbourhood transmission is slightly inconclusive as it is shifted compared to the two other – this component only makes up a small part of the overall transmission though.

4.3 Results of the mTSIR mimicking HHH model

When $\alpha = 1$ and the auto-regressive part is just $I_{t,k-1} + m_{t,k}$ we can, as described in Sect. 3.7, mimic the behaviour of the mTSIRmodel by using a HHH model with

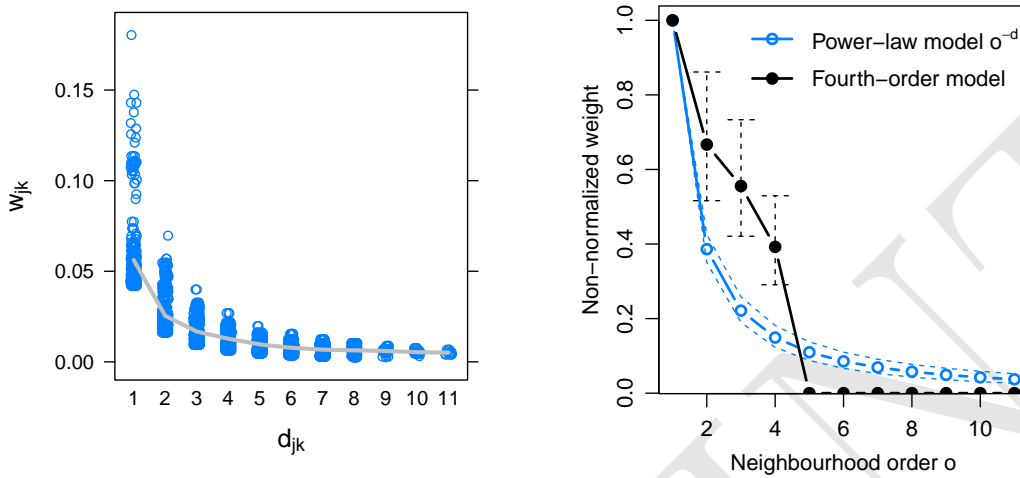


Figure 6: (left) Normalized weights as a function of d_{jk} . (right) Difference between the power-law model and a model containing individual coefficients for each neighbourhood order 1-4 and zero weight thereafter.

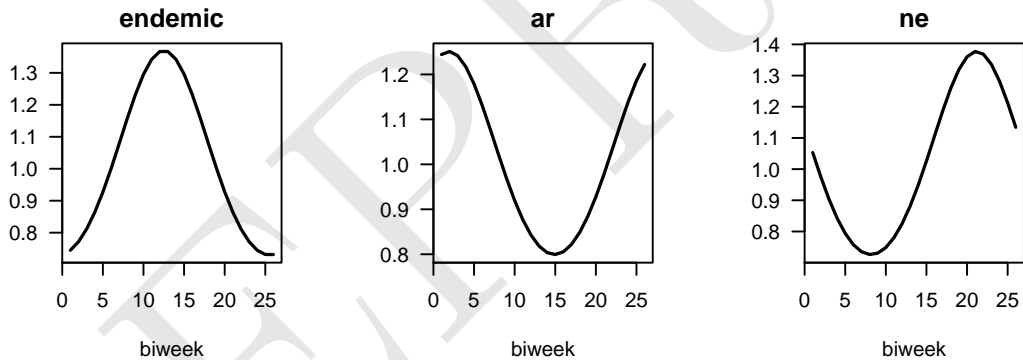


Figure 7: Illustration of the seasonal terms of the endemic, auto-regressive and neighbourhood component of the powerlaw HHH model.

$$\log(\lambda_{t,k}) = \beta_1^{(\lambda)} \sin\left(\frac{2\pi t}{26}\right) + \beta_2^{(\lambda)} \cos\left(\frac{2\pi t}{26}\right) + \log\left(\frac{S_{t-1,k}}{N_{t,k}}\right),$$

$$\log(\phi_{t,k}) = \log(\theta) + \beta_1^{(\phi)} \sin\left(\frac{2\pi t}{26}\right) + \beta_2^{(\phi)} \cos\left(\frac{2\pi t}{26}\right) + (\tau_1 - 1) \log(N_{t,k}) + \log(S_{t-1,k}),$$

and $w_{jk} = d_{jk}^{-\rho}$, which corresponds to a power-law distance relationship. In both these components the last term of the predictor represents an offset – an alternative would have been to use terms of the type $\beta_3^{(\cdot)} \log(S_{t-1,k})$ in order to address additional non-linearity in the susceptibles as suggested in the original TSIR formulations (Finkenstädt and Grenfell, 2000; Finkenstädt et al., 2002). The results obtained from fitting this model to the 60 cities data are given in Tab. 2.

