LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Michael Höhle

Spatio-temporal epidemic modelling using additive-multiplicative intensity models

Technical Report Number 041, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Spatio-temporal epidemic modelling using additive-multiplicative intensity models

Michael Höhle[(1,2,3)*]

[(1)] Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany
[(2)] Department of Mathematical Sciences, Technische Universität München, Munich, Germany
[(3)] Munich Center of Health Sciences, Germany

Abstract

An extension of the stochastic susceptible-infectious-recovered (SIR) model is proposed in order to accommodate a regression context for modelling infectious disease surveillance data. The proposal is based on a multivariate counting process specified by conditional intensities, which contain an additive epidemic component and a multiplicative endemic component. This allows the analysis of endemic infectious diseases by quantifying risk factors for infection by external sources in addition to infective contacts. Simulation from the model is straightforward by Ogata's modified thinning algorithm. Inference can be performed by considering the full likelihood of the stochastic process with additional parameter restrictions to ensure non-negative conditional intensities.
As an illustration we analyse data provided by the Federal Research Centre for Virus Diseases of Animals, Wusterhausen, Germany, on the incidence of the classical swine fever virus in Germany during 1993-2004.

1 Introduction

Today, infectious diseases remain a threat to human and animal health. Emerging and re-emerging pathogens – like SARS, influenza, hemorrhagic fever among humans or foot and mouth disease and classical swine fever among animals – keep public authorities on go. As a consequence there has been an interest in human, veterinary and plant epidemiology to gain insight into disease dynamics by the use of stochastic modelling. Typical epidemic models are variations of the so called susceptible-infectious-recovered (SIR) model described in e.g. Becker (1989) and Andersson and Britton (2000). These models are well investigated for homogeneous populations and software exists to apply them in practice (Höhle and Feldmann, 2007).

In this paper we are especially interested in heterogeneous populations. For such populations there has been a development in the literature of explaining heterogeneity by covariates (Lawson and Leimich, 2000; Neal and Roberts, 2004; Diggle, 2006). We generalize such trends by attempting a regression view of infectious diseases where the dynamics of the disease are quantified by covariates. Our work thus contains a contribution to the cooperation of health researchers, epidemiologists and statisticians on determining ecological drivers of such infectious disease dynamics. Inspired by the work of Diggle (2006) a spatial SIR model is formulated based on conditional intensities. By considering the possible location of events as known beforehand, e.g. farms where outbreaks can occur, the dynamics of the disease can be described by a marked temporal point process.

A shortcoming of the model in Diggle (2006) is that only a single outbreak of the disease is modelled. Scheel et al. (2007) compensate for this by adding a single source representing infection

*Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany. Email address: hoehle@stat.uni-muenchen.de

from unknown sources and which is infective through the entire observation period. We generalize their formulation and decompose the conditional intensity function in an endemic and an epidemic component. Such modelling is similar to models used for count data time series in public health surveillance (Held et al., 2005). The proposed model has conditional intensities similar to the hazard rates in the additive-multiplicative model known from survival analysis (Lin and Ying, 1995; Sasieni, 1996; Martinussen and Scheike, 2002). However, contrary to survival modelling the stochastic processes of individuals in the context of epidemic modelling interact with each other: Once an infection occurs other individuals have a higher risk of becoming infected. We thus have a mutually exciting multivariate point process model, where direct disease transmission between individuals is quantified by the spatial distance and where infections from external sources are modelled in dependence of covariates. To our knowledge this explicit formulation of epidemic modelling as a two component point process is new and provides an useful modelling tool for practical applications.

To illustrate application we show how to use the model on classical swine fever virus (CSFV) data from two federal states in Germany. CSFV is a highly contagious virus disease infecting domestic pigs and wild boars. It has a great economic impact when occurring in countries with large industrialized pig populations. A major problem in the eradication of the disease in Germany is that the virus in certain areas has become endemic in the wild boar population. Using genetic typing investigations have shown, that 59% of the primary outbreaks in domestic pigs were due to indirect contact to infected wild boars (Fritzemeier et al., 2000). It is not clear how the exact transmission occurs – possible CSFV transmission routes could be the direct contact between free-ranging or inappropriately restricted domestic swine, introduction of infected carcasses and feed, or indirect transmission through contaminated equipment and persons. We use the proposed model to quantify the disease transmission between domestic pigs and wild boars.

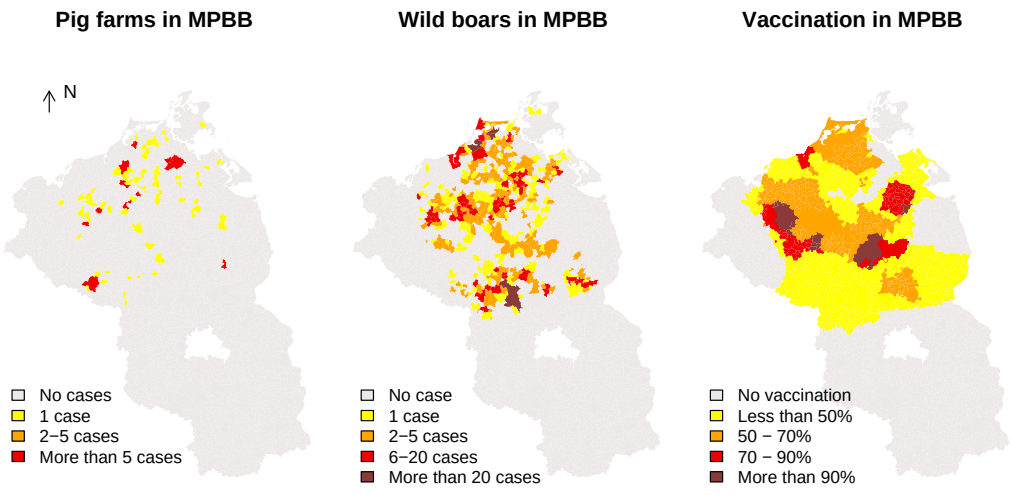
This paper is organized as follows. Section 2 introduces the classical swine fever virus data as motivating example. Section 3 presents the extension of the SIR model, whereas Sections 4 and 5 discuss simulation and inference in the proposed model. Section 6 gives results for the CSFV data and a discussion finalizes treatment.

2 The CSFV data

The local veterinary authorities of the federal states Mecklenburg-western Pomerania (MP) and Brandenburg (BB) provided information on all outbreaks of classical swine fever among domestic pig farms to the Federal Research Centre for Virus Diseases of Animals, Wusterhausen, Germany. The study period was 1993-2004 for the two federal states. For each infected farm the detection date, the size and its spatial location at municipality level was provided. The actual spatial coordinates are known to the authorities, but can not be used directly by us due to restrictions on privacy protection.