By transformation we find the corresponding mTSIR model parameters to be $\hat{\theta} = 2.5 \cdot 10^{-5}$ (95% CI: $1.9 \cdot 10^{-5}$ - $3.4 \cdot 10^{-5}$), $\hat{\tau}_1 = 1.10$ (95% CI: 1.08-1.13) and $\hat{\rho} = 0.84$ (95%

	Estimate	2.5 %	97.5 %
ar.sin(2 * pi * t/26)	1.35	1.20	1.50
ar.cos(2 * pi * t/26)	4.12	4.05	4.18
ne.1	-10.58	-10.87	-10.29
ne.log(N)	0.10	0.08	0.13
ne.sin(2 * pi * t/26)	-0.20	-0.25	-0.15
ne.cos(2 * pi * t/26)	-1.31	-1.35	-1.27
newweights.d	0.84	0.75	0.93
overdisp	2.72	2.68	2.76

Table 2: Parameter estimates and 95% CIs for the mTSIR mimicking HHH model.

CI: 0.75-0.93). It also becomes clear that the clumping parameter of the negative binomial is estimated to be at a very different scale than suggested by the mTSIR model. As an additional sensitivity analysis we investigate whether the addition of an endemic component of the form $\log(\nu_{t,k}) = \log(S_{t,k-1}) + \alpha^{(\nu)}$ is mandated. This would indicate that there is additional influx of measles cases from unspecified sources, e.g. imports from other cities than the 60 included in the analysis. However, and as expected for measles, the AIC of the original model is not improved by this extension (difference in the loglikelihood almost immeasurable). Altogether, no substantial improvement over the best fitting model in Sect. 4.2 is seen though.

5 Discussion

Focus in this manuscript was on infectious disease data where disease location was known to occur at a fixed and known set of units (e.g. cities or county centroids) hence ultimately leading to the analysis of multivariate time series of counts. Underlying, the epidemic process may, however, be continuous in space: For infectious diseases such continuous-space continuous-time models have been developed in, e.g., [Lawson and Leimich \(2000\)](#) as well as [Meyer et al. \(2012\)](#). Other recent advances have been a more direct focus on the actual who-infected-who transmission chains manifesting as spatio-temporal point patterns. This could for example be based on a covariate augmented network model of the underlying contact structure in the population ([Groendyke et al., 2012](#)). Another path is to supplement the reconstruction with available microbiological information about the disease strain in the cases, e.g., by taking the difference in RNA sequences as additional genetic distance metric supplementing spatial distance ([Aldrin et al., 2011](#); [Ypma et al., 2012](#)). Others have taken the collapse of geographical space even further by visualizing and computing only effective-distances based on, e.g., mobility or population flux data ([Brockmann and Helbing, 2013](#)). Spatio and spatio-temporal methods have also been used for the detection of disease clusters ([Kulldorff, 2001](#); [Diggle et al., 2005](#); [Lawson and Kleinman, 2005](#)) Another application of spatial methods in infectious disease epidemiology has been the source identification of large food-borne outbreaks. Here, point-patterns of cases are aligned with knowledge about the spatial and spatio-temporal distribution of food items or appropriate proxies ([Manitz et al., 2014](#); [Kaufman et al., 2014](#)).

In a world of commuting and travelling, 'location' of a case is at best a 'most likely' or 'average' of the individuals location over time. In the digital age new data sources emerge allowing an even more exact spatio-temporal tracking of individuals. Hence, location is destined to become a more dynamic concept in future spatio-temporal epidemic models. In general and for stigmatizing diseases in particular, positional accuracy is destined to be in conflict with data protection of the individual. It is a matter of discussion and evaluation

how the right balance is to be found between the two.

Addressing spatial heterogeneity can be one aspect to improve on the usefulness of epidemic models, but certainly also complicates the analysis. When such models are used to support the decision making processes on intervention, control and vaccination, it is important to remember how formidable a task the modelling of infectious diseases is and where the limitations lie. From a statistician's point of view transmission modelling in practice misses a stronger emphasis on parameter estimation, uncertainty handling and model selection. While the modelling community has published high-impact semi-mechanistic models, the concurrent advances in the field by the statistical community have gone somewhat unnoticed. In the last 10 years, though, mechanistic modelling, data science and statistical inference for dynamic processes appear to have synthesized more. This means that Markov Chain Monte Carlo, approximate Bayesian computation and plug-and-play simulation methods are now standard instruments for spatially enriched epidemic models (Mugglin et al., 2002; McKinley et al., 2009; He et al., 2010; Jandarov et al., 2014). Furthermore, flexible open-source R packages have become available for visualizing, analysing and simulating such epidemic models. In particular, the R package `surveillance` (Höhle et al., 2015) implements both the two component SIR model and the HHH as illustrated in Meyer et al. (2014). However, with powerful computational tools at hand it is also worthwhile to discuss, how complex models need to be in order to answer the questions they are designed for. Long computation times often imply little room for model diagnostics or model selection, which is equally important. The power of mathematical models lies in their abstraction as May (2004) reminds us. What to include and what to abstract upon in order to make a model *relevant* requires an interdisciplinary approach.

Acknowledgments

Sebastian Meyer, University of Zurich, is thanked for his input and comments on the part of HHH modelling in this chapter.

References

- Abrams, S., P. Beutels, and N. Hens (2014). Assessing mumps outbreak risk in highly vaccinated populations using spatial seroprevalence data. *American Journal of Epidemiology* 179(8), 1006–1017.
- Aldrin, M., T. M. Lyngstad, A. B. Kristoffersen, B. Storvik, Ø. Borgan, and P. A. Jansen (2011). Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *Journal of the Royal Society Interface* 8(62), 1346–1356.
- Anderson, R. M. and R. M. May (1991). *Infectious diseases of humans: dynamics and control*. Oxford University Press.
- Andersson, H. and T. Britton (2000). *Stochastic Epidemic Models and their Statistical Analysis*, Volume 151 of *Springer Lectures Notes in Statistics*. Springer-Verlag.
- Bailey, N. T. J. (1950). A simple stochastic epidemic. *Biometrika* 37(3/4), 193–202.
- Bailey, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley.

References

- Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Griffin.
- Becker, N. G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall/CRC.
- Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1–59. (with discussion).
- Bjørnstad, O. N., B. F. Finkenstädt, and B. T. Grenfell (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model. *Ecological Monographs* 72(2), 169–184.
- Bjørnstad, O. N. and B. T. Grenfell (2008). Hazards, spatial transmission and timing of outbreaks in epidemic metapopulations. *Environmental and Ecological Statistics* 15, 265–277.
- Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88(421), 9–25.
- Britton, T. (2010, May). Stochastic epidemic models: a survey. *Mathematical Biosciences* 225(1), 24–35.
- Brockmann, D. and D. Helbing (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science* 342(6164), 1337–1342.
- Cliff, A. and P. Haggett (2004). Time, travel and infection. *British Medical Bulletin* 69, 87–99.
- Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics* 65, 1254–1261.
- Daley, D. J. and J. Gani (1999). *Epidemic Modelling*. Cambridge University Press.
- Deardon, R., X. Fang, and G. Kwong (2015). Statistical modelling of spatio-temporal infectious disease transmission. In D. Chen, B. Moulin, and J. Wu (Eds.), *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases*. John Wiley & Sons.
- Diekmann, O. and J. A. P. Heesterbeek (2000). *Mathematical epidemiology of infectious diseases*. Wiley.
- Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research* 15(4), 325–336.
- Diggle, P. J., B. Rowlingson, and T.-L. Su (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 16, 423–34.
- Diggle, P. J. and B. S. Rowlingson (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A* 157(3), 433–440.
- Dohoo, I., W. Martin, and H. Stryhn (2010). *Veterinary Epidemiologic Research* (2nd ed.). VER Inc.