Data on the number of farms for each municipality and their mean size is taken from a 2005 production survey in the two federal states. This information is assumed to be representative for the study period of 1993-2004. As part of the CSFV surveillance routines sera from a random sample of about 5% of the wild boars shot were investigated for CSFV (Staubach et al., 2002).

In Mecklenburg-w.P. the number of infected pig farms during the study period was 67 (corresponding to 9% of the total farms in the federal state), while a total of 826 infected wild boars were found in the sample of the hunting bag. For Brandenburg 14 farms were infected (1% of the total farms in the state) and 287 CSFV infected wild boars were found. Molecular typing revealed that the MP and BB outbreaks among domestic pigs originate from the same strains (Fritzemeier et al., 2000). This information and the geographic proximity made us treat MP and BB as one common area in the further analysis. A municipality was declared to have infected wild boars at time t, if shot animals were diagnosed with CSFV within 60 days before t and 30 days after t. Figure 1



(a) (b) (c)

Figure 1: Spatial distribution of CSFV incidence among (a) pig farms and (b) wild boars in the municipalities of Mecklenburg-western Pomerania and Brandenburg. Panel (c) shows for each municipality the relative duration of vaccination areas compared to the length of the study period (13 years).

shows the spatial distribution of CSFV incidence among (a) pig farms and (b) wild boars for the MPBB data.

As part of the eradication strategy, attempts were made to vaccinate the wild boar population through oral immunisation (Kaden et al., 2005). Vaccination was performed by a suspension in a bait of corn, which was manually distributed in the selected areas using a density of 30-40 baits/km^2. A municipality was declared to be part of a vaccination area at time t if a vaccination area covers the municipality within 60 days before and 60 days after t. Figure 1(c) shows the relative duration of the vaccination period compared to the study period 1993-2004 for each municipality.

Figure 2 shows the corresponding temporal incidence of the reported CSFV cases among domestic pig farms and wild boars. The pig farm time series has two peaks as a result of two epidemics during 1993-1995 and 1997-1998. Note also the strong seasonality of the hunting bag data and the gaps in the domestic pig series, which indicate that several outbreaks have taken place.

Important questions one would like to answer is whether there is time-wise interaction between CSFV infection among wild and domestic pigs and if there is a difference in the infectious behaviour in regions being part of vaccination areas. Ways of answering such questions would be a classical two-by-two table analysis possibly stratified by time (Lachin, 2000) or an analysis using autoregressive logistic models (Diggle et al., 2002). Instead we propose a more mechanistic continuous time model using a multivariate counting process, which allows for explicit modelling of the disease components and takes censoring into account. Thus the dependence of data due to the infective nature of the disease is taken into consideration and no choice of appropriate discretization of the time scale is required.

3 Spatial SIR Model

Our modelling framework is the stochastic susceptible-infectious-recovered (SIR) compartmental model described in e.g. Andersson and Britton (2000). With little effort this modelling is extendable

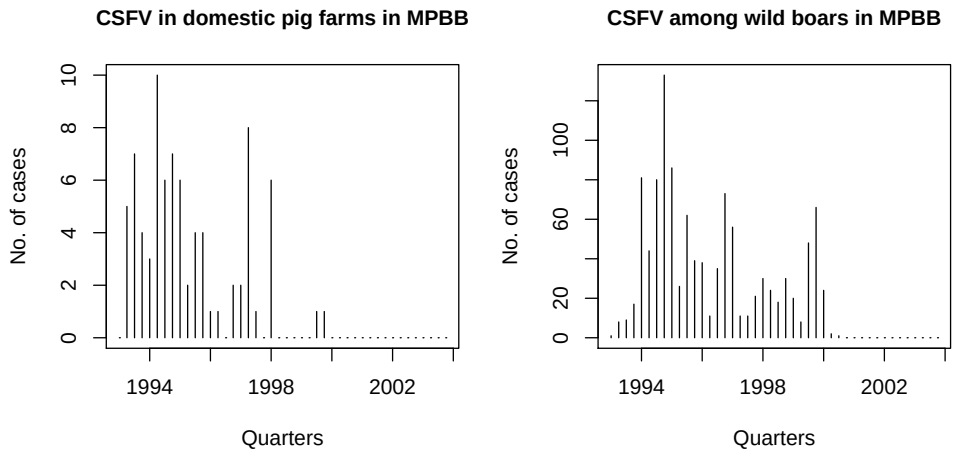


Figure 2: Number of CSFV infections among pig farms and wild boars per quarter in Mecklenburg-western Pomerania and Brandenburg.

to e.g. S-Exposed-IR (SEIR) models containing an incubation time before an individual becomes infectious. Also, we will assume a closed population of size n, but generalizations to dynamic populations are possible.

Introducing notation we let $S(t)$ represent the set of all susceptible individuals just before time t and let $I(t)$ denote the set of infectious individuals just before time t. Two transitions are of interest: susceptibles becoming infectious and infectious individuals recovering. The durations between these two events for each individual, i.e. the length of the infectious period, can be assumed to be a fixed constant (as often done in practical applications) or realizations of independent and identically gamma distributed random variates, i.e. $T^I \sim \text{Ga}(\gamma^I, \delta^I)$.

Given the event history $\mathcal{H}_t$ up to but not including time t, the conditional intensity function at t for a state change from susceptible to infectious of individual $1 \leq i \leq n$ is assumed to be

$$\lambda_i(t|\mathcal{H}_t) = Y_i(t) \cdot [e_i(t|\mathcal{H}_t) + h_i(t)], \tag{1}$$

where $Y_i(t)$ is an at risk indicator for individual i. For example $Y_i(t) = \mathbb{1}_{(i \in S(t))}$, where $\mathbb{1}_{(\cdot)}$ is the indicator function. If right-censoring occurs because the epidemic is only observed until time T, then $Y_i(t) = \mathbb{1}_{(i \in S(t) \wedge t \leq T)}$. With the risk indicator it is thus easily possible to model re-infection as in a SIR-Susceptible (SIRS) model. To ease notation we omit the event history in the terms $S(t|\mathcal{H}_t)$ and $I(t|\mathcal{H}_t)$, but keep it in $e_i(t|\mathcal{H}_t)$ to stress dependence on the history of the process for this term.

Model (1) constitutes the regression framework by splitting the conditional intensity into endemic and epidemic components. The epidemic component $e_i(t|\mathcal{H}_t) \geq 0$ is included additively to underline the superposition of two stochastic processes. Similarly, $h_i(t) \geq 0$ represents the endemic risk for becoming infected, specifically it does not depend on the internal history of the process. This risk of infection from external sources can be time varying due to seasonality of the disease, but also spatial and individual heterogeneity exists due to e.g. population density, vegetation, control measures or the existence of disease vectors. For $h_i(t)$ we use a framework similar to the Cox model by expressing the endemic risk using a time-dependent baseline risk $\exp(h_0(t))$ with possible time dependent $q \times 1$ external covariate vector $\boldsymbol{z}_i(t)$ acting in a multiplicative fashion on the base risk:

$$h_i(t) = \exp\left(h_0(t) + \boldsymbol{z}_i(t)^T \boldsymbol{\beta}\right).$$

Note that all spatial heterogeneity has to be expressed through covariates, the baseline depends on time only. To ensure identifiability, $\boldsymbol{z}_i(t)$ can not contain an intercept term.

An advantage with respect to interpretability of (1) is that if conceptually $h_i(t) = 0$ the resulting $\lambda_i(t|\mathcal{H}_t)$ corresponds to the conditional intensity function of an ordinary SIR model. Thus, if external sources of infection can be ruled out then the conditional intensity should be able to become zero, e.g. when there are no infectious individuals. On the other hand, if there is a baserisk then the conditional intensity should also stay positive even when there are no infectives. Had the combination of h_i and e_i instead been multiplicatively, an $e_i(t|\mathcal{H}_t)$ of zero would have resulted in $\lambda_i(t|\mathcal{H}_t)$ being zero making no further infections possible.

As mentioned, $e_i(t|\mathcal{H}_t)$ represents the epidemic individual-to-individual transmission of the disease. It is assumed that an adequate model is a distance weighted sum over the infective individuals:

$$e_i(t|\mathcal{H}_t) = \sum_{j \in I(t)} f(||\boldsymbol{s}_i - \boldsymbol{s}_j||).$$

Thus $f(u)$ is a parametric function of the e.g. Euclidean distance between the position $\boldsymbol{s}_i$ of individual i and the position $\boldsymbol{s}_j$ of individual j. Concerning $f(u)$ we will assume that the distance function can be represented by a linear basis expansion:

$$f(u) = \sum_{m=1}^{p} \alpha_m B_m(u), \tag{2}$$

with the B_m's being known functions. Thus $h_i(t) = 0$ and $f(u) = \alpha_1$ corresponds to the standard homogeneous SIR model with transmission parameter $\alpha_1 > 0$. Similarly, the grid based model in Höhle et al. (2005) can be described as $f(u) = \alpha_1 \mathbb{1}_{(u=0)} + \alpha_2 \mathbb{1}_{(0<u\leq N_4)}$, where $\alpha_1, \alpha_2 > 0$ and N_4 is the distance between immediate neighbours. Less straightforward are distance functions such as $f(u) = \alpha \exp(-\rho u)$ used in e.g. Diggle (2006) – here one has to resort to a linearization through a Taylor expansion. Another possibility is to use B-splines to represent $f(u)$ (Dierckx, 1995). Because $f(u)$ in all cases represents a distance kernel, an important restriction is that $f(u) \geq 0$ for all u within a predefined range $[a, b]$ covering all data points.

By interchanging the summation over infectious individuals and B-spline terms, an ordinary additive structure with time-varying covariates is obtained for the epidemic component:

$$e_i(t|\mathcal{H}_t) = \sum_{m=1}^{p} \alpha_m \sum_{j \in I(t)} B_m(||\boldsymbol{s}_i - \boldsymbol{s}_j||) = \sum_{m=1}^{p} \alpha_m \, x_{im}(t) = \boldsymbol{x}_i(t)^T \boldsymbol{\alpha},$$

where $x_{im}(t) = \sum_{j \in I(t)} B_m(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$ and with $\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))$ being $\mathcal{H}_t$-predictable, because $I(t)$ is left-continuous. Thus if (2) applies the resulting model in (1) has resemblance to additive-multiplicative hazard models known from survival analysis (Lin and Ying, 1995; Martinussen and Scheike, 2006):

$$\lambda_i(t|\mathcal{H}_t) = Y_i(t) \cdot \left[\boldsymbol{x}_i(t)^T \boldsymbol{\alpha} + \exp(h_0(t)) \exp\left(\boldsymbol{z}_i(t)^T \boldsymbol{\beta}\right)\right]. \tag{3}$$

By conditioning on the past as covariates we can model time to infection for an individual i using an additive-multiplicative model with time-varying covariates. This conditioning approach has similarities to modelling time series using autoregressive regression models or the piecewise Cox-model of Scheel et al. (2007).

Note however that in contrast to the survival context of the additive-multiplicative model in our application the multivariate counting process described by all individuals now has dependent paths. Another difference in our approach compared with the available inference and implementation for the additive-multiplicative hazard model described in Martinussen and Scheike (2006) is that $\boldsymbol{\alpha}$ is time constant and of direct interest. Depending on our choice of distance function $f(u)$ there are also non-negative constraints on $\boldsymbol{\alpha}$ to ensure a positive intensity function. Thus we proceed along the lines of (Lin and Ying, 1995) with additional care for parameter constraints.

Given $\mathcal{H}_t$ the overall conditional intensity function for the next transition of a susceptible individual is $\lambda^*(t|\mathcal{H}_t) = \sum_{i=1}^{n} \lambda_i(t|\mathcal{H}_t)$. This quantity will be the key component for model simulation and inference in Section 4 and 5.

The endemic component $h_i(t)$ consists of a location-independent base risk for infection from unknown sources and a component allowing the modelling of covariate effects. In the multiplicative approach of Diggle (2006) $h_0(t)$ is left unspecified and inference is performed for $\boldsymbol{\beta}$ based on the partial likelihood similar to Cox-regression. However, in our additive parametrisation a partial likelihood approach is only possible if $h_0(t)$ is also a multiplicative part of $e_i(t|\mathcal{H}_t)$ as e.g. in Scheel et al. (2007). Stressing the superposition of independent epidemic and endemic processes we prefer to have $h_0(t)$ only in the endemic component and use a parametric model for it, or as proposed in this paper, a piecewise constant function with and without smoothing.

4 Simulation

Obtaining process realisations by simulation is an important tool for model checking and prediction, because analytical results for the proposed type of complex stochastic process are not available. Furthermore, displaying results of simulated epidemics helps understand the dynamics of the disease and shows, whether the proposed model has the desired behaviour.

To simulate from the above continuous time stochastic process in the time interval $[0, \tau]$ several algorithms are possible. One option is to base simulation for the above marked point process on an adaptation of the inversion method described in Nicolai and Koning (2006). However, if time depending covariates only change value at a discrete set of time points and piecewise upper bounds can be found for the overall conditional intensity $\lambda^*(t|\mathcal{H}_t)$ at appropriate intervals of $[0, \tau]$ a faster alternative is *Ogata's modified thinning* algorithm (Daley and Vere-Jones, 2003) for marked point processes.