- Eggo, R. M., S. Cauchemez, and N. M. Ferguson (2011, Feb). Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *Journal of the Royal Society Interface* 8(55), 233–243.
- Erlander, S. and N. F. Stewart (1990). *The Gravity Model in Transportation Analysis: Theory and Extensions*. VSP.
- Fahrmeir, L. and T. Kneib (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press.
- Finkenstädt, B. F., O. N. Bjørnstad, and B. T. Grenfell (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* 3(4), 493–510.
- Finkenstädt, B. F. and B. T. Grenfell (2000). Time series modelling of infectious diseases: a dynamical systems approach. *Journal of the Royal Statistical Society, Series C* 49(2), 187–205.
- Geilhufe, M., L. Held, S. O. Skrovseth, G. S. Simonsen, and F. Godtliebsen (2014, May). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal* 56(3), 363–382.
- Gibson, G. J. and E. Renshaw (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics applied in Medicine & Biology* 15, 19–40.
- Giesecke, J. (2002). *Modern infectious disease epidemiology* (2nd ed.). Hodder Arnold.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81(25), 2340–2361.
- Gog, J. R., S. Ballesteros, C. Viboud, L. Simonsen, O. N. Bjornstad, J. Shaman, D. L. Chao, F. Khan, and B. T. Grenfell (2014). Spatial Transmission of 2009 Pandemic Influenza in the US. *PLoS Computational Biology* 10(6), e1003635.
- Grenfell, B. T. (2006). TSIR analysis of measles in England and Wales. <http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm>. Available from the Internet Archive as <http://web.archive.org/web/20060311051944/http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm>.
- Grenfell, B. T., O. N. Bjørnstad, and J. Kappey (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414(6865), 716–723.
- Grenfell, B. T., A. Kleczkowski, C. A. Gilligan, and B. M. Bolker (1995). Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Statistical Methods in Medical Research* 4(2), 160–183.
- Groendyke, C., D. Welch, and D. R. Hunter (2012, Sep). A network-based analysis of the 1861 Hagelloch measles data. *Biometrics* 68(3), 755–765.
- He, D., E. L. Ionides, and A. A. King (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface* 7(43), 271–283.

References

- Held, L., M. Höhle, and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling* 5, 187–199.
- Held, L. and M. Paul (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* 54(6), 824–843.
- Held, L. and M. Paul (2013). Statistical modeling of infectious disease surveillance data. In N. M. M’ikanatha, R. Lynfield, C. A. Van Beneden, and H. de Valk (Eds.), *Infectious Disease Surveillance* (2nd ed.). Wiley-Blackwell.
- Hens, N., Z. Shkedy, M. Aerts, C. Faes, P. Van Damme, and P. Beutels (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*. Springer.
- Herzog, S. A., M. Paul, and L. Held (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology and Infection* 139(4), 505–515.
- Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical Journal* 51(6), 961–978.
- Höhle, M., E. Jørgensen, and P. O’Neill (2005). Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society, Series C* 54(2), 349–366.
- Höhle, M., S. Meyer, and M. Paul (2015). *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*. R package version 1.8-4.
- Isham, V. (2004). Stochastic models for epidemics. Technical Report 263, Department of Statistical Science, University College London. Available as <https://www.ucl.ac.uk/statistics/research/pdfs/rr263.pdf>.
- Jandarov, R., M. Haran, O. Bjørnstad, and B. Grenfell (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society: Series C* 63(3), 423–444.
- Kaufman, J., J. Lessler, A. Harry, S. Edlund, K. Hu, J. Douglas, C. Thoens, B. Appel, A. Kasbohrer, and M. Filter (2014). A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. *PLoS Comput. Biol.* 10(7), e1003692.
- Keeling, M. J., O. N. Bjørnstad, and B. T. Grenfell (2004). Chapter 17 - metapopulation dynamics of infectious diseases. In I. Hanski and O. E. Gaggiotti (Eds.), *Ecology, Genetics and Evolution of Metapopulations*, pp. 415 – 445. Burlington: Academic Press.
- Keeling, M. J. and P. Rohani (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.
- Kermack, W. O. and A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society, Series A* 115, 700–721.
- Klinkenberg, D., J. de Bree, H. Laevens, and M. C. M. de Jong (2002). Within- and between-pen transmission of classical swine fever virus: a new method to estimate the basic reproduction ratio from transmission experiments. *Epidemiology and Infection* 128, 293–299.