Denote by $L_c = \{c_1, \ldots, c_d\}$ the time points where the covariate vector $\boldsymbol{z}(t)$ changes for at least one individual i. Furthermore, denote by $L_r(t) = \{r_1, \ldots, r_n\}$ the $\mathcal{H}_t$-predictable set of recovery times with $r_i = \infty$ if i is not infected before t. Here, predictability is ensured, because the length of the infectious period T^I is either deterministic or can be simulated for each individual beforehand. Finally, let

$$L(t) = \min \Big\{ \{c_j \in L_c : c_j > t\} \cup \{r_i \in L_r(t) : r_i > t\} \Big\} \tag{4}$$

be the time after t of the next external change in the overall conditional intensity. Thus in the interval $(t, L(t)]$ the terms in the overall conditional intensity are constant except for $\exp(h_0(t))$. Hence, if an upper bound for $h_0(t)$ can be found for this interval it is easy to compute an upper bound $M(t)$ for $\lambda^*(t|\mathcal{H}_t)$. This leads to the following simulation algorithm, where i_K and r_K are the time of infection and recovery of individual K, respectively.

Algorithm 1: *Ogata's modified thinning* algorithm for marked point processes

1 Given current time t, update $L(t)$ and calculate local upper bound $M(t)$ for the overall conditional intensity $\lambda^*(t|\mathcal{H}_t)$;
2 Generate proposed waiting time $T \sim \text{Exp}(M(t))$;
3 **if** $t + T > L(t)$ **then**
4 | let $t = L(t)$;
5 **else**
6 | let $t = t + T$;
7 | Accept t with probability $\lambda^*(t|\mathcal{H}_t)/M(t)$; otherwise goto step 1;
8 | Draw index K of the next infective from the set $\{1, \ldots, n\}$ with respective probabilities $\lambda_i(t|\mathcal{H}_t)/\lambda^*(t|\mathcal{H}_t)$, $1 \leq i \leq n$;
9 | Update the event history and set $i_K = t$ and $r_K = t + T^I_K$;
10 goto step 1

With Algorithm 1 it is thus possible to quantify parameter uncertainty using a parametric bootstrap. Furthermore, prediction of quantities such as time to next event or risk for a specific individual to become infected within a given time can now be performed.

5 Inference

Inference for the proposed spatial SIR model will be based on a counting process formulation (Andersen et al., 1993). Interest will focus on the estimation of parameters relevant to individuals becoming infected.

Denote by $N_i(t), i = 1, \ldots, n$, the counting process, which for individual i counts the number of changes from state susceptible to state infectious. The corresponding intensity of $N_i(t)$ is $\lambda_i(t)$ as given in (1) with dropped $\mathcal{H}_t$ to slim notation. By $N(t) = (N_1(t), \ldots, N_n(t))$ we denote the multivariate counting process.

With $h_0(t)$ known and $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ the loglikelihood function of N when observed up to time τ is (Andersen et al., 1993; Martinussen and Scheike, 2006)

$$l(\boldsymbol{\theta}, \tau) = \sum_{i=1}^{n} \left\{ \int_0^\tau \log(\lambda_i(t)) dN_i(t) - \int_0^\tau \lambda_i(t) dt \right\},$$

where $dN_i(t) = N_i((t + dt)-) - N_i(t-)$ is the increment over the small time interval $[t, t + dt)$. A big advantage of the above counting process notation is that re-infections of an individual i are easily handled by appropriate specification of the at risk process $Y_i(t)$.

As a consequence, the $p + q$ dimensional score process has the form

$$S(\boldsymbol{\theta}, \tau) = \begin{bmatrix} \frac{\partial l(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\alpha}} \\ \frac{\partial l(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \int_0^\tau \frac{\boldsymbol{x}_i(t)}{\lambda_i(t)} dM_i(\boldsymbol{\theta}, t) \\ \sum_{i=1}^{n} \int_0^\tau \frac{\exp(\boldsymbol{z}_i(t)^T \boldsymbol{\beta}) \boldsymbol{z}_i(t) \exp(h_0(t))}{\lambda_i(t)} dM_i(\boldsymbol{\theta}, t) \end{bmatrix}$$

where $dM_i(\boldsymbol{\theta}, t) = dN_i(t) - \lambda_i(t)dt$. The expected information matrix can be estimated by

$$\mathcal{I}(\boldsymbol{\theta}, \tau) = \sum_{i=1}^{n} \int_0^\tau \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log(\lambda_i(t)) \right)^{\otimes 2} dN_i(t),$$

with $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for the column vector $\boldsymbol{a}$.

In case data are collected over a long observational period, dealing with censoring of the individuals is straightforward in the counting process framework. If $C_i(t)$ is the censoring process of individual i, e.g. if censoring occurs at time t_0 one has $C_i(t) = \mathbb{1}_{(t \leq t_0)}$, one would instead of $Y_i(t)$ operate with $Y_i^C(t) = C_i(t) \cdot Y_i(t)$.

The maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \arg\sup_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \tau)$ for $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\beta} \in \mathbb{R}^q$ can be found by bound constrained optimization of $l(\boldsymbol{\theta}, \tau)$ using e.g. a limited memory BFGS algorithm with gradient $S(\boldsymbol{\theta}, \tau)$ (Byrd et al., 1995). If operating with a B-spline based distance function f, it is straightforward to force the distance function to be monotone decreasing, because monotonicity constraints can easily be transformed into sufficient constraints on the α_m's as described in e.g. Dierckx (1995). Likelihood inference can still be performed by constrained optimization algorithms handling the inequality constraints between the α_m's, e.g. as in Lange (1994).

5.1 Penalized likelihood

If $h_0(t)$ is completely unspecified one modifies the above score process such that $h_0(t)$ disappears from the estimating functions as in Lin and Ying (1995). If a parametric representation is used then these coefficients enter $\boldsymbol{\beta}$ and the $\exp(h_0(t))$ in the nominator of the $\boldsymbol{\beta}$-part of the score process disappears.

As an alternative we shall use a non-parametric model for $h_0(t)$ based on the penalized likelihood framework with a set of r degree zero B-splines for $h_0(t)$ as in Fahrmeir and Klinger (1998). This adaptation provides an alternative to the non-parametric estimate of $\exp(h_0(t))$ in the additive-multiplicative intensity model. With fixed knot positions $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{r+1})$ the following non-parametric model for $h_0(t)$ is assumed:

$$h_0(t) = \sum_{j=1}^{r} \beta_{0j} B_{0j}(t), \text{ where } B_{0j}(t) = \mathbb{1}_{(\kappa_j \leq t < \kappa_{j+1})}.$$

Furthermore, letting $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0r})^T$ and redefining $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \beta_1, \ldots, \beta_q)^T$ the penalized loglikelihood and score function are

$$pl(\boldsymbol{\theta}, \tau) = l(\boldsymbol{\theta}, \tau) - \frac{\lambda}{2} \sum_{j=k+1}^{r} (\Delta^k \beta_{0j})^2 = l(\boldsymbol{\theta}, \tau) - \frac{1}{2} \lambda \boldsymbol{\beta}_0^T \boldsymbol{S}_0 \boldsymbol{\beta}_0,$$

$$pS(\boldsymbol{\theta}, \tau) = S(\boldsymbol{\theta}, \tau) - \lambda \sum_{j=k+1}^{r} \Delta^k \beta_{0j} \quad = S(\boldsymbol{\theta}, \tau) - \begin{bmatrix} \mathbf{0} \\ \begin{pmatrix} \lambda \boldsymbol{S}_0 \boldsymbol{\beta}_0 \end{pmatrix} \\ \mathbf{0} \end{bmatrix}$$

where λ is the smoothing parameter. Typically, one would use a first or second order difference penalty on $\boldsymbol{\beta}_0$ (Eilers and Marx, 1996). Hence, $\boldsymbol{\beta}_0^T \boldsymbol{S}_0 \boldsymbol{\beta}_0$ is the matrix equivalent of the penalty term, where the k'th order difference of $\boldsymbol{\beta}_0$ – recursively defined as $\Delta^k \beta_{0j} = \Delta^{k-1}(\beta_{0j} - \beta_{0(j-1)})$ and $\Delta^0 = 1$ – is penalized. An important question is how to choose λ in order to obtain an appropriate amount of smoothing for the data. General criterion for this selection known from e.g. generalized additive models (GAMs) do not immediately apply: there is no way to quantify effective number of parameters using a modified AIC criterion and cross validated criterion does not make sense as observations are not independent. Sect. 5.3 gives a proposal on how to deal with the selection of the smoothing parameter λ in our context.

5.2 Parameter uncertainty and model selection

Operating within a likelihood framework means that for the unconstrained parameters usual asymptotic results can be used to compute Wald, score or likelihood ratio tests (LRTs) (Andersen et al., 1993). Inversion of these tests provides confidence intervals. However, the constrained parameters of the epidemic component need special care. A reparametrization using e.g. $\psi_m = \log \alpha_m$ would remove the non-negative constraints on the α_m's, but to investigate the need for an epidemic component one would test $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ against $H_1 : \boldsymbol{\alpha} \geq \mathbf{0}$, which – reparametrization or not – is at the border of parameter space. Here we have adopted the notation that a test of H_0 against H_1 is to be read as testing H_0 against $H_0 \backslash H_1$. A large sample approximation of the LRT statistic under the nullhypothesis in this constrained setting with nuisance parameters is the chi-bar-square distribution (Silvapulle and Sen, 2005, Sect. 4.3):

$$LRT = 2 \left[\sup_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \in \mathbb{R}^q} l(\boldsymbol{\theta}, \tau) - \sup_{\boldsymbol{\alpha} = 0, \boldsymbol{\beta} \in \mathbb{R}^q} l(\boldsymbol{\theta}, \tau) \right] \overset{a}{\sim} \bar{\chi}^2(\mathcal{I}^{\alpha\alpha}(\boldsymbol{\beta})),$$

where $\mathcal{I}^{\alpha\alpha}(\boldsymbol{\beta})$ is the $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ block of $\mathcal{I}^{-1}(\boldsymbol{\theta}, \tau)$ evaluated at $\boldsymbol{\theta} = (\mathbf{0}^T, \boldsymbol{\beta}^T)^T$ and $\bar{\chi}^2(\boldsymbol{V})$ for the positive definite matrix $\boldsymbol{V}$ is defined by the following transformation of $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{V})$:

$$\bar{\chi}^2(\boldsymbol{V}) = \boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{Z} - \min_{\boldsymbol{\alpha} \geq \mathbf{0}} (\boldsymbol{Z} - \boldsymbol{\alpha})^T \boldsymbol{V}^{-1} (\boldsymbol{Z} - \boldsymbol{\alpha}).$$

Note that for a specific realization of $\boldsymbol{Z}$ the minimization problem in the 2nd term can be solved by quadratic programming, which makes it possible to compute e.g. $P(\bar{\chi}^2(\boldsymbol{V}) \leq t)$ for $t > 0$ by simulation. Because the true value of $\boldsymbol{\beta}$ is not known a point estimate of the p-value in the constrained LRT is obtained by replacing $\boldsymbol{\beta}$ in the above with the estimate of $\hat{\boldsymbol{\beta}}$ under H_0.

Operating within a likelihood framework also means that model selection is possible by e.g. AIC. However, parameter constraints reduce the average increase in the maximized loglikelihood - thus the penalty for constrained parameters should be smaller than the factor two used in the ordinary definition of AIC. One-sided AIC (OSAIC) suggested by Hughes and King (2003) is such a proposal when p out of $k = p + q$ parameters have non-negative constraints:

$$\text{OSAIC} = -2l(\boldsymbol{\theta}, \tau) + 2\sum_{g=0}^{p} w(p, g)(k - p + g),$$

where $w(p, g)$ are p-specific weights. In case of $p = 1$ the weights are $w(1, 0) = w(1, 1) = \frac{1}{2}$ and the total penalty is thus $2(k-1)+1$. For $p = 2$ constrained parameters with joint covariance matrix V the weights are $w(2, 0) = \frac{1}{2}\pi^{-1}\arccos(\rho_{12})$, $w(2, 1) = \frac{1}{2}$ and $w(2, 2) = \frac{1}{2} - \frac{1}{2}\pi^{-1}\arccos(\rho_{12})$ where ρ_{12} is the correlation coefficient $v_{12}(v_{11}v_{22})^{-\frac{1}{2}}$ between the two parameters. This or higher order weights can also be computed by the simulation approach suggested in Silvapulle and Sen (2005, Sect. 3.5).