-
- Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A* 164, 61–72.
- Lau, M. S. Y., G. Marion, G. Streftaris, and G. J. Gibson (2014). New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of The Royal Society Interface* 11(93).
- Lawson, A. and K. Kleinman (Eds.) (2005). *Spatial and Syndromic Surveillance for Public Health*. Wiley.
- Lawson, A. and P. Leimich (2000). Approaches to the space-time modelling of infectious disease behaviour. *IMA Journal of Mathematics Applied in Medicine and Biology* 17(1), 1–13.
- Liu, W., H. W. Hethcote, and S. A. Levin (1987). Dynamical behavior of epidemiological models with nonlinear incidence rates. *Journal of Mathematical Biology* 25, 359–380.
- Madden, L. V., G. Hughes, and F. van den Bosch (2007). *The Study of Plant Disease Epidemics*. APS Press.
- Manitz, J., T. Kneib, M. Schlather, D. Helbing, and D. Brockmann (2014). Origin Detection During Food-borne Disease Outbreaks - A Case Study of the 2011 EHEC/HUS Outbreak in Germany. *PLoS Currents* 6.
- May, R. M. (2004). Uses and abuses of mathematics in biology. *Science* 303(5659), 790–793.
- McKinley, T. J., A. R. Cook, and R. Deardon (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* 5(1), Article 24.
- Meyer, S., J. Elias, and M. Höhle (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* 68(2), 607–616.
- Meyer, S. and L. Held (2014). Modelling power-law spread of infectious diseases. *Annals of Applied Statistics* 8(3), 1612–1639.
- Meyer, S., L. Held, and M. Höhle (2014). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Available as <http://arxiv.org/pdf/1411.0416>.
- Morton, A. and B. Finkenstädt (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series C* 54(3), 575–594.
- Mugglin, A. S., N. Cressie, and I. Gemmell (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine* 21(18), 2703–2721.
- Neal, P. and G. O. Roberts (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* 5(2), 249–261.
- O’Neill, P. D. and N. G. Becker (2001). Inference for an epidemic when susceptibility varies. *Biostatistics* 2(1), 99–108.
- O’Neill, P. D. and G. O. Roberts (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A* 162, 121–129.
-

References

- Paul, M. and L. Held (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* 30(10), 1118–1136.
- Paul, M., L. Held, and A. M. Toschke (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 27, 6250–6267.
- Pitzer, V. E., C. Viboud, L. Simonsen, C. Steiner, C. A. Panozzo, W. J. Alonso, M. A. Miller, R. I. Glass, J. W. Glasser, U. D. Parashar, and B. T. Grenfell (2009). Demographic variability, vaccination, and the spatiotemporal dynamics of rotavirus epidemics. *Science* 325(5938), 290–294.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Renshaw, E. (1991). *Modelling Biological Populations in Space and Time*. Cambridge University Press.
- Riley, S., K. Eames, V. Isham, D. Mollison, and P. Trapman (2014). Five challenges for spatial epidemic models. *Epidemics*. Epub ahead of print, DOI doi:10.1016/j.epidem.2014.07.001.
- Scheel, I., M. Aldrin, A. Frigessi, and P. Jansen (2007). A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming. *Journal of the Royal Society Interface* 4, 699–706.
- Smyth, G. K. (1998). Numerical integration. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*, pp. 3088–3095. Wiley.
- Viboud, C., O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* 312(5772), 447–451.
- Weidemann, F., M. Dehnert, J. Koch, O. Wichmann, and M. Höhle (2014). Bayesian parameter inference for dynamic infectious disease modelling: rotavirus in Germany. *Statistics in Medicine* 33(9), 1580–1599.
- Wilking, H., M. Höhle, E. Velasco, M. Suckau, and T. Eckmanns (2012). Ecological analysis of social risk factors for rotavirus infections in Berlin, Germany, 2007-2009. *International Journal of Health Geographics* 11(1), 37.
- Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC.
- Xia, Y., O. N. Bjørnstad, and B. T. Grenfell (2004, Aug). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *American Naturalist* 164(2), 267–281.
- Ypma, R. J., A. M. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W. M. van Ballegooijen (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. Biol. Sci.* 279(1728), 444–450.