5.3 Residual analysis

The OSAIC provides a criterion for selecting the best model from a set of competing models. However, additional graphical checks should be used in order to investigate whether the model is able to reproduce important features of the data. In our case we will use a residual analysis for point processes as described in Ogata (1988). Here, the estimated cumulative intensity

$$\hat{\Lambda}^*(t) = \int_0^t \hat{\lambda}^*(s|\mathcal{H}_s)ds$$

is used to transform the observed S→I event-times $\{t_i\}$ of the spatial SIR process to the time scale $\{\tau_i = \Lambda^*(t_i)\}$. If the estimated overall conditional intensity $\hat{\lambda}^*(t|\mathcal{H}_t)$ describes the true underlying conditional intensity well, the sequence $\{\tau_i\}$ should behave like a stationary Poisson process with intensity 1. As a consequence, the variables

$$Y_k = \tau_k - \tau_{k-1} = \Lambda^*(t_k) - \Lambda^*(t_{k-1}), \quad k = 2, \ldots,$$

should be realisations of independent and identical Exp(1) distributed variables and hence $U_k = F(Y_k) = 1 - \exp(-Y_k) \overset{\text{iid}}{\sim} U(0, 1)$. A graphical check of this property is to plot the observed u_k against the empirical CDF $\hat{F}_U(u)$ and compare with the straight line. A Kolmogorov-Smirnov (K-S) test against the uniform distribution provides a test for this property by using the test statistic

$$D = \sup_{0 \leq u \leq 1} \left| \hat{F}_U(u) - u \right|.$$

This test statistic also provides a first criterion for selecting the smoothing parameter λ from Sect. 5.1: select λ such that the K-S test statistic D is minimized.

6 Results

For the CSFV data in MPBB we use a SEIR model with assumed removal at detection, and – based on the findings of the 1997-1998 CSFV epidemic in the Netherlands – a fixed incubation

time of 7 days and a fixed infectious period of 4.6 days (Stegeman et al., 1999). Given the long observational period and the few cases the between-farm transmission of the disease is expected to play a minor role. Hence, only basic models for the between farm transmission are investigated in order to quantify whether any between farm transmission occurs at all.

As a first attempt we use the parametric model $h_0(t) = \beta_{\text{intercept}} + \beta_{\text{t}} \cdot t$ for the baseline hazard function. In practice this means that $h_0(t)$ is a piecewise constant function with change-points at all event times. A homogeneous transmission is used for the epidemic component and interest is in investigating the endemic and epidemic components for the CSFV data. The time-varying covariates boars and vacc are used as explanatory covariables in the endemic component and the resulting model selection based on OSAIC is shown in Tab. 1. The results in the table underline the necessity of the epidemic component, because these models obtain a better OSAIC. Alternatively, this can be investigated by a test of $H_0 : \alpha_1 = 0$ vs. $H_1 : \alpha_1 \geq 0$ in e.g. model 1 from Tab. 1 using the described LRT test procedure ($p = 0.079$). Hence, at a $\alpha = 0.05$ level of significance there is insufficient evidence for a homogeneous farm-to-farm transmission. Furthermore, presence of wild boars and vaccination areas appear to provide additional explanatory power.

model	$B_1(u) = \mathbb{1}_{(0 \leq u)}$	intercept	t	boars	vacc	OSAIC
1	+	+	+	+	+	1966.53
2	+	+	+	+	-	1981.14
3	+	+	+	-	-	2023.43
4	+	+	-	-	-	2107.52
5	-	+	+	+	+	1967.53
6	-	+	+	+	-	1987.67
7	-	+	+	-	-	2033.45
8	-	+	-	-	-	2119.26

Table 1: Model selection based on OSAIC, which brings out model 1 as the one having smallest OSAIC. Symbol + indicates presence of the term in the model.

If the distance kernel in model 1 is replaced by a kernel $B_1(u) = \mathbb{1}_{(0 \leq u < 50km)}$ and $B_2(u) = \mathbb{1}_{(50km \leq u)}$ (denoted model 9) this results in an OSAIC of 1956.90. Dropping B_2 in model 9 and only admitting short-distance spread by B_1 (denoted model 10) improves OSAIC to 1955.82. Alternatively, tests for $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ in model 9 and 10 yield p-values of 6.71e-04 and 1.82e-04, respectively.

Based on the above OSAIC computations we thus settle for model 10 as basis for a more detailed analysis. The parameter estimates are: $\alpha_1 = 3.69 \cdot 10^{-5}$, $\beta_{\text{intercept}} = -10.66$, $\beta_{\text{t}} = -0.61$, $\beta_{\text{boars}} = 2.36$ and $\beta_{\text{vacc}} = 1.34$. This means that in a situation with no infectives, i.e. the epidemic component being zero, the intensity in areas with infected boars is $\exp(\beta_{\text{boars}}) = 10.54$ times larger as in areas with no infected wild boars. Similarly, the intensity in vaccination areas is increased by a factor of $\exp(\beta_{\text{vacc}}) = 3.83$. At first this effect of vaccination might be surprising. However, the vaccination areas were carefully selected by authorities as those areas with high risk of infection. Thus without any additional available covariates explaining this selection, vaccination is an indicator for an increased risk. 95% confidence intervals for $\beta_{\text{intercept}}, \beta_{\text{t}}$, β_{boars} and β_{vacc} in model 10 based on profile loglikelihoods are shown in Figure 3. Both β_{boars} and β_{vacc} are thus seen to be significant at the $\alpha = 0.05$ level.

A residual analysis as shown in Fig 6(a) however reveals that model 10 fails to capture some behaviour of the data around the first third of time. The Kolmogorov-Smirnov test against uniformity of the residuals is rejected at the $\alpha = 0.05$ significance level, because $p = 0.023$. To improve on this and to gain additional insight into the endemic component we replace $\beta_{\text{intercept}}$ in model 10 with a piecewise constant function with a total of $r = 8$ degree zero B-splines as described in Sect. 5.1 while keeping the term $\beta_{\text{t}} \cdot t$ in the baseline hazard. Knots are based on the respective octiles of the observed event times and the smoothing parameter λ is determined by a grid search shown in Fig. 4. The K-S test statistic $D(\lambda)$ is minimized by a value of $\lambda \approx 0.25$.

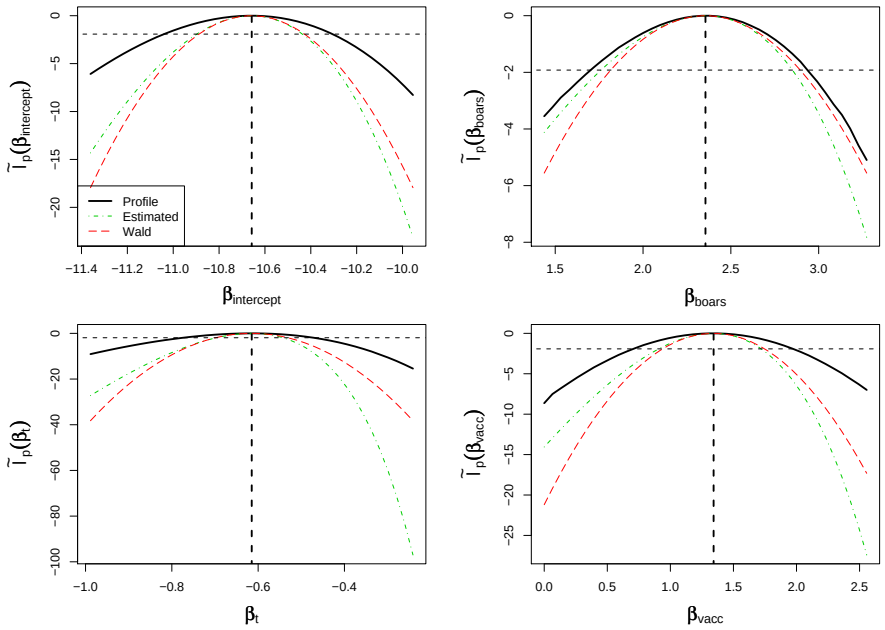


Figure 3: Profile loglikelihood, Estimated likelihood and a quadratic approximation for each parameter. Also shown are the intersection with $-\frac{1}{2}\chi^2_{0.95}(1)$, which yields LRT based 95% confidence intervals.

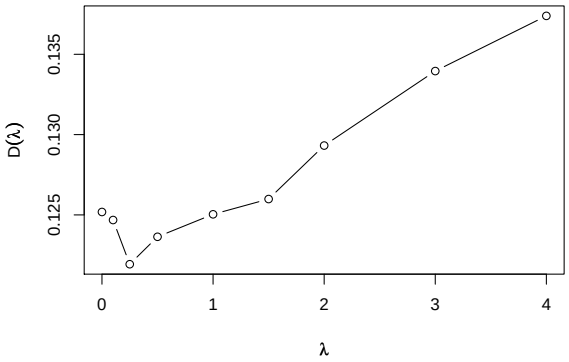


Figure 4: K-S test statistic $D(\lambda)$ as a function of the smoothing factor λ.

Figure 5 shows the estimated penalized $\beta_{\text{intercept}}(t)$ term of the baseline hazard together with a pointwise 95% Wald confidence interval. Also shown are the total intensity and a plot of the proportion $\sum_{i=1}^{n} e_i(t|\mathcal{H}_t)/\lambda^*(t|\mathcal{H}_t)$, which illustrates how large a proportion the epidemic intensity makes up of the overall intensity. Notice the peaks in the baseline hazard around days 600 and 1500, which could not be handled by the single constant β_0 baseline hazard in model 10. Figure 6(b) shows the improved residuals with the K-S test now having a p-value of 0.372.

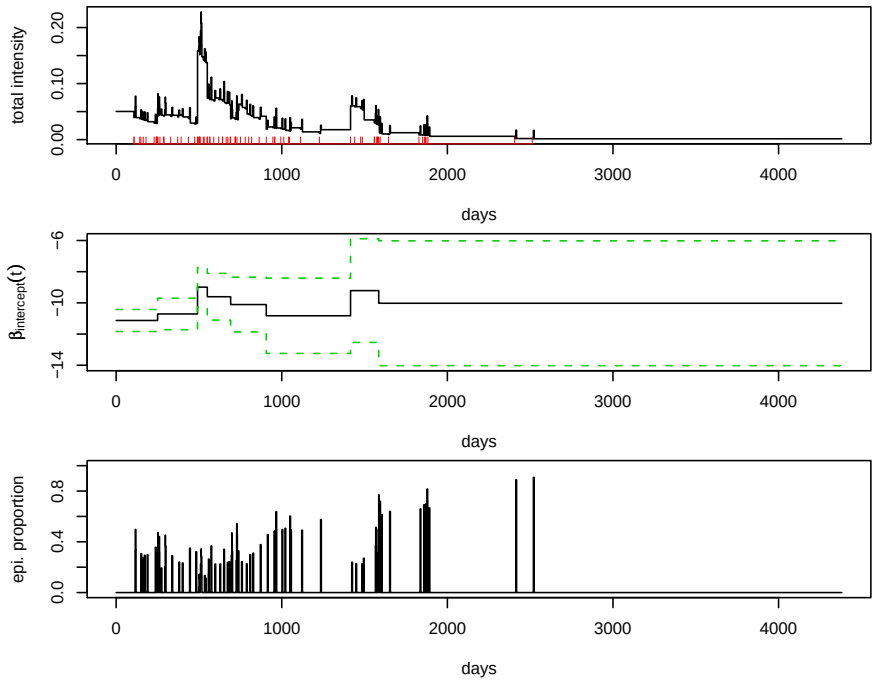


Figure 5: Plot of the total intensity $\lambda^*(t|\mathcal{H}_t)$, the piecewise constant intercept part of $h_0(t)$ (together with a 95% CI) and the epidemic proportion for the piecewise exponential model. A rug in the top figure shows the 81 observed S→E event times in days since January 1st, 1993.

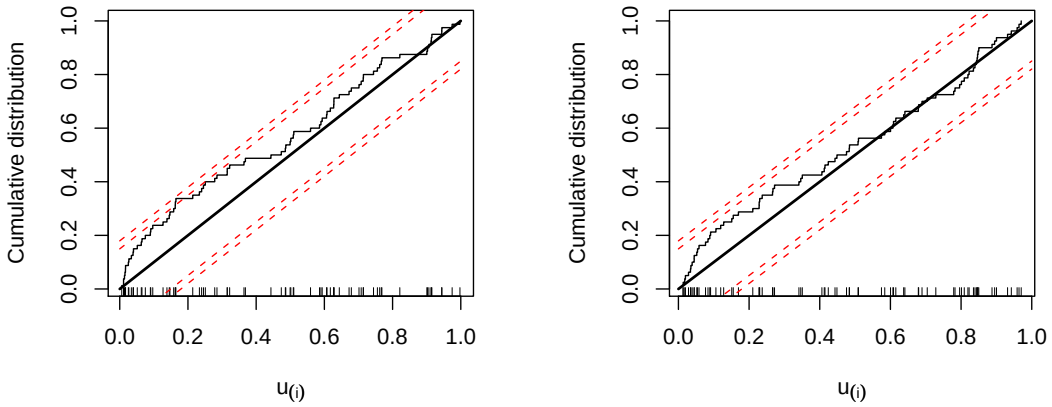
7 Discussion

The presented spatial SIR model is a step towards a regression approach in stochastic epidemic modelling of spatio-temporal infectious disease surveillance data. Combining a Cox model and a spatial heterogeneous SIR model covers endemic and epidemic components. Prediction of e.g. time to next event or probability of infection within 1 year is possible by simulation.

In our work we have assumed full observability of process events. However, in many realistic settings only partial observability might be the case. Besides imputation of the missing observations a solution could be to treat model inference in a Bayesian setting as e.g. in Höhle et al. (2005). A Bayesian setting would also allow for a natural formulation of the non-negative constraints on the $\boldsymbol{\alpha}$ using prior distributions and would yield credibility regions for the baseline hazard by formulating the penalization as a prior. However, design and implementation of an efficiently mixing Markov chain Monte Carlo (MCMC) sampler would require a careful analysis while still depending on evaluating $l(\boldsymbol{\theta}, \tau)$, which can become quite time consuming for large data sets. For our application and to provide a routinely usable general regression framework for infectious disease surveillance data we thus prefer for now an efficient implementation of constrained maximum likelihood over a Bayesian setting.

One possible modelling extension could be to expand the epidemic component with additional terms in order to make infectivity depend on covariates as e.g. in Lawson and Leimich (2000), where the strength of infectivity of an individual is a time depending function. However, for the illustrating CSFV application this was not of immediate interest.

Had we used a multiplicative composition between h_i and e_i, the framework of Kneib and Fahrmeir (2007) could have immediately been used, which also would have made allowance for spatial effects in $h_i(t)$ through e.g. Gaussian Markov random fields. However, the additive structure of the intensity gives a more realistic behaviour for infectious diseases. Future research has to show how their mixed model based approach to hazard estimation could be adopted to the proposed



(a) model 10

(b) penalized model

Figure 6: Empirical distribution of U_k compared to the CDF of a uniform distribution for (a) model 10 and (b) the penalized baseline hazard model. Also shown are the 95% and 99% error bounds derived from the Kolmogorov-Smirnov test statistic.

additive-multiplicative model.

The presented methods will be available in the new version of the R-package for epidemic modelling `RLadyBug` available from the comprehensive R Archive Network (CRAN). A preliminary version of the software and its documentation can be found in (Meyer and Höhle, 2008).

8 Acknowledgements

Thanks to Sebastian Meyer for valuable programming contributions and discussions on the simulation and inference of the proposed model and to Inga Tschöpe who performed many of the initial data manipulations on the CSFV data. Also thanks to Ludwig Fahrmeir and Jesper Møller for providing suggestions and comments and Christoph Staubach, Federal Research Institute for Animal Health, Germany for providing us with the data.

References

Andersen, P., Borgan, O., Gill, R., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer.

Andersson, H., Britton, T., 2000. Stochastic Epidemic Models and their Statistical Analysis. Vol. 151 of Springer Lectures Notes in Statistics. Springer-Verlag.

Becker, N. G., 1989. Analysis of Infectious Disease Data. Chapman & Hall/CRC.

Byrd, R., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific and Statistical Computing 16 (5).

Daley, D., Vere-Jones, D., 2003. An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods. Springer.

Dierckx, P., 1995. Curve and Surface Fitting with Splines. Oxford University Press.

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., Heagerty, P., Liang, K.-Y., Zeger, S. L., 2002. Analysis of Longitudinal Data, 2nd Edition. Oxford University Press.

Eilers, P., Marx, B., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11 (2), 89–121.

Fahrmeir, L., Klinger, A., 1998. A nonparametric multiplicative hazard model for event history analysis. Biometrika 85 (3), 581–592.

Fritzemeier, J., Teuffert, J., Greiser-Wilke, I., Staubach, C., Schlüter, H., Moennig, V., 2000. Epidemiology of classical swine fever in Germany in the 1990s. Veterinary Microbiology 77, 29–41.

Held, L., Höhle, M., Hofmann, M., 2005. A statistical framework for the analysis of multivariate infectious disease surveillance data. Statistical Modelling 5, 187–199.

Höhle, M., Feldmann, U., 2007. RLadyBug – an R package for stochastic epidemic models. Computational Statistics and Data Analysis, Special Issue on Statistical Software 52 (2), 680–686.

Höhle, M., Jørgensen, E., O'Neill, P., 2005. Inference in disease transmission experiments by using stochastic epidemic models. Journal of the Royal Statistical Society Series C 54 (2), 349–366.

Hughes, A., King, M., 2003. Model selection using AIC in the presence of one-sided information. Journal of Statistical Planning an Inference 115, 397–411.

Kaden, V., Hänel, A., Renner, C., Gossger, K., 2005. Oral immunisation of wild boar against classical swine fever in Baden-Württemberg: development of seroprevalences based on the hunting bag. Eur J Wildl Res 51, 101–107.

Kneib, T., Fahrmeir, L., 2007. A mixed model approach for geoadditive hazard regression. Scandinavian Journal of Statistics 34, 207–228.

Lachin, J. M., 2000. Biostatistical Methods - The Assessment of Relative Risks. Wiley.

Lange, K., 1994. An adaptive barrier method for convex programming. Methods and Applications of Analysis 1 (4), 392–402.

Lawson, A., Leimich, P., 2000. Approaches to the space-time modelling of infectious disease behaviour. IMA Journal of Mathematics Applied in Medicine and Biology 17 (1), 1–13.

Lin, D., Ying, Z., 1995. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. The Annals of Statistics 23 (5), 1712–1734.

Martinussen, T., Scheike, T., 2002. A flexible additive multiplicative hazard model. Biometrika 89 (2), 283–298.

Martinussen, T., Scheike, T., 2006. Dynamic Regression Models for Survival data. Statistics for Biology and Health. Springer.

Meyer, S., Höhle, M., 2008. Towards spatio-temporal epidemic modelling in R. Tech. rep., Department of Statistics, Ludwig-Maximilians-Universität München, Germany, expected to be available from Oct 2008.

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

Nicolai, R., Koning, A., 2006. A general framework for statistical inference on discrete event systems. Tech. Rep. 45, Econometric Institute Report, Erasmus University Rotterdam.

Ogata, Y., 1988. Statistical models for earthquake occurences and residual analysis for point processes. Journal of the American Statistical Association 83 (401), 9–27.

Sasieni, P. D., 1996. Proportional excess hazards. Biometrika 83 (1), 127–141.

Scheel, I., Aldrin, M., Frigessi, A., Jansen, P., 2007. A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming. Journal of the Royal Society Interface 4, 699—706.

Silvapulle, M. J., Sen, P. K., 2005. Constrained Statistical Inference. Wiley.

Staubach, C., Schmid, V., Knorr-Held, L., Ziller, M., 2002. A Bayesian model for spatial wildlife disease prevalence data. Preventive Veterinary Medicine 56, 75–87.

Stegeman, A., Elbers, A. R. W., Bouma, A., de Smit, H., de Jong, M. C. M., 1999. Transmission of classical swine fever virus within herds during the 1997-1998 epidemic in the Netherlands. Preventive Veterinary Medicine 42, 201–218